This is to certify that the
thesis entitled

RATERS' CULTURAL BACKGROUND AS
RATER-TEXT INTERACTION

presented by

CHEN WANG

has been accepted towards fulfillment
of the requirements for the

MA degree in Teaching English to Speakers of Other Languages

Paula Winke
Major Professor's Signature

May 5, 2010
Date

DATE DUE	DATE DUE	DATE DUE

CULTURAL EXPERIENCE AS RATER-TEXT INTERACTION

By

Chen Wang

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

MASTER OF ARTS

Teaching English to Speakers of Other Languages

2010

ABSTRACT

CULTURAL EXPERIENCE AS RATER-TEXT INTERACTION

By

Chen Wang

This study investigates whether raters' cultural influence is a source of rater-text interaction in L2 writing assessment. Twenty-seven ESL/EFL teachers participated in the study. They were divided into three groups: raters who are American but have never been to China; raters who are American and have taught English in China for more than one year; and raters who are Chinese native speakers. All raters received rater training before rating. In the rating experiment, all raters assigned a score on a scale from 1 to 5, judged the best and worst feature of the essays, and provided some comments. No significant differences were found between the scores given by the two groups of American raters, but significant differences between the Chinese raters and the two groups of American raters were obtained. The qualitative data, which consisted mainly of the raters' comments, demonstrated that each group of raters attended to different aspects of the essays. The paper also discusses the possible influences of raters' cultural background on the scoring of second writing assessment. Finally, limitations of the study and suggestions for future studies of raters in writing assessment are given by the researcher.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

Researchers in the field of writing assessment have identified several factors that may influence the test scores of test takers, including the writing task, the written text itself, the scale being used, characteristics of the raters, characteristics of the writers other than their writing ability, and various contextual factors (Weigle, 2002). These variables may cause problems concerning test validity. For example, the scores may be different for examiners depending on the tasks they write about or the rater who evaluates their essays. In order to increase the validity of the writing tests, research should focus on identifying and explaining the source of variability in authentic writing assessment contexts and how they influence the accuracy of essay test scores (Broad, 2003; Cumming, 1997; Deville & Chalhoub-Deville; Huot, 2002). The overall aim of such research is to ultimately inform test developers and those who use test scores in how to check for, avoid, or correct such construct-irrelevant variance that may appear in test sores.

Much research on second language writing assessment has focused on the study of raters' variability that may cause a potential source of measurement error. It is well known that human raters introduce subjective factors into the procedure of scoring, for instance, the raters themselves may become tired, prefer one writing style over another, or have personal experience that influences the way they assign the scores. Rater variability may pose test fairness issues and reduce the test's reliability and validity (Kondo-Brown, 2002). Test takers may get different scores depending on the raters who rate or when they rate. It is of great importance to identify the sources of rater variability, and "if found to exist, to address them accordingly" (Johnson & Lim, 2009, p.486). The commonly used measures for statistically estimating rater variability are inter-rater reliability (rater agreement) and intra-rater reliability (self-consistency). Weigle (2002) defines inter-rater reliability as "the tendency of different raters to give the same scores to the same scripts" and intra-rater reliability as "the tendency of a rater to give the same score to the same script on different occasions" (p. 135). However, "the measures of inter- and intra- rater reliability only tell us the product of the writing assessment, yet nothing do we know about the process" (Johnson & Lim, 2009, p.486). Connor-Linton (1995b) explained that "if we do not know what raters are doing (and why they are doing it), then we do not know what their ratings mean" (p.763).

The study of rating process can address many questions in the area of writing assessment. Cummings (1990) argues that the understanding of the rating process could reveal the validity of different rating scales and procedures. However, he mentioned that a focus on rater reliability is not adequate. Lumley (2002) stated that the investigation of testing validity and reliability is concerned with how the process of rating is conducted. Studying the rating process can improve our understandings of the backwash effects of different rating scales and procedures, and may illuminate the kinds of evaluation that are consistent with instructional goals and teacher beliefs about the construction of writing scales (Connor-Linton, 1995b).

CHAPTER 2

Literature

Rater Bias Pattern

One area of the research that has been done to address problems with the rating process is rater bias patterns. There are a range of different forms of rater variability (Bachman & Palmer, 1996; Mc Namara, 1996; Weigle, 2002; Lumley, 2005): raters may interpret the scoring rubrics differently; raters may show various degrees of leniency and severity in their scoring procedure; raters may comply with the rating scales in different degrees. This category of study is rather quantitative, characterized by a large number of essays that have been rated, and uses advanced statistical tools, such as two-mode clustering and multi-faceted Rasch measurement. Eckes (2008) hypothesized that experienced raters may fall into different categories that value scoring rubrics in different degrees. In his study, sixty-four raters participated in a large-scale writing assessment and were asked to indicate on a four-point scale how much importance they attached to a certain score criteria that covered many aspects of a writing, such as fluency, completeness, and grammatical correctness. Eckes used many-facets Rasch analysis and two-mode clustering and found six distinguishable rater types. For example, the "syntax" type refers to raters who had a strong focus on language ability and task completeness.

Schaefer (2008) studied rater bias of 40 native English speakers who rated essays similar to the TOEFL Test of Written English (TWE). The essays were examined on a rating scale that contained six categories: Content, Organization, Style and Quality of Expression, Language Use, Mechanics, and Fluency. Schaefer found recurring rater bias through a Rasch analysis and he concluded that there were several subgroups of bias patterns. One subgroup of raters may rate Content and/or Organization severely and Language Use and/or Mechanics leniently and vice versa. Another subgroup of raters showed more severity to more advanced writers and more leniency to lower level writers, and vice versa.

Kondo-Brown (2002) used four experienced Japanese English teachers to rate students' essays. Although the inter-rater reliability and intra-rater reliability was very high, the severity or leniency pattern was significant between raters. She found that raters were more likely to show severity to extremely low or high ability writers and suggested more clearer criterion or rater training are needed for test takers in the extreme levels.

Rating Procedure and Rater Background

In response to Cumming's (1990) and Connor-Linton's (1995b) call for more research to be done to study the rating procedure mainly through analyzing the think-aloud protocol from raters while they rate the essays, Cumming, Kantor, and Powers (2002) focused on the exploration of rater's decision making-procedures. A general sequence of decision-making through the rating process was identified and proved to be almost consistent when applying the framework to a following study using different raters and different writing tasks. Cumming, Kantor and Powers found that "raters attended more extensively to rhetoric and ideas (compared to language) in compositions they scored high than in compositions they scored low" (p.67). Teachers who taught English as second/ foreign language, compared to EMT (English-mother-tongue) teachers, attended to language of the essays more extensively than the idea and organization. Most teacher raters were aware that their teaching experience and rating experience had some influence on their rating behavior.

When scoring the essays, raters play two roles: first as a reader; then as a judge. Like other readers, raters will also bring their reading expectations and using some reading strategies while scoring essays. Several studies addressed the relationship between background of the readers and the effects of they brought to the rating process. Kobayashi and Rinnert (1996) conducted a study that focused on reader's different backgrounds, such as their L1, academic status, amounts of writing instruction, and their experience in evaluating EFL students' essays. Four groups of participants joined the study and they were Japanese students who had not received formal English writing instruction, Japanese students who were experienced English writers, Japanese EFL teachers and native speaker English teachers. It was found that culturally influenced rhetorical patterns made a difference in the raters' evaluations. For example, Japanese students without English writing instruction gave better ratings to essays with a Japanese rhetorical pattern and native English teachers gave higher scores to essays with American rhetorical patterns. However, there was no significant difference in the evaluations between native English teachers and experienced Japanese EFL teachers. In the wake of the previous study, Rinnert and Kobayashi expanded their study in 2001, when they used four groups of readers in Japan with a different amount of exposure to ESL writing to investigate if their findings from 1996 would be replicated. The researchers worked out a list of different characteristics between English writing and Japanese writing and manipulated 16 different essays with various textural features of Japanese and English, such as "overall organization moving from general to specific " and "a thesis statement aiming to convince the reader of a particular position" as typical features in English writings, in contrast with "overall organization moving from specific to general" and "no strong specific position taken by the writer, thus leaving more up to the reader" as typical feature in Japanese writing(Rinnert & Kobayashi, 2001, p.192). They found that more experienced readers had a tendency to favor the features of the second language (L2) while less experienced readers showed a preference for the features of their first language (L1).

Clearly, human raters may exhibit various interactions with the reading passages they rate or have biases when scoring essays from ESL learners. Researchers have contributed extensively to the study of raters' background on their rating process. Although, the raters' background is a complicated issue, in order to construct a finely-tuned study, researchers often target one or two traits of raters' characteristics. To summarize, factors such as professional background, linguistic background, and cultural background of raters could alone or together have an effect on the rating procedure.

Professional Backgrounds

It is easy to understand that professional background may influence raters' behavior. Rating experience, rater expert and lay rater, clearly will be different in their performance. Teaching experience plays a role as well, whether raters are English-as-a-first-language, or English-as-a-second language teachers may make a difference in their reaction to the ESL essays.

Shohamy, Gordon, and Kraemer (1992) had four groups of participants who were distinguished from each other in their professions, whether they were EFL teachers or native English speakers with other occupations, and the training experience, whether they received training before rating the essays. Through an ANOVA, significant differences were only found between raters who differed on training. Professional background in this study did not yield any significant difference and the researchers argued for the importance of rater training before rating is actually performed.

Brown (1991) intended to find if there was any difference between the scores given by teachers who are English teachers and those given by ESL teachers. Sixteen teachers in total participated and rated 112 essays from students who were native speakers of English or ESL learners. The result was that no significant differences were found between the two groups of raters in general. However, from the qualitative data, in which raters were required to label the best and the worst features of the essay, researchers found that raters may reach the same score from different interpretations of the essays. For example, English teachers considered Syntax Cohesion and Syntax more importantly, whereas ESL raters valued Organization more importantly, even though both groups assigned similar scores to the test takers.

Weigle, Boldt, and Valsecchi (2003) conducted a pilot study on different faculty members' response to students' essays. The 16 raters represented four different professional backgrounds: ESL, English, history, and psychology. They all graded two types of essays, text-responsible writing (TR) and non text-responsible writing (NTR) essays in which the former content is about students' response about their understanding of specific texts, usually a short article, and the latter is writing about their personal ideas and arguments on a general topic. Raters across all different backgrounds ranked content as the most important criterion for their judgments. Yet, raters from the English department appeared to favor grammar more than other raters. For different writing tasks, English teachers and ESL teachers treated the tasks in a different way in that they valued content to be more important in TR essays and grammar more important in NTR essays. However, the history teachers treat all the tasks evenly. The psychology teachers did not demonstrate a clear pattern.

Linguistic Background

With more and more ESL teachers who are non-native speakers becoming raters of second language writing, there have been several studies exploring the potential different perceptions held by native speaking (NS) raters and non-native (NNS) speaking raters.

Shi (2002) conducted a study to examine NS and NNS teacher's evaluation on essays from Chinese college students. Twenty-three teachers in each category participated in the study and ranked 10 essays according to the quality of the writing. The raters were required to write and rank three reasons for their rating. No significant differences were found between the two groups of the raters; however, they differed from each other in the comments concerning the positive and negative features of the essays. NS teachers gave more positive comments regarding content and language, while NNS teachers showed more negative comments on the organization and length of the essays. It was found that NS teachers tended to give lower scores even though they gave more positive comments than NNS speakers.

In a recent study, Johnson and Lim (2009) employed a quantitative approach to find if there is any rater-bias pattern caused by the different L1 background of raters, and more importantly by the L1 background of the second language writers. Seventeen raters, including four NNS raters, scored the writing part of Michigan English Language Assessment Battery (MELAB). It is important to note that most of the raters were considered to be experienced because they all had received rater training and had gotten a certificate to serve as a long-term rater in the University of Michigan language assessment program. The essay pool was from a large variety of examinees from different language backgrounds. The researcher applied FACTETS and found that no rater bias could be attributed to the language backgrounds of the raters. The small number of NNS raters, including one Korean rater who was born in Korea but mostly was raised in the US, may account for the result. Clearly, there is a need to conduct research with larger size of NNS raters.

An Educational Testing Service report (Erdosy, 2004) investigated the rating procedure of four raters who represented a mixed linguistic and professional background. Concurrent verbal protocols were used to produce qualitative data to explain the source of potential rater variability. The researcher mentioned the limited generalizability of this study but he found that without a scoring rubric, the raters were very likely to refer to their teaching experience. NNS raters may refer to their experience as an ESL learner. It was further argued that cultural and linguistic backgrounds may be viewed as more important than academic backgrounds in influencing the raters' behavior. Rating experience may impact rating strategies but have limited effects on the construction of the rating criteria. Erdosy suggested future study should involve more rater participants to make the results more generalizable.

Cultural Background

A few studies found that raters' background may interact with the text characteristics of second language learners. Kaplan (1966) suggested that writers' rhetorical patterns varied across cultures. Despite many problems with his characterization of the way different cultures write, contrastive rhetoric has become an important area of investigation. It was found that students with little L2 writing experience are likely to refer to their L1 knowledge in their L2 writing (Connor, 1996). Rinnert and Kobayashi (2001) proposed five distinctive features of Japanese writing and English writing. For example, "a thesis statement aiming to convince the reader of a particular position" in English writing is compared with "no strong specific position taken by the writer thus leaving more up to the writer" in Japanese (p. 192). The findings were consistent with earlier studies that Japanese students with less exposure to English writing favored Japanese rhetorical patterns. Hamp-Lyons (1989) found that experience with other languages could change NS raters' reactions to the English writing of learners from those language communities. The teachers were found to be more tolerant of features of L1 patterns in students' English writing.

In addition, some other studies have reported the different intellectual traditions held by different cultures. Ballard and Clanchy (1991) identified different learning approaches and thinking patterns between students from Asia and the western world. Asian students were encouraged by the education system to adopt a reproductive learning approach featured by rote learning and memorization. Challenging the teacher or text-book and critical thinking was less favored in the classroom. In contrast, students in Western society have adopted a speculative approach of learning in terms of applying critical thinking, questioning and evaluation. Wang (2009) compared test takers' perspectives on the differences of GRE writing and TOEFL writing and found that students hold negative opinions about GRE writing because topics in the writing prompts required critical thinking and logical reasoning, which they had not be taught in schools in China. Wang found that the Chinese test takers often avoided describing a person or event related to Chinese culture as a seemingly preferred or even learned test taking strategy, because they thought the American raters would have a hard time understanding their example.

Research Question

The previous research identified various rater bias patterns which could be partially attributed to raters' backgrounds which could interact with the text features of students' essays. As raters' professional backgrounds and linguistic backgrounds have been addressed in several studies, I think that there is a need for a study that explores the impact of raters' cultural experience on their rating procedures given by limited research focus in this area. It is suggested by the literature that the experiment should involve more raters in experiments to make findings more generalizable and robust. Hence, the current study will use a larger sample of participants and will address two variables related to the raters' backgrounds: cultural and linguistic backgrounds. The research question that guides this study is:

Do raters with different amounts of exposure to and experience with the culture of test takers differentially score the test takers' essays?

I hypothesize that raters' differing amounts and types of exposure to the test takers' culture will influence the scores the raters assign to the test takers' essays. Furthermore, I expect that the more experience the raters have with the culture, the easier it will be for them to understand the content of the L2 learners' essays. Consequentially, the ones who have more exposure and experience with the culture may be more tolerant of prevalence of the test takers' L1 rhetorical features, which could result in the raters assigning higher scores or producing more positive comments concerning the quality of the essays.

CHAPTER 3

Method

Participants

An email for invitation for participating in the study was sent to potential participants by the researcher. For participants who were around the university where the experiment took place, the researcher met the participant face to face and monitored the whole procedure. For participants who lived far away and were unable to meet with the researcher, they were sent all the materials with detailed instructions to ensure that they followed the experiment procedures step by step. Nine participants in the study completed the study on their own, while 18 participants met the researcher individually to finish the study.

Three groups of participants were invited to participate in the study. They were ESL/EFL teachers who were Americans and had never been to China, labeled as American raters with no Cultural Experience (ARNCE), Americans who have taught English as a foreign language in China for at least one year, labeled as American raters with cultural experience (ARCE), and Chinese ESL/EFL teachers who were born in China and received their secondary education there, labeled as Chinese raters (CR). The backgrounds of the three groups of raters varied in terms of their academic majors, teaching experience, and their familiarity with TOEFL rating rubrics for writing. All the participants filled out a background questionnaire (See appendix A). In addition, all the raters were required to self-evaluate their experience with rating essays and using TOEFL writing scales to be one of the following: new; having some experience; being expert-like. Table 1 shows that the three groups of participants were comparable in their academic background. Except for the group of American raters with cultural experience, which had three raters in other majors, the raters in other groups held a degree related to ESL, English, or Education. Regarding teaching experience, Chinese raters exceeded the other groups, with an average of 13.6 years of English teaching, while the other two groups are similar, 4.4 years by American raters without cultural experience averaging, and 3.4 years by American raters with cultural experience. Regarding rating experience, Chinese raters had the most experienced raters among which six raters considered themselves to be expert-like. American raters with cultural experience are the least experienced raters, with only one rater had some rating experience and the rest raters were new about it. Regarding to the familiarity of using TOEFL rubrics, American raters with no cultural experience was more experienced with five raters having some experience and one being expert-like.

Generally, American raters with cultural experience have the least professional experience with rating and using TOEFL rubrics. When interpreting the findings, the researcher was aware of these mediating variables and treated them with cautions. More details are given in the discussion and limitations part of this paper.

Table 1

Rater Backgrounds

Variables	Categories	ARNCE (n = 8)	ARCE (n = 10)	CR (n = 9)
Gender	Male	1	4	2
	Female	7	6	7
Major	ESL/English/Education	8	7	9
	Others		3	
Degrees	Bachelor	0	5	3
	Master or above	8	5	6
Teaching (yrs)	Group average	4.4	3.4	13.6
Experience with rating essays	New	1	7	0
	Some experience	6	3	3
	Expert	1	0	6
Experience with TOEFL rubrics	New	2	9	4
	Some experience	5	1	3
	Expert	1	0	1

Materials

The essay prompt used in this study was based on the TOEFL independent writing tasks. Students were asked to write an essay about whether they agreed or disagreed with a statement concerning the effects of globalization on their home culture (see Appendix B). They were told to develop an argument for their stance. The topic of globalization was selected internationally to promote test takers to write from their own cultural perspectives.

Nineteen essays were collected from students with various backgrounds. The essays were written by 16 Chinese international students and 3 international students from other countries, such as Iran, Korea, and Saudi Arabia. 16 students were from Intensive English Program (IEP) in the English Language Center and three students were graduate students majored in Communication, Journalism and Engineering. The reason why essays from other foreign students were included was to distract the raters' attention from the focus of the research, that is, to see how they reacted to the scoring of Chinese students' essays, but they were not included in the analysis. The lengths of the essays varied, ranging from 198 words to 444 words, with an average of 295 words.

Seven essays, including five benchmark essays and two sample practice essays used in the training session in this study were from the TOEFL writing benchmark essays (ETS, 2005). The seven essays were written by TOEFL test takers and the essay topic was about telling the truth. In the rating experiment, a score sheet was used and each rater put all of their response on the score sheet (see Appendix C).

The holistic rating rubric of TOEFL independent writing task (ETS, 2004) was used in the scoring procedure in this study (see Appendix D). The rationale for choosing this rating rubric is that most of the performance tests have adopted a similar holistic rating scale because it is easy and fast to use, and also mimics the natural reading process (Weigle, 2002). Additionally, the writing prompt was based on the current structure of TOEFL writing prompts, thus it was appropriate to have raters use a standard TOEFL rating rubric to rate the essays.

Procedure

Collecting and preparing students' essays.

The students were presented with the writing prompt and instructions (see Appendix B). They first read the prompt and then were given 40 minutes to write the essay in class. After that, the researcher typed the essays with all spelling and punctuation errors left uncorrected. The essays were typed to avoid any influence from hand writing on the rating process. The order of the 20 essays was randomized before they were given to the raters, and they all got the essays in the same order.

The rating experiment.

Raters first filled out a biographic questionnaire which inquired about their professional, cultural, and linguistic backgrounds. After that, they went through a rater training session. They were given the TOEFL rating rubric and TOEFL benchmark essays from different levels (1 to 5), which were about a different topic from the one in this study. The benchmark essays and the sample essays were written by TOEFL test takers. The raters read these materials to become familiar with the rating scales and features of essays at different levels. Then the raters were given two extra sample essays on that topic to practice using the TOEFL rating rubric. The raters rated an essay, and then were provided with the correct rating. The trainer went over reasons why the correct rating was correct. The trainer and the raters discussed discrepancies between their ratings and the correct rating. This process was repeated with the second sample essay. For the participants who could not be present in the experiment, the training session was conducted on their own. The researcher provided the reasons for correct ratings for the two practice essays, hoping that the participants could realize their discrepancies between their ratings and the correct rating and thus make some adjustments.

After the training session, raters began to rate 19 essays. They were told to refer to the rating rubric at any time. There were three main tasks for the raters during the rating procedure: give a score (from 1 to 5) to each essay, provide a comment in two or three sentences based on the strangeness and weakness of each essay, and mark the best and worst features of each essay they could perceive. The features of an essay included Content, Organization, Coherence, and Language. However, the researcher did not provide any supporting materials to explain these four features of an essays but rather left them to the participants to decide.

Finally, raters filled out a follow-up questionnaire that inquires into the importance of different essay features they attached to (see Appendix E). They were asked whether they agreed or disagreed with a statement that the adequacy or inadequacy of Chinese cultural knowledge and Chinese language play a role in their rating process.

Analysis

Quantitative data.

The quantitative data consisted of the essay scores given by three groups of raters. These score assignments served as the primary variable to estimate any differences in score assignment among the rater groups. The research question was whether the three groups of raters would assign different scores to the essay written by the Chinese students. Repeated measures analysis of variance[1] (ANOVA) was used with the rater category as the between-subject variable and 16 Chinese students' essay scores as the repeated levels. This was to find if there was any significant difference among the groups. Post-hoc tests were conducted to find the source of group differences.

Qualitative data.

Two types of qualitative data existed in this study. First was the best and worst feature circled by each rater, and second was the raters' comment. I coded and analyzed the best and worst features to find out whether the three groups of raters approached the essay in the same way because even when the same score was given for an essay, the raters may have had different reasons for assigning those scores.

When analyzing the best and worst feature of an essay, I found that a few raters circled more than one best or worst features; for example, when the raters thought content and language were equally good for an essay, they would circle both features. When analyzing the data, I counted them all. In addition, some raters did not circle any best feature, because they thought nothing was good in the essay. Because of this, I added a "noun" category. While coding the data for each group, only when more than half of the raters agreed on an answer did I view it as a valid answer for that group. Otherwise, I assumed that no agreed-upon answer had been achieved and instead I marked the response as *n/a*.

To code the raters' comments concerning students' essays, I typed all the comments into a spreadsheet. To make the comments codable, I first segmented the comments for coding. For example, if one rater wrote "good organization and good language," I segmented it into two items. After segmentation, there were 960 comments total. To code the data, I borrowed a coding scheme from Shi (2001) (See Appendix F). Shi divided the comments from her study into five major categories: general, content, organization, language, and length. Under the major categories of content, organization, and language, Shi had further subcategories. I tried to code the comments into the subcategories, but changed some of them to fit this data better. In total, I used twelve categories in this study (see Figure 1). The categories were: (1) general (general); (2) content (general); (3) content (ideas); (4) content (arguments); (5) organization (general); (6) organization paragraphs; (7) organization (coherence); (8) language (general); (9) language (general); (10) language (intelligibility); (11) language (fluency); (12) fluency (fluency). Before coding all of the data, I first coded 25% of the data. A second rater also coded this 25% of the data. Inter-rater reliability was calculated at that point (after we had each coded 25% of the data). Using a percentage matching test, I calculated inter-rater reliability in this manner: if I and the second rater agreed both on the main category and on the subcategory for a comment, agreement was marked as 100%; but if I and the second rater only agreed on the main category and not the subcategory, then agreement was marked as 50%. Finally, I marked us as agreeing at only 25% if the subcategories matched but not the main categories. This was possible when comments were talking about something general, for example a coding like organization (general) and content (general). The initial inter-rater reliability was 70%. After that, we discussed the 30% of the data we did not code in the same manner. Though consensus, we reached agreements on the correct coding of those data. I then coded the rest of the data on my own.

As Shi (2001) did in coding the data, I used the key word, usually noun phrases to identify the categories of the comments. The adjectives in the comment usually denoted whether it was a positive or negative comment, and the content word gave the clue of comment categories. I first identified the positive or negative nature of the comment and then categorized it based on the key word.

After coding and organizing the comments into 12 categories, I calculated the simple statistics for the negative and positive comments for each group in general to get an overview of the leniency and harshness of the raters' comments. Next, I calculated the count for the comments in 12 categories in each rater group. Such analysis could reveal raters' scoring procedure as to which aspect of the essay they judged positively or negatively. If comparing the data among three rater groups, we could know the group difference on raters' positive or negative judgment on the 12 categories of essay aspects.

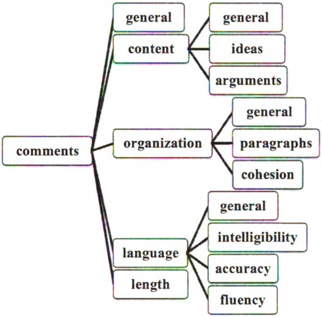

Figure 1. Categories of raters' comments

CHAPTER 4

Findings and Results

Reliability

I conducted a test of inter-rater reliability to ensure that all rater groups were rating reliably. If reliability was high enough, the differences between the groups' score assignments would be not related to unreliable raters. I used Cronbach's alpha to estimate reliability. The reliability for raters in total is .784, and the rate for American raters without cultural experience is .784, for American raters with cultural experience is .829, and for Chinese raters is .787. The degree of reliability for the three groups of raters was almost at the same level. This indicates that the raters were very similar in terms of reliability.

Quantitative Data

One way, [1]repeated measures ANOVAs with rater category as the between subject variable were calculated to measure the differences among the scores given by the three groups of raters. Table 2 shows that the scores were significantly different among the three groups of raters (df (2, 15) = 12.707, F = 6.943, $p < .05$). As can be shown in Table 3 Post-hoc comparisons using the Scheffe test indicated that the mean scores were significantly different between American raters without cultural experience and the Chinese raters (M = 0.5, SD = 0.164, p < .05). Significant differences were also found between means of the American raters with cultural experience and the Chinese raters (M = 0.53, SD = 0.164, p < .05). There was no significant difference found between the two American rater groups (M = 0.03, SD = 0.160, p = .983).

Table 2

Tests of Between-Subjects Effects

Source	Sum of Squares	df	Mean Square	F	Sig.
Intercept	2970.531	1	2970.531	1623.114	.000*
Rater Category	25.415	2	12.707	6.943	.004*
Error	43.923	24	1.830		

Table 3

Post hoc Test for Multiple Comparisons

						95% Confidence Interval	
	Rater Category	Rater Category	Mean Differenc e	Std. Error	Sig.	Lower Bound	Upper Bound
Scheffe	ARNCE	ARCE	-.03	.160	.983	-.45	.39
		CR	.50*	.164	.021*	.07	.93
	ARCE	ARNCE	.03	.160	.983	-.39	.45
		CR	.53*	.155	.009*	.12	.93
	CR	ARNCE	-.50*	.164	.021*	-.93	-.07
		ARCE	-.53*	.155	.009*	-.93	-.12

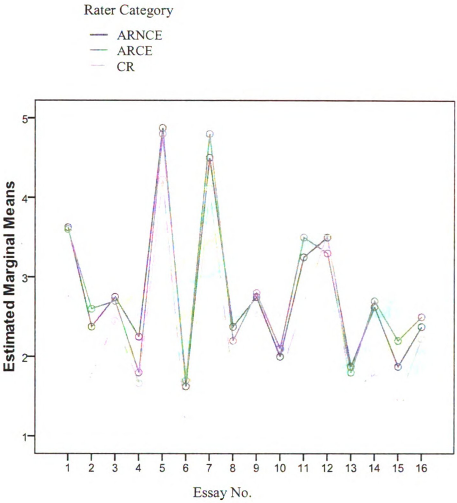

Figure 2. Estimated marginal means

As can be seen from the multi comparisons, post hoc test results illustrated in Table 3, raters who were Chinese ESL/EFL teachers scored significantly differently from the two groups of American raters. Figure 1 demonstrates that the Chinese raters were harsher than the other groups of raters in scoring Chinese students' essays because the Chinese raters tended to give a lower score than American raters with no cultural experience (*M* = 0.5, *SD* = 0.164, $p < 0.05$) and American raters with cultural experience (*M* = 0.53, *SD* = 0.164, $p < 0.05$). Among the American rater groups, there were not any significant differences in score assignments, and Table 4 illustrates that the American raters with cultural experience made score assignments that were slightly higher than the score assignments from American raters with no cultural experience (*M* = .03, *SD* = .160, *p* = .983).

Qualitative Data

Best and worst features.

All the raters circled the best and worst feature for each essay after assigning a score. The features of the writing were content, organization, coherence and grammar. Table 5 showed the most frequently circled features by each group of raters. For the best feature, there were three essays for which all groups of raters had the same perception. For example, essay No. 1 was best for organization, No. 3 for content and No.9 for organization. For the worst feature, the raters in different groups agreed on four essays, and all because of the essay writers' poor language. Still for most of the essays, raters across all groups had different perceptions concerning the best and worst features. Given for that, it can be reasoned that raters in the three groups did not approach the essay from the same aspect, or in other words, the raters had scored very differently because they had diverse perceptions of the best or worst feature for each essay.

Although it varied a great deal, the data showed that content and organization were judged as much better than other aspects of the essays by all raters, and language was perceived as the worst features by all raters. When looking further into different rater groups, American raters with no cultural experience chose content as the best feature for nine essays out of sixteen. This number was higher than the other two groups, which might suggest that American raters with no cultural experience generally valued content more positively than the other raters. It was obvious that

Chinese raters perceived grammar as the worst feature for ten essays out of sixteen. This may suggest that Chinese raters were stricter with language of the essay than other raters. No clear tendency was found for American raters with cultural experience, since they had too diverse opinions.

Table 4

Best Features and Worst Features Identified by Raters

Essay No.	Best features			Worst features		
	ARNCE	ARCE	CR	ARNCE	ARCE	CR
1	Orga	Orga	Orga	Lang	Lang	Lang
2	Orga	Cont	-	Lang	Lang	Lang
3	Cont	Cont	Cont	Lang	-	Lang
4	Lang	-	Cont	Cont	-	Lang
5	Lang	Cont	Cont	Cohe	Lang	Lang
6	Cont	-	None	-	-	Cont
7	Cont	-	Cont	Orga	Orga	-
8	Cont	Orga	None	Cohe	Cont	Cont
9	Orga	Orga	Orga	Lang	Lang	Lang
10	Lang	Orga	0	Cont	Cont	-
11	Cont	Cont	Orga	Lang	Cohe	Lang
12	Cont	Cont	Orga	Lang	Lang	Lang
13	Orga	Orga	Cont	Cont	Cohe	Cont
14	Cont	-	None	-	Cont	Lang
15	Cont	-	Cont; Lang	-	Orga	Lang
16	Cont	-	-	-	Lang	-

Note. The codes are used to label four features of the essay. Cont = *content*; Orga = *organization*; Cohe = *cohesion*; Lang = *language*. Dashes indicate that no agreed answers by that rater group.

Rater's comments.

The raters' comments were coded into negative or positive comments first, and then were put into 12 categories according to the coding scheme developed in Shi's (2001) study. The total number of comments coded for American raters with no cultural experience was 328, for American raters with cultural experience was 393, and for Chinese raters was 342. Table 6 shows that raters across the three groups generally provided more negative comments than positive comments, probably due to the low proficiency of essay writers. Overall, American raters with cultural experience had provided the largest amount of positive comments. On the contrary, Chinese raters had the most negative comments in terms of both number and frequency. But still, two groups of American raters were very close in their proportion of giving positive or negative comments. American raters with cultural experience commented a little more positively than American raters with no cultural experience, indicated by the percentage of 40% compared to 36%. It is important to remember that in the quantitative analysis, there were not any significant differences between the two groups of American raters. From an overview of the number and frequency of the comments for these two groups, the number was too close to demonstrate that American raters with cultural experience may judge the essays more positively than those with no cultural experience. More in-depth analysis of the nature of the comments thus is required.

Figure 2 shows the count of positive comments coded into 12 categories. My focus was still trying to find out the differences between American raters with and without cultural experience, if there were any. Figure 1 shows American raters with experience gave about two times more comments than the other raters on general aspect of the writing, 24 items compared to 13 and 14 in other groups. This might be indicating that American raters with cultural experience were slightly more satisfied with the essay writers in this level than other groups of raters, but more evidence is needed, especially negative comments from American raters with cultural experience should be taken into account. Regarding the comments about language, American raters with cultural experience offered much more comments on the general quality and intelligibility of the language in the essays than other raters. However, American raters with cultural experience commented less positively on accuracy of language compared with other raters. A close observation of the Chinese raters shows that they offered many fewer positive comments on content (general) and content (ideas) for an essay than other raters. Raters across all the three groups provided almost the equal amount of comments on content (argument) and organization (general).

Figure 3 shows the number of negative comments coded into 12 categories. The first impression from the figure is that the negative comments from Chinese raters outnumbered the other groups in several categories, for example, general (general), content (arguments) and especially language (accuracy) where the number almost doubled in size from the other groups. American raters with cultural experience provided many more negative comments on content (general), language (general) and language (intelligibility), the same as they did when providing positive comments. Now it is clear that American raters with cultural experience tended to offer more comments in both positive and negative way in such categories as content (general), language (general) and language (intelligibility), thus there still lacks evidence to prove that American raters with cultural experience valued more positively than those without experience in terms of content and other aspects of the writing. However, American raters with cultural experience commented less negatively on language (accuracy) than the other groups, as indicated by 26 comments, compared to 38 by American raters without cultural experience, and 64 for Chinese raters.

Table 5

An Overview of Raters' Comments

Comments	ARNCE		ARCE		CR	
	(N = 8)		(N = 8)		(N = 9)	
	Count	Frequency	Count	Frequency	Count	Frequency
Positive	119	36%	157	40%	100	29%
Negative	209	64%	236	60%	242	71%
Total	328		393		342	

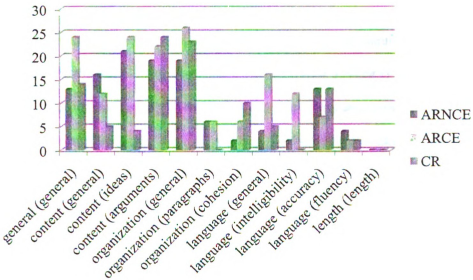

Figure 3. Raters' positive comments

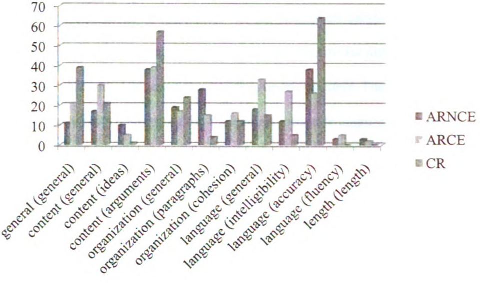

Figure 4. Raters' negative comments

CHAPTER 5

Discussion

In the area of second language writing assessment, a large number of studies have focused on the study of rater variability that may cause a potential source of measurement error. Human raters, used in scoring second language writing, may often introduce subjective factors into the procedure of scoring. Raters' backgrounds or experience could influence their ways in approaching the essays. All of these factors may lead to rater variability, which may pose test fairness issues and reduce the test's reliability and validity. Johnson and Lim (2009) claim that it is of great importance to identify the sources of rater variability, and "if found to exist, to address them accordingly" (p. 486). The current study manipulates the amount of raters' cultural experience to study its influence on the rating procedure. This study will promote our understanding of the possible source of rater variability and will shed light on teacher's different perspectives of evaluating writing by second language learners.

It was found in this study that the Chinese raters were harsher than the other two groups of American raters, given the statistically significant differences between the ratings given by the Chinese raters and the American raters. The Chinese raters were particularly strict concerning the L2 learners' accurate language and grammar, evidenced by their enormous negative comments on the language aspect of the essays. The result is not consistent with a previous study. Shi (2001) did not find any significant differences in the score assignments between the two groups of raters, NES (native-English-speakers) raters and NNS (non-native-English-speakers) raters. Probably this is due to the two studies' different experiment treatments. For example, Shi did not have rater training sessions or use any rating rubrics when raters gave evaluations to the students' essays. On the other hand, this study involved rater training and rating rubrics. More importantly, Shi asked the raters to rank the 10 students' essays in an order of 1 to 10. However, raters in the current study had a 5 point scale (1-5) in assigning scores.

The result for the qualitative part is also inconsistent with the Shi's (2001) study. The major difference is that Shi found NNS raters comment much less on the accuracy of the language for the writing because "being nonnative speakers, the Chinese teachers might shy away from making qualitative judgments or comments about the English language of their students" (p. 312). However, the current study found that Chinese raters gave more negative comments than the two groups of American raters on the language, especially on the accuracy of the language.

The fact that Chinese raters in this study assigned lower scores than American raters did probably is because the raters' past teaching and learning experiences play a role in the scoring procedure. More than half of the Chinese raters have taught college

English in China, and the proficiency of the students they taught might have been higher than that of the students taking courses in the English Language Center who just graduated from high school, those who participated in this study. It can be argued that the Chinese raters have much more experience in rating essays on a different level; they might judge lower level students more negatively than they should when assigning scores. Their enormous negative comments on language accuracy perhaps are related to how grammar accuracy is treated in English education in China. Grammar learning is emphasized heavily when a student learns English, and thus English teachers in China attach great importance to the grammar accuracy accordingly. As Eckes (2008) found in his study, experienced raters may fall into different categories that value categories within rating rubrics to different degrees. Shaefer (2008) also found a similar rater bias pattern in which one group of raters rated content and organization more severely, while the other group rated language and mechanics more severely.

However, in terms of language assessment, the raters who weigh the language more importantly than other aspects of the language may decrease the validity of the test because they should approach an essay with balanced attention to each aspect of the writing (Weigle, 2002). The holistic scores commonly used in writing assessment may increase the possibility of overemphasizing one aspect of the writing, especially for raters who have the tendency to overemphasize one aspect of the criteria being evaluated. The current study did not find any significant differences between the two groups of American raters with and without cultural experience in China. But American raters with experience have a trend to be more lenient than the raters without experience. The quantitative analysis identified a 0.03 higher score (the mean differences) given by the raters with experience than the raters without experience.

The frequency of giving positive comments in general is also slightly higher for raters with experience, for example, 40% for raters with experience, compared to 36% for raters without experience. The result is backed up by a question in the questionnaire that asked whether the raters think their cultural experience in China plays a role in the rating procedure. About four American raters with experience in China said that they thought their experience had an effect on their evaluation of the essays. The arguments they provided were various. For example, one rater said that "I probably gave the essays a higher rating if they seemed more truthful to my idea of how globalization has affected Chinese culture." Another rater responded, "I recognize some thoughts that may be poorly supported by heresy (heresay) and also have an easier time understanding the writer's perspective." One rater even thought that since she was aware that her knowledge about Chinese language and culture might make her give higher scores, she had to compensate by judging the essays too strictly. Within the raters who said "No" to this question, two raters stated that they could find some common and repeated "Chinglish forms" (expressions that were mistranslated from Chinese to English). However, they were not certain whether they would be biased toward the essay writers.

Clearly, American raters who taught English in China had more possibilities to refer to their common knowledge of Chinese culture and their teaching experience when rating the essays than their peers who never went to China. They adopted various strategies when comprehending and judging students' essays, for example, their cultural experience in China, their linguistic knowledge about the Chinese language, and their experience with reading lots of "Chinglish" errors. Some of them were even aware of an influence on the rating and adapted some compensation strategies to overcome the perceived problems. The current study is consistent with

Hamp-Lyons's study (1989) in which Hamp-Lyons found that experience with other language could change native speaker raters' reactions to the English writing of learners from those language communities. Despite their thoughts while rating essays seemed exceptionally complicated, there is not enough evidence to prove that the raters with experience in China are more lenient raters regarding the marking of features of essays. The result of this study shows that this area is in need of further inquiry.

This study has implications for rater training. The inter-rater reliability is reasonably high even though the ESL/EFL teachers had diverse backgrounds in terms of teaching and rating experience. An easy-to-use rating rubric is essential to the rating tasks because all raters in the study referred to the rubric time and time again to clearly distinguish between a "2-point-essay" or a "3-point-essay." However, in other situations, some raters shared their experience that they were quite ambivalent in deciding a final score, because "the language is high 2" but "the content is low 3." As Hamp-Lyons (1991) pointed out, second language writers may have developed unbalanced skills in different aspects of writing. The holistic scale again fails to represent the multi-trait features of an essay. In rater training sessions, it is important to remind the raters they will expect to have such undetermined moments and they should look further in the essay for other evidence to assign a score and not be arbitrary; not something between "2" and "3" but a "2" or "3."

CHAPTER 6

Conclusion

The purpose of this study is to understand whether raters' cultural influence may affect the score assigned to students' essays. The raters in this study were divided into three groups: American ESL/EFL teachers without teaching experience in China: American ESL/EFL teachers with teaching experience in China; and Chinese ESL/EFL teachers. All raters received rater training and performed rating tasks that included assigning a holistic score (1 to 5) to essays. They also judged the best and worst feature of the essays, and offered comments. The findings are that there were significant differences in the scores assigned by the Chinese raters and the two groups of American raters. Chinese raters were harsh raters and gave lower scores than the other raters. American raters with teaching experience in China had a trend of being more lenient raters than American raters without teaching experience in China; American raters with experience assigned slightly higher scores and offered more frequent, positive comments than American raters without experience. This study is important because it discovers an area of rater variability that has not been addressed in the literature. A trend was found in this study that raters with extensive cultural experience with the target culture (the one being learned by the test takers) are more lenient raters than those who do not have such experience. Clearly, the results from this study call for future inquiry in this area.

Considering raters' cultural experience and investigating its effect on the rating procedure should be part of any large-scale testing program's rating procedure. A larger, more finely-tuned experimental study on these aspects of the rating process should be conducted in the future. In the current study, the majority of the raters with teaching experience in China stayed in China for one or two years. If the raters had spent more than five years teaching in China, the results may have been quite different. On the other hand, when recruiting rater participants, researchers in the future should find raters with similar professional backgrounds. For example, to truly investigate the influence of cultural background, the amount and quality of the raters' teaching and rating experience should be the same; likewise, their academic backgrounds and current employment should be comparable. Controlling these factors would make the results more robust and generalizable. In addition, the researcher is not sure if Chinese raters are harsher on Chinese essay writers or if they are just harsher raters overall. A comparison group of non- Chinese raters of comparable size would be needed to test for that in the future studies. Nonetheless, this study, despite its weaknesses, demonstrated that cultural background plays a role and rater training programs will need to address it.

Note:

[1] Repeated measures analysis of variance: It is used to analyze the score differences among the three groups in the study, because there is a between-subject variable (the cultural experience), and a within-subject variable (19 essays). The raters are independent, but the measurement across the essays is not since the same people were measured 19 times (Tabachnick & Fidell, 2007).

APPENDIX A

Questionnaire of Demographic Information

1. ***Basic information***

   Email________________________________

   Gender (circle one): Male / Female   Age (in years):_________

   Job position (if applicable):___________________________

   Major in undergraduate study:_______________________________________

   Major in graduate study (if applicable):__________________________________

2. ***Language background***

   What is your native language? (If you grew up with more than one language, please specify.)

   ____

   Can you speak another language or other languages? What is it or what are they?_

3. ***Teaching and Rating experience:***

   a) How many years have you taught English as a second or foreign language?___

   b) Could you list all institutions where you have worked?

   ____

   ____

   ____

   ____

   c) Have you been an instructor for any writing courses? If yes, what was the level of the class and for how long did you teach it?

   ____

   ____

   ____

   d) Do you have experience in rating essays written by ESL/EFL learners? (Please rate your experience in choosing one description below "Yes".)

   Yes________   No_________

i. I am new at it.

ii. I've got some experience.

iii. I am very experienced.

e) Are you familiar with rating rubrics or scales for TOEFL writing? (If yes, please rate your experience in choosing one description below "Yes".)

Yes________ No_________

i. I am new at it.

ii. I've got some experience.

iii. I am very experienced.

f) Have you used TOEFL rating rubrics or scales in the writing courses you taught?

4. ***Cultural knowledge:***

a) Have you ever been to China? If yes, can you make a list of when and for how long you travelled there?

____

____

____

____

b) Can you describe the degree of your knowledge of Chinese culture by choosing one statement below?

i. I know nothing about Chinese culture.

ii. I know just a few things about Chinese culture.

iii. I know enough knowledge of Chinese culture that I am confident to work or study
there without trouble.

iv. I know Chinese culture almost at the same level as a native Chinese speaker does.

c) Can you name all Chinese traditional festivals you know and on which day they are? You will use Chinese lunar calendar if you can.

____

____

____

____

d) Can you speak Chinese? If yes, can you rank your Chinese language ability?

	Reading proficiency	Writing proficiency	Speaking fluency	Listening ability
Very poor				
Fair				
Good				
Very good				
Native-like				

e) Where did you learn Chinese and what was highest level of Chinese course you took?

____

____

____

____

APPENDIX B

Writing Prompt

The Writing Task:

Directions: You will have 40 minutes to write an essay about the following topic. The length of the essay should be at least 300 words in order to be an effective essay.

Do you agree or disagree with the following statement? "Globalization is impacting the culture of some countries in negative ways." Use specific details and reasons to support your answer.

APPENDIX C

Scoring Sheet

Now, you will begin the official rating section. You will rate 19 essays consecutively. It is a bit different from rating the sample essays. Here are three steps that you will do:

1. You will assign a score to each essay.
2. You will provide a two-to-three sentence comments about the essay.
3. You will circle the best and worst feature of each essay.

Please put all your response, including the essay score, your comments and chose of the best and worst feature of each essay on the score sheet, NEVER on the essays. You can refer to the writing rubrics anytime you want during your rating procedure. Now, you can begin. When you finish rating, please tell the researcher you are done.

Essay NO.	Score	Best Feature	Worst Feature	Comments
		Content	Content	
		Organization	Organization	
		Coherence	Coherence	
		Language	Language	
		Content	Content	
		Organization	Organization	
		Coherence	Coherence	
		Language	Language	
		Content	Content	
		Organization	Organization	
		Coherence	Coherence	
		Language	Language	
		Content	Content	
		Organization	Organization	
		Coherence	Coherence	
		Language	Language	
		Content	Content	
		Organization	Organization	
		Coherence	Coherence	
		Language	Language	
		Content	Content	
		Organization	Organization	
		Coherence	Coherence	
		Language	Language	

APPENDIX D

TOEFL Independent Writing Rating Rubrics

Score	Task Description
5	An essay at this level largely accomplishes all of the following: • effectively addresses the topic and task. • is well organized and well developed, using clearly appropriate explanations, exemplifications, and/ or details. • displays unity, progression, and coherence. • displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors.
4	An essay at this level largely accomplishes all of the following: • addresses the topic and task, though some points may not be fully elaborated. • is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/ or details. • displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections. • displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning.
3	An essay at this level largely accomplishes all of the following: • addresses the topic and task using somewhat developed explanations, exemplifications, and/ or details. • displays unity, progression, and coherence, though connections of ideas may be occasionally obscured. • may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning. • may display accurate but limited range of syntactic structures and vocabulary.

2	An essay at this level largely accomplishes all of the following: • limited development in response to the topic and task. • inadequate organization or connection of ideas. • inappropriate or insufficient explanations, exemplifications, or details to support or illustrate generalizations in response to the task. • a noticeably inappropriate choice of words, or word forms. • an accumulation of errors in sentence structure and/ or usage.
1	An essay at this level largely accomplishes all of the following: • serious disorganization or underdevelopment. • little or no detail, or irrelevant specifics, or questionable responsiveness to the task. • serious and frequent errors in sentence structure or usage.

APPENDIX E

Follow-up Questionnaire of Rating Behavior

a. Could you order the importance of each aspect of writing when you rate an essay? (for example, 1-most important and 4-least important)

A. Grammar and Language use

B. Content

C. Organization and Development

D. Coherence and Unity

1. ______________________
2. ______________________
3. ______________________
4. ______________________

b. Do you think your adequacy or lack of Chinese cultural knowledge play a role in the way you rate the essays?

c. If you have any idea after rating the essays, you can write it here.

APPENDIX F

Coding Scheme

MAJOR CATEGORIES	SUB-CATEGORIES	DEFINITIONS	EXAMPLES OF POSITIVE/NEGATIVE COMMENTS
General	General	General comments on overall quality of writing.	• This is an excellent essay (ARNCE-2). • This is very well written (ARNCE-6).
Content	General	General comments on content	• The writer makes good point (ARCE-4). • The essay lacks enough content to prove the author's points (ARCE-4).
	Ideas	General or specific comments on ideas and thesis.	• The ideas in this article were very clear (ARCE-5). • The writer had great ideas (ARCE-5). • Great clear statements (ARCE-8).
	Arguments	General or specific comments on aspects of arguments such as balance, use of comparison, counter arguments, support, uses of details or examples, clarity, unity, maturity, originality, relevance, logic, depth, objectivity, conciseness, development and expression developed.	• Use examples to address the topic (CR-8). •Developed the topic to some extend but not enough (CR-7). • Has some good examples (ARCE-3). • Great examples (ARNCE-5) • Inadequate development (ARNCE-6). • Good arguments were made (ARCE-5). • Very little examples (ARCE-9).
Organization	General	General comments on organization	• Well organized (ARCE-9). • It is organized well (CR-4).
	Paragraphs	Comments on the macro level concerning paragraphs	• But they should be broken down into subpoints and paragraphs (ARCE-10). • Some paragraphs seem oddly

		introductions, and conclusions.	short (ARCE-3). • Needs to know how to group paragraphs together (ARNCE-7). • This could be slightly more effective if some of the short paragraph were combined or expanded (ARNCE-2).
	Transitions	Comments on the micro level concerning transitions, coherence, and cohesion.	• Coherence is a problem (ARNCE-5). • Some problems with cohesion and connectives (ARNCE-6). • Cohesive devices not effective (ARCE-3).
Language	General	General comments on language	• Language is poor (ARCE-8). • The use of language was not very good (ARCE-5).
	Intelligibility	Comments on whether the language is clear or easy to understand.	• English usage is sometimes confusing (ARCE-9). • The language usage at times the writer's opinion was initially unclear (ARCE-5).
	Accuracy	General comments on accuracy or specific comments on word use, grammar and mechanics.	• Serious and frequent errors in sentence structure or usage (CR-4). • Serious errors in usage (CR-5). • There is almost no error in sentence structure (CR-4). • Many language errors in grammar and syntax (ARCE-9).
	Fluency	Comments on fluency, conciseness, maturity, naturalness, appropriateness, and vividness of language.	• Too many words, he could've said it more concisely (ARCE-7). • Sentence variety, command of complex vocabulary and structures (ARCE-3). • Sometimes redundant (ARCE-9). • Demonstrate range of vocabulary (CR-2).
Length	Length	Comments on whether the writer has fulfilled the word limit.	• It was quite short (ARNCE-2).

BIBLIOGRAPHY

Ballard, B. & Clanchy, J. (1991). Assessment by misconception: Cultural influences and intellectual traditions. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (p. 241-278). Norwood, NJ: Ablex.

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

Broad, B. (2003). *What we really value*: 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.

Connor, U. (1996). Contrastive rhetoric: Cross-cultural aspects of second language writing. New York: Cambridge University Press.

Connor-Linton, J. (1995). Looking behind the curtain: what do L2 composition ratings really mean? *TESOL Quarterly 29*, 762–65.

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language Testing and Assessment* (pp. 51–63). Dordrecht, Netherlands: Kluwer.

Cummings, A. (1990). Expertise in evaluating second language compositions. *Language Testing 7*, 31–51.

Cummings, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, 86, 67–96.

Deville, G., & Chalhoub-Deville, M. (2006). Old and new thoughts on test score variability: Implications for reliability and validity. In M. Chalhoub-Deville, C.A. Chapelle & P. Duff (Eds.), *Inference and generalizability in applied linguistics: Multiple perspectives* (pp. 9–25). Amsterdam: Benjamins.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*, 155-185.

Educational Testing Service. (2004). *iBT/Next generation of TOEFL Test independent writing rating rubrics (Scoring standards).* Princeton, NJ: Educational Testing Service.

Educational Testing Service. (2005). *TOEFL iBT writing sample responses, Princeton.* NJ: Educational Testing Service.

Erdosy, U. (2004). *Exploring Variability in Judging Writing Ability in a Second Language: A Study of Four Experienced Raters of ESL Compositions.* (TOEFL

Report 70). Princeton, NJ: Educational Testing Service.

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

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (p. 241-278). Norwood, NJ: Ablex.

Hinkel, E. (1994). Native and nonnative speakers' pragmatic interpretations of English text. TESOL Quarterly, 28, 353-376.

Johnson, J. & Lim, G. (2009). The influence of rater language background on writing performance assessment. *Language Testing, 26,* 485-505.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? Language Testing, 19, 246–76.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*, 275-304.

Kondo-Brown, K. (2002). A facets analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19,* 3-31.

Kobayashi, H. & Rinnert, C. (1996). Factors affecting composition evaluation in an EFL context: cultural rhetorical pattern and readers' background. *Language Learning 46,* 397–437.

Kobayashi, T. (1992). Native and nonnative reactions to ESL compositions. *TESOL Quarterly 26,* 81–112.

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

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

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*, 465-493.

Schoonen, R. (2005). Generalizability of writing scores: an application of structural equation model. *Language Testing, 22*, 1-30.

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

Tabachnick, B. & Fidell, L. (2007). *Using multivariate statistics (5th ed).* Toronto: Allyn & Brown.

Wang, C. (2009). Chinese test-takers' perspectives on GRE and TOEFL writing tests. Paper presented at the Michigan TESOL conference, Grand Rapids, Michigan.

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

Weigle, S. C., Boldt, H. & Valsecchi, M. I. (2003). Effects of tasks and rater background on the evaluation of ESL student writing: A pilot study. *TESOL Quarterly, 37(2)*, 345-354.