Language and Communication Quarterly, 1: 2-24

http://www.untestedideas.com/lcq.html
ISSN: 2168-7633 (Print)
ISSN: 2168-7641 (Online)

ORIGINAL RESEARCH ARTICLE

Examining the Impact of Rater Educational Background on ESL Writing Assessment: A Generalizability Theory Approach

Ramy Elorbany
Graduate of Niagara University

Jinyan Huang*
Niagara University

Using generalizability theory as a theoretical framework, this study investigated the impact of raters' educational background on the assessment of K-12 ESL students' writing. Twenty teacher candidates (ten TESOL majors and ten non-TESOL majors) from universities in western New York and southern Ontario participated in this study. The 20 participants were asked to rate three ESL essays holistically on a 1-10 point scale (1 being the lowest and 10 being the highest with permission to use half points). The results indicate that raters' TESOL-related educational background did impact their rating of ESL essays. The TESOL teacher candidates marked the three ESL essays more consistently and reliably than their non-TESOL counterparts. Important implications are discussed.

In North America, the number of immigrants has been on the rise since the 1970's (Statistics Canada, 2011; U.S. Department of Homeland Security, 2010). The increase of the immigrant population and subsequently the increasing number of

*Correspondence should be sent to: Dr. Jinyan Huang, 325 F Academic Complex, College of Education, Niagara University, Niagara University, NY 14109. Email: *jhuang@niagara.edu*

English-as-a-second-language (ESL*) students in North American K-12 schools would have major implications for both the Canadian and American education systems. In accordance with the values which both Canada and the United States share, each ESL student must be provided with a fair and equal opportunity to learn (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999).

Providing a fair and equal opportunity for K-12 ESL students is not a simple task. A fair and equal education must be built on valid and reliable assessments (AERA, APA, NCME, 1999). However, research has started to show that there exist concerns about the reliability, validity, and ultimately, fairness of the assessment of K-12 ESL students' learning, the assessment of their writing, in particular (Huang, 2008, 2012). The assessment of ESL writing becomes problematic due to the following two major reasons. First, many factors affect ESL students' writing in English; for example, their English language proficiency, their writing proficiency, the impact of their first language on their ability to write, their home culture, etc. (Gamaroff, 2000; Huang, 2008; Ruetten, 1994; Sweedler-Brown, 1993; Vann, Meyer, & Lorenz, 1984). Second, raters may take these factors into consideration differently, causing more variability and less reliability in the assessment of ESL students' writing than the assessment of the native English-speaking students' writing (Connor-Linton, 1995; Huang, 2008, 2009, 2012; Sweedler-Brown, 1993).

These problems consequently lead to fairness concerns about the assessment of K-12 ESL students' writing (Huang, 2008; 2012). Further, research has shown that K-12 ESL students experience considerable challenges in meeting course expectations due to their linguistic and cultural differences (Barkaoui, 2010; Huang, Cunningham, & Finn, 2010; Huang, Smith, & Smith, 2011; Ruetten, 1994). These differences can prevent ESL students from communicating clearly with their teachers and from meeting expectations that are commonly understood by their native English-speaking peers within North American K-12 schools (Huang et al., 2010; Huang et al., 2011; Jenkins, 2000; Johnson, Penny, Gordon, Shumate, & Fisher, 2005; Trice, 2001; Wiggins, 1993), Thus exacerbating the fairness concerns about the assessment of their writing.

Fairness is an essential pillar of education; a pillar which is upheld and strongly supported by North American values. Educational organizations, institutions, and individual professionals should make assessments as fair as possible for test takers of different races, genders, and ethnic backgrounds (AERA, APA, NCME, 1999; Joint Advisory Committee, 1993). Failure to provide fair assessment for ESL students is an infringement on their right to receive an equal opportunity to learn (Huang, 2008; Huang & Foote, 2010; Johnson et al., 2005; Popham, 2011).

---

*ESL students refer to students who enter North American K-12 schools with little or no previous knowledge of English and have received education in the language of their home country. They can also be Canadian- or American-born students who are from homes and/or communities in which English is not widely used and who therefore have limited proficiency in English.

THE IMPACT OF RATER CHARACTERISTICS ON ESL WRITING ASSESSMENT

There are a number of factors that affect the assessment of ESL writing and these factors can be categorized into three major categories: student-related, task-related, and rater-related. Of the three categories, rater-related factors are the most precarious in efforts to achieve fairness in assessment. This is because rater-related factors may jeopardize the reliability and in turn validity and fairness of writing assessment (Gamaroff, 2000; Vann et al., 1984; Johnson et al, 2005; Huang & Foote, 2010; Huang, 2008, 2009, 2012).

The impact of rater characteristics on rating has been explored for decades and by a number of researchers in the field (e.g., Brown, 1991; Huang, 2009; Kobayashi, 1992; Sakyi, 2000; Santos, 1988; Weigle, 1994). Huang (2009) reviewed 20 empirical studies and identified a number of rater characteristics that impact ESL writing assessment. These rater-related factors include "raters' academic disciplines, professional experiences, linguistic backgrounds, tolerance for error, perceptions and expectations, and rater training" (Huang, 2009, p. 4). The following section is a brief summary of the literature.

Raters' Academic Discipline

Many studies indicate that raters' academic discipline impacts their rating of ESL students' writing (Brown, 1991; Santos, 1988; Song & Caruso, 1996; Vann et al., 1984; Weigle, Boldt, & Valsecchi, 2003). For example, Mendelsohn and Cumming (1987) conducted a study in which they compared engineering professors with ESL professors in terms of differences in using criteria when they marked ESL papers. They found that engineering professors attributed more importance to language use than to rhetorical organization in rating the effectiveness of ESL papers; whereas ESL professors attributed more importance to rhetorical organization.

Further, Santos (1988) examined 178 non-ESL professors' scoring of two ESL students' written compositions. Of the 178 professors, 96 were in the humanities/social science field and 82 were in the physical sciences field. Although no significant difference was found in terms of using the rating criteria (content and language) between the two groups of professors, the study revealed that the physical science professors were more severe raters than the humanities/social science professors.

Raters' Professional Experience

Raters' professional experience is another factor that can impact their rating of ESL students' writing (e.g., Barkaoui, 2010; Hamp-Lyons, 1996; Vaughan, 1991). Barkaoui (2010) investigated the rating differences between 11 novice and 14 experienced raters. Each rater was asked to rate 12 ESL essays both holistically and analytically. The findings indicate that with the holistic scale the "experienced raters referred more frequently to

rhetorical and ideational aspects (i.e., ideas, organization, development) than did the novices, but with the analytic scale, they referred to linguistic features (syntax and morphology) and text length more often" (Barkaoui, 2010, p.11). Furthermore, Song and Caruso (1996) reported that the more years of experience a rater had in teaching; the more lenient he or she would be in rating ESL essays holistically.

Raters' Linguistic Background

Interestingly, raters' linguistic background was found to impact their rating of ESL students' writing as well. Kobayashi (1992) compared 145 native English raters with 124 native Japanese raters in the rating of ESL compositions. The raters were all at the professional, graduate, or undergraduate level. Not only did the researcher discover that raters' academic rank was a factor that had affected their rating of ESL compositions, but the researcher also found that native English raters identified more errors in ESL compositions and were much severer with grammatical errors than native Japanese raters.

Raters' Perception and Expectations

A few researchers have examined the impact of raters' perception and expectations on the rating of ESL compositions (e.g., Casanave & Hubbard, 1992; Janapolous, 1995). Casanave and Hubbard (1992) conducted a graduate faculty survey study on the writing requirements of first year doctorate students including at an American university. The graduate faculty members were asked to provide information on the criteria they would use to evaluate ESL and native English-speaking students' writing, their writing problems as well as information regarding their expectations of the first year doctorate students. There were 85 participants in the study who represented 28 departments of the two fields of humanities/social sciences and science/technology. When asked to rank features of writing which would affect rating, raters from both fields had similar responses. Both groups considered discourse level criteria (e.g., quality of content, development of ideas, and adequate treatment of topic) to be of high importance while word- and sentence-level criteria (e.g., accuracy of grammar, size of vocabulary, and spelling and punctuation) to be of low importance. However, humanities/social science faculty considered all of these features to be more influential to their rating than did science/technology faculty. Furthermore, when asked to rank the importance of writing skills (on a scale of 1-5) at the three different stages of the doctorate program (first year, second year, all subsequent years), the humanities/social science faculty ranked the importance of writing skills higher than did their science/technology counterparts.

Raters' Tolerance for Error

Raters' tolerance for error has been a rater characteristic of interest to many researchers in the field of language assessment (e.g., Janopoulos, 1992; Santos, 1988; Sweedler-Brown, 1993; Vann et al., 1984). Sweedler-Brown (1993) conducted a study examining how grammatical and syntactic features would affect raters' holistic marking of ESL writing. Only two out of 18 essays with grammatical and syntactic errors received a passing grade. The same raters were asked to rate these essays after the errors had been corrected; and surprisingly, 17 out of 18 received a passing grade.

Similar studies show findings suggesting that raters from different academic backgrounds may place different weights on such errors as grammatical or sentence level errors (Janopoulos, 1992; Santos, 1988; Vann et al., 1984). Furthermore, a number of studies show that among those from the social science, education, humanities, biological and agricultural, physical and mathematical sciences, and engineering departments, professors from the social science department were the most tolerant of ESL writing errors (Janopoulos, 1992; Santos, 1988; Vann et al., 1984). However, Sweedler-Brown (1993) discovered that rater training could positively decrease the differences in tolerance for error among the raters.

Rater Training

Rater training is seen by many researchers as an important tool to minimize rater variation in the rating of ESL writing (e.g., Davidson, 1991; Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981; Weigle, 1994). Weigle (1994) conducted a study in which raters were required to rate ESL written compositions before and after rater training. The findings indicate that a desired change in raters could be achieved by training the raters. Similarly, Ruetten (1994) and Sweedler-Brown (1993) suggest that training be used to minimize rater variability and increase inter-rater and intra-rater reliability.

Most researchers indicate that rater training does have a positive impact on the rating of ESL writing. However, Charney (1984) and Gere (1980) believe that "holistic training procedures alter the process of scoring and reading and distort the raters' ability to make sound choices concerning writing ability" (Huot, 1990, p. 202).

To sum up, many studies have examined the impact of several rater-related factors on the assessment of ESL students' writing (Brown, 1991; Huang, 2009; Kobayashi, 1992; Sakyi, 2000; Santos, 1988; Weigle, 1994). Very few studies, however, have directly examined how raters' ESL-related educational background affects their rating of ESL writing. In other words, would raters with ESL-related educational background score ESL writing differently than those with no ESL-related educational background? This study was intended to bridge the research gap.

THEORETICAL FRAMEWORKS GUIDING WRITING ASSESSMENT RESEARCH

The three approaches to detecting rating variability in the field of performance assessment (e.g., writing) are multi-facet Rasch approach, the classic test theory (*CTT*) approach, and the generalizability (*G*-) theory approach. Although the multi-facet Rasch approach is also a viable alternative it was not considered within the scope of this study and has traditionally been used in more large-scale testing situations.

The *G*-theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) approach is more powerful than *CTT* for the detection of rater variability (Shavelson, Baxter, & Gao, 1993). *G*-theory extends the framework of *CTT* in order to take into account the multiple sources of variability that can have an effect on test scores (Shavelson & Webb, 1991).

In addition to having the capability of identifying multiple sources of variance, *G*-theory can also pin point how much each variable affects the true score. Furthermore, *G*-theory identifies not only multiple sources of error, but also the interaction of these sources of error (Shavelson & Webb, 1991).

G-theory has also been employed specifically in the field of ESL assessments (e.g., Huang, 2008; 2011, 2012; Huang & Foote, 2010). Using *G*-theory as a theoretical framework, Huang (2008) examined the rating variability and reliability of scores assigned to ESL essays and native English (NE) essays in large-scale standardized writing assessments in Canada. Later, using *G*-theory, Huang and Foote (2010) investigated the rater variation differences between ESL papers and NE papers in the context of classroom assessment at an American university. In both studies, the results show that the writing scores assigned to ESL essays were less consistent than the scores assigned to NE essays. These findings indicate that there are threats to the fairness of ESL writing assessment in both large-scale standardized and classroom assessment contexts.

Since *G*-theory provides a more powerful theoretical framework than *CTT* and it has been increasingly used in the research of second language writing assessments, it was used as the framework of this present study.

PURPOSE OF THE STUDY

Using *G*-theory as a theoretical framework, this study examined the effects of raters' educational background on the rating variability and reliability of K-12 ESL students' writing. In other words, it examined whether there were differences in rating variability and reliability between raters with ESL-related educational background (i.e., TESOL*-majored teacher candidates) and raters with no ESL-related educational background (i.e., non-TESOL majored teacher candidates).

---

*TESOL: Teaching-English-to-speakers-of-other-languages

Specifically, the following three research questions guided this study: 1) Are there significant differences between the writing scores assigned by raters with and without TESOL educational background? 2) What are the sources of score variation contributing relatively more to the variability of the scores assigned by raters with TESOL educational background in contrast to raters without TESOL educational backgrounds? And 3) does the reliability (e.g., generalizability coefficients for norm-referenced score interpretations and dependability coefficients for criterion-referenced score interpretations) of the scores assigned by raters with TESOL educational background differ from the reliability of the scores assigned by raters without TESOL educational background?

METHODOLOGY

The Selection of Writing Samples

The writing samples used for this study were obtained from three Grade 10 ESL students studying at a western New York high school. The students were asked to respond to the following writing task:

> *Write an essay about how you can get better at something you like to do and how you might share your talent with others. Give examples of people who might teach you how to develop your skills.*

The three ESL students were not asked to complete the writing task for the purpose of this study; however, permission was obtained from the necessary parties and ethical procedures were followed. The three ESL students were from different parts of the world and did not share the same native language. Further, it is important to note that the three essays were already scored by the classroom teachers, who were certified ESL teachers in the State of New York. The three selected ESL essays were considered to be of intermediate quality and were similar in length.

The Selection of Raters

All 20 raters participating in this study were selected through convenience sampling method from volunteer teacher candidates. Table 1 presents a summary of the 20 raters. These raters were either graduate or undergraduate teacher candidates studying at American or Canadian universities. It is important to note that these universities were located in either the State of New York or the Province of Ontario. The 20 participants included 9 Canadian citizens and 11 American citizens.

As shown in Table 1, among the 20 participants ten were TESOL major teacher candidates and the other ten were non-TESOL major teacher candidates. All participants were native speakers of English with no experience of formal teaching. Most of the participants were aged between 20 and 30. The 20 participants included 13 undergraduate

teacher candidates and seven graduate teacher candidates; four male teacher candidates and 16 female teacher candidates.

Table 1
A Summary of the 20 Raters

Rater Information		Rater Gender	
		Male	Female
Age	20-24	2	10
	25-29	2	5
	≥30		1
Level of Education	Undergraduate	1	12
	Graduate	3	4
Education Focus	TESOL	1	9
	Non-TESOL	3	7
Citizenship	American	1	10
	Canadian	3	6
	Total	4	16

The Rating Procedure

The rating rubric used in this study was based on well-accepted and agreed upon criteria (content, organization, and mechanics) in the field of ESL writing assessment (Huang & Foote, 2010). As soon as the 20 participants were informed of the study and had read and signed the consent forms, they were given a package containing the scoring rubric and the three ESL essays. A brief explanation of the rubric and the task at hand were provided by the principal researcher. Following the explanation were several minutes allotted for a 'question period' to ensure that all participants understood their tasks and what was expected of them. All 20 participants were then asked to rate the three ESL essays holistically on a 1-10 point scale (1 being the lowest and 10 being the highest with permission to use half points). They were required to score the essays individually to avoid discussion amongst the raters. This resulted in 3 papers (*p*) and 60 scores, each paper receiving twenty different scores from twenty different raters (*r*).

SPSS Statistical Analyses

Descriptive statistical analyses (the mean and standard deviation) as well as paired samples *t*-tests for the writing scores given by both rater groups (i.e., TESOL raters and non-TESOL raters) were conducted. These statistical analyses were conducted to compare the score means and standard deviations of the two rater groups and determine if there were significant differences between the scores given by TESOL raters in comparison to non-TESOL raters for each ESL essay.

The *G*-Theory Analyses

Using the *G*-theory framework, data were further analyzed in the following stages: 1) a person-by-rater (nested within experience) mixed effects *G*-study; 2) a person-by-rater random effects *G*-study for raters with TESOL background; 3) a person-by-rater random effects *G*-study for raters with no TESOL background; 4) the calculation of *G*-coefficients for norm-referenced score interpretations; and 5) the calculation of dependability coefficient for criterion-referenced score interpretations.

A person-by-rater (nested within experience) mixed effects G-study. A person-by-rater nested within experience (*p x r: e*) mixed effects *G*-study analysis was conducted. The purpose of this *G*-study was to obtain variance component estimates for the five sources of variation: person or paper (*p*), experience (e), rater nested within experience (*r: e*), person-by-experience (*p x e*), and person-by-rater nested within experience (*p x r: e*).

Person-by-rater random effects G-study for raters with TESOL background. A person-by-rater (*p x r*) random effects *G*-study for raters with TESOL background was conducted. The purpose of this *G*-study was to obtain variance component estimates for the following three sources of variation: person or paper (*p*), rater (*r*), and person-by-rater (*p x r*). These variance components were further used to calculate both the *G*-coefficients and dependability coefficients for the scores assigned by raters with TESOL background.

Person-by-rater random effects G-study for raters no TESOL background. Similar to the above analysis, a person-by-rater (*p x r*) random effects *G*-study for raters with no TESOL background was conducted.

The information obtained from the above *G*-analyses was used to compare teacher candidates with TESOL educational background and those with no TESOL educational background in terms of score variability. A difference between the two groups was expected as the literature indicated that rater professional experience may influence the rating of ESL writing.

The calculation of G-coefficients for norm-referenced score interpretations. A *G*-coefficient is the ratio of the universe score variance to itself plus relative error variance ($\mathrm{E}\rho^2 = \frac{\sigma^2_\rho}{\sigma^2_\rho + \sigma^2_\delta}$). Using this formula, the *G*-coefficients for norm-referenced score interpretations were calculated for each rater group (i.e., raters with TESOL educational background versus raters with no TESOL educational background). In norm-referenced test contexts, each examinee's score on the test is interpreted relative to the scores of all other examinees who took the test.

The calculation of dependability coefficients for criterion-referenced score interpretations. A dependability coefficient is the ratio of the universe score variance to itself plus absolute error variance ($\Phi = \frac{\sigma^2_\rho}{\sigma^2_\rho + \sigma^2_\Delta}$). Using this formula, the dependability

coefficients for criterion-referenced score interpretations were calculated for each rater group (i.e., raters with TESOL educational background versus raters with no TESOL educational background). In criterion-referenced test contexts, each examinee's score is interpreted relative to a performance standard.

Computer Programs

Microsoft EXCEL was used for data preparation. Further, SPSS was used to conduct both descriptive and inferential (i.e., paired samples *t*-tests) statistical analyses. SPSS is a popular data-analysis program used by researchers in social sciences. SPSS can be used for manipulating data, analyzing data, and generating graphs and tables.

In addition, the computer program GENOVA (Crick & Brennan, 1983) was used for the *G*-studies for the data set. GENOVA is a computer program used to estimate the variance components for the main and interaction effects as well as their standard errors where the design is balanced.

RESULTS

Descriptive Results

As previously mentioned each paper was holistically rated by 20 raters (ten with TESOL educational background and ten with no TESOL educational background) on a 1-10 point scale. Table 2 provides the descriptive statistical results (i.e., the mean and standard deviation of the scores of each paper assigned by the 20 raters).

Table 2
Descriptive Results

	TESOL		Non-TESOL	
Essay	M	SD	M	SD
1	8.20	.422	7.60	.937
2	7.00	.236	6.75	1.03
3	6.05	.497	5.10	.966

As shown in Table 2, all three papers received consistently lower scores from the non-TESOL raters than from the TESOL raters, suggesting that the non-TESOL raters were generally more severe than the TESOL raters. Table 2 also shows that the standard deviations for essays #1 and #3 scores assigned by the non-TESOL raters were approximately two times larger than the standard deviations of the scores assigned by the TESOL raters; for essay #2 scores, the standard deviation for the non-TESOL raters was almost five times larger than the standard deviation for the TESOL raters. All these results

suggest that the non-TESOL raters were considerably more variant and less consistent in scoring the three ESL essays than the TESOL raters.

Paired Samples *t*-Tests Results

The paired samples *t*-tests were conducted to compare the mean score differences between the TESOL raters and the non-TESOL raters. The results are presented in Table 3.

As shown in Table 3, essays #1 and #2 did not receive significant different scores from the TESOL raters than from the non-TESOL raters. However, essay #3 received a significantly higher score from the TESOL raters than from the non-TESOL raters ($p < .05$). Further, as reported earlier, the standard deviations for all three essay scores assigned by the non-TESOL raters were larger than those by the TESOL raters. These results suggest that raters' TESOL-related educational background did impact the rating of these ESL essays.

Table 3
Paired Samples t Tests Results

	Paired Differences					
	M	SD	SEM	*t*	*df*	Sig.
Pair 1 TESOL-NTESOL* (Essay #1)	.60	1.10	.348	1.724	9	.119
Pair 2 TESOL-NTESOL (Essay #2)	.25	1.03	.327	.764	9	.464
Pair 3 TESOL-NTESOL (Essay #3)	1.30	1.30	.411	2.310	9	.046**

*NTESOL: non-TESOL; **significant at the .05 level

G-Theory Analyses Results

The following section presents the *G*-analyses results. Specifically, the person-by-rater (nested within experience) mixed effects *G*-study results were first presented. The results for the person-by-rater random effects *G*-studies were then presented for raters with and without TESOL background, respectively. Finally, the *G*-coefficients and dependability coefficients for each rater group (i.e., raters with TESOL educational background versus raters with no TESOL educational background) were reported.

The person-by-rater (nested within experience) mixed effects G-study

The person-by-rater nested within experience *p x r: e* mixed effects *G*-study yielded the following five sources of variation: person or paper (*p*), experience (*e*), rater nested within experience (*r:e*), person-by-experience (*pe*), and person-by-rater nested within experience (*pr: e*). Table 4 presents the results.

The results presented in Table 4 show that the person (*p*), the object of measurement yielded the largest variance component (64.67% of the total variance), suggesting that the three selected ESL essays are very different in terms of quality. The residual (*pr:e*) yielded the second largest variance component (20.71% of the total

variance). The residual contains the variability due to the interaction between experience, raters, person, and other unexplained systematic and unsystematic sources of error. Educational background (*e*) yielded the third largest variance component (7.12% of the total variance), suggesting that there was a large difference in the writing scores that could be attributed to raters' educational background. Rater nested within experience (*r:e*) yielded the fourth largest variance component (6.59% of the total variance), suggesting that raters within educational background differed from one another in terms of rating severity or leniency. The person-by-education (*pe*) variance component was clearly the smallest variance component. It explained less than 1% of the total variance.

Table 4
Variance Components for Mixed Effects p x r:e G-Study Results

Source of Variability	df	σ^2	%
p	2	1.3271	64.67
e	1	0.1461	7.12
r:e	18	0.1352	6.59
pe	2	0.0188	0.92
pr:e	36	0.4250	20.71
Total	59	2.0522	100

The person-by-rater random effects G-study for raters with and without TESOL background

The person-by-rater (*p x r*) random effects *G*-study for raters with and without TESOL background yielded the following variance components for each rater group: person (*p*), rater (*r*), and person-by-rater (*pr*). Table 5 presents the results for these two *G*-studies.

Table 5
Variance Components for Random Effects p x r G-studies Design

Language Group	Source of Variability	df	σ^2	%
TESOL	*p*	2	1.1417	85.62
	r	9	0.0000	0.00
	pr	18	0.1917	14.38
	Total	29	1.3334	100
Non-TESOL	*p*	2	1.5500	61.75
	r	9	0.3019	12.03
	pr	18	0.6583	26.22
	Total	29	2.5102	100

The *p* x *r* random effects *G*-studies results for both TESOL and non-TESOL raters are presented in Table 5. As shown in Table 5, the results for the TESOL raters show that person (*p*), the object of measurement yielded the largest variance component (85.62% of the total variance), suggesting that the three ESL essays were extremely different in quality as marked by the TESOL raters. The residual yielded the second largest variance

(14.38% of the total variance). The residual contains the variability due to the interaction between raters and papers, and other unexplained systematic and unsystematic sources of error. However, the rater (*r*) variance component was 0, indicating that the TESOL raters rated the three ESL essays extremely consistently.

As shown in Table 5, the results for non-TESOL raters show that person (*p*), the object of measurement yielded the largest variance component (61.75% of the total variance), suggesting that the three ESL essays were very different in quality as marked by the non-TESOL raters. The residual yielded the second largest variance (26.22% of the total variance). Again, the residual contains the variability due to the interaction between raters and papers, and other unexplained systematic and unsystematic sources of error. However, the rater (*r*) variance component yielded the third largest variance component (12.03% of the total variance). This result indicates that non-TESOL raters differed considerably from one another in terms of leniency of marking the ESL essays.

The calculation of G-coefficients for norm-referenced score interpretations

Using the formula $\mathrm{E}\rho^2 = \dfrac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma_\delta^2}$, the *G*-coefficients for each rater group (i.e., raters with TESOL educational background versus raters without TESOL educational background) were calculated for norm-referenced score interpretations. The results are presented in Table 6.

As shown in Table 6, the *G*-coefficient obtained for non-TESOL raters for the current 10-rater scenario was .96, whereas the *G*-coefficient for the TESOL teacher candidates was .98. However, in classroom assessment context, where only one rater scores each student paper, the *G*-coefficient for the non-TESOL rater group was .70; in contrast, the *G*-coefficient for the TESOL rater group would be .86. Comparing the reliability of one rater with TESOL background to another rater with no TESOL background reveals a much larger difference in terms of reliability.

Figure 1 displays the differences in *G*-coefficients between TESOL and non-TESOL raters. The figure clearly displays that the TESOL raters are less variant and more consistent and reliable than the non-TESOL raters in the context of norm-referenced score interpretations.

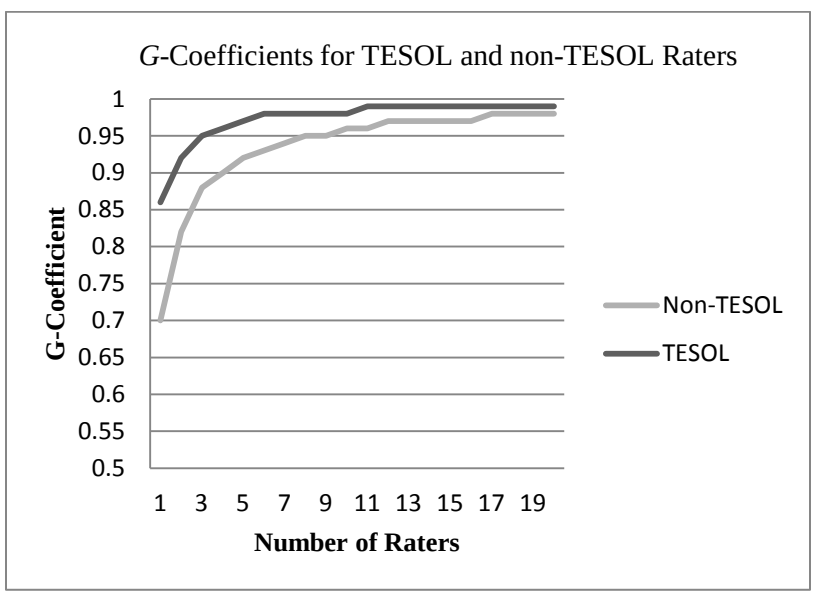
Table 6

Summary of G-coefficients for TESOL and Non-TESOL Raters

Number of Papers	Number of Raters	*G*-Coefficients	
		TESOL	NTESOL
3	1	.86	.70
3	2	.92	.82
3	3	.95	.88
3	4	.96	.90
3	5	.97	.92
3	6	.98	.93
3	7	.98	.94
3	8	.98	.95
3	9	.98	.95
3	10	.98	.96
3	11	.99	.96
3	12	.99	.97
3	13	.99	.97
3	14	.99	.97
3	15	.99	.97
3	16	.99	.97
3	17	.99	.98
3	18	.99	.98
3	19	.99	.98
3	20	.99	.98

Figure 1

G-coefficients for Different Number of Raters with and without TESOL Background



The calculation of Phi-coefficients for criterion-referenced score interpretations

Using the formula $\Phi = \frac{\sigma_\rho^2}{\sigma_\rho^2 + \sigma_\Delta^2}$, dependability (Phi) coefficients for each rater group (i.e., raters with TESOL educational background versus raters without TESOL educational background) were calculated for criterion-referenced score interpretations. The results are presented in Table 7.

As shown in Table 7, the Phi-coefficient obtained for non-TESOL raters for the current 10-rater scenario was .94, whereas the Phi-coefficient for the TESOL teacher candidates was .98. However, in classroom assessment context, where only one rater scores each student paper, the Phi-coefficient for the non-TESOL raters was .62; in contrast, the Phi-coefficient for the TESOL raters would be .86. Again, comparing the reliability of one rater with TESOL background to another rater with no TESOL background reveals a much larger difference in terms of reliability.

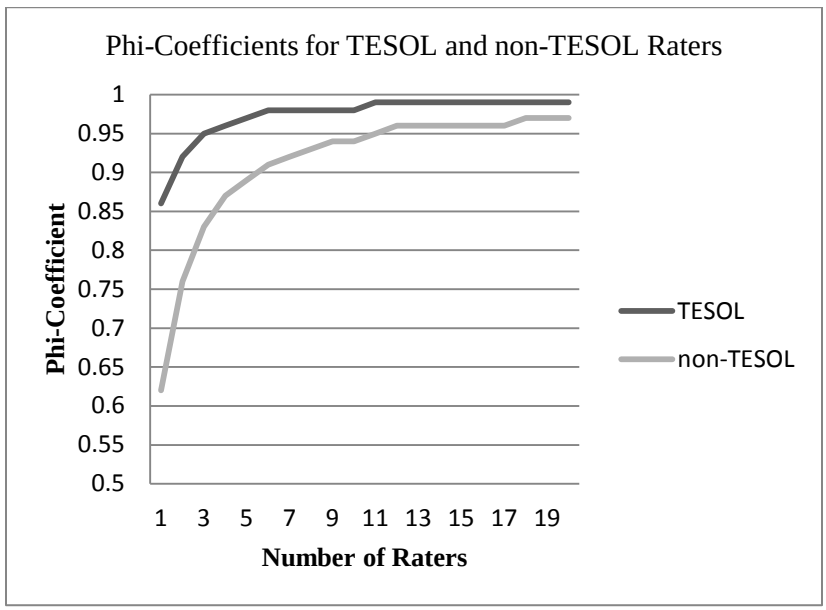
Table 7
Summary of Dependability Coefficients for TESOL and Non-TESOL Raters

Number of Papers	Number of Raters	Dependability Coefficients	
		ESL	NESL
3	1	.86	.62
3	2	.92	.76
3	3	.95	.83
3	4	.96	.87
3	5	.97	.89
3	6	.98	.91
3	7	.98	.92
3	8	.98	.93
3	9	.98	.94
3	10	.98	.94
3	11	.99	.95
3	12	.99	.96
3	13	.99	.96
3	14	.99	.96
3	15	.99	.96
3	16	.99	.96
3	17	.99	.96
3	18	.99	.97
3	19	.99	.97
3	20	.99	.97

Figure 2 displays the differences in Phi-coefficients between TESOL and non-TESOL raters. Similar to Figure 1, this figure clearly displays that the TESOL raters are much less variant and much more consistent and reliable than the non-TESOL raters in the context of criterion-referenced score interpretations.

To sum up, the results of the *G*-theory analyses presented above indicate that raters' ESL-related educational background did impact their rating of ESL students' essays. Raters with TESOL educational background did mark ESL essays more consistently and reliably than those without any TESOL educational background.

Figure 2
Phi-coefficients for Different Number of Raters with and without TESOL Background



DISCUSSION AND CONCLUSIONS

The first research question attempted to determine if there would be significant differences between the writing scores assigned by raters with and without ESL educational background. Descriptive statistical analysis and paired samples *t*-tests indicate no significant differences in the scores assigned to ESL essays #1 and #2. However, the scores assigned to essay #3 by non-TESOL raters were significantly lower than the scores assigned by TESOL raters. Further, the standard deviations for essays #1 and #3 scores assigned by the non-TESOL raters were approximately two times larger than the standard deviations of the scores assigned by the TESOL raters; for essay #2 scores, the standard deviation for the non-TESOL raters was almost five times larger than the standard deviation for the TESOL raters. All these results suggest that the non-TESOL raters were considerably more variant and less consistent in scoring the three ESL essays than the TESOL raters.

The second research question aimed to examine the sources of score variation contributing relatively more to the scores assigned by raters with TESOL educational background in contrast to raters without TESOL educational background. First, the results show that the desired variance component person (*p*), which was the object of measurement, explained 85.62% of the total score variance for the TESOL rater group; however, the same variance component explained only 61.75% of the total variance for the non-TESOL rater group. Second, the results show that for the non-TESOL rater group, the undesired rater (*r*) variance component accounted for 12.03% of the total variation in comparison to 0% for the TESOL rater group. Finally, the residual variance component was also greater for the non-TESOL rater group (26.22% of total variance) than for the TESOL rater group (14.38% of total variance), indicating that there was more unexplained score variance for the non-TESOL raters than for the TESOL raters.

The last research question aimed to find if the reliability (i.e., generalizability coefficients for norm-referenced score interpretations and dependability coefficients for criterion referenced score interpretations) of the scores assigned by raters with TESOL educational background differ from the reliability of the scores assigned by raters without TESOL educational background. The findings indicate that the TESOL raters are much less variant and much more reliable than the non-TESOL raters in both the norm-referenced and criterion-referenced score interpretation contexts. In other words, in both score interpretation contexts, the non-TESOL raters were much less consistent and reliable than the TESOL raters.

To sum up, the results of this study indicate that raters' TESOL-related educational background did impact their rating of ESL students' essays. The TESOL raters did mark ESL essays more consistently and reliably than their non-TESOL counterparts. This is perhaps because TESOL raters have systematically learned about how ESL students acquire their English language skills as well as how they should help develop their English language abilities and assess these abilities in the classrooms.

Limitations

This study was limited in the following four ways. First, the writing samples selected for analysis were not representative of ESL writing of different genres and quality levels. In other words, only one genre of writing (i.e., descriptive) was selected for this study; further all three ESL essays were obtained from a single high school and they were all of intermediate quality. As Huang (2008) indicated, different types of writing impact the scoring variability and reliability of ESL writing. Furthermore, task-related factors have been found to be responsible for students' writing performance, the rating process, and its reliability (Jennings, Fox, Graves, & Shohamy, 1999; Reid 1990; Weigle, 1999).

Second, this study only examined the impact of rater educational background on the assessment of ESL writing. Many studies have shown that multiple factors jointly affect the assessment of ESL writing (Brown, 1991; Huang, 2008, 2011, 2012; Huang & Foote,

2010; Santos, 1988; Song & Caruso, 1996; Vann et al., 1984; Weigle et al., 2003). The ignorance of other factors might have led to the large residual variance component in the *G*-theory analyses.

Third, the 20 raters of this study did not receive any formal training before scoring these ESL essays. As indicated in the literature, rater training is an important procedure that can minimize rater variation in the assessment of ESL writing (e.g., Davidson, 1991; Jacobs et al., 1981; Weigle, 1994). However, the fact that classroom teachers scoring students' essays in authentic settings do not usually receive formal training provides a counter argument to this potential limitation (Huang & Foote, 2010).

Finally, only holistic scoring was employed in this study. Research has shown that the use of different scoring methods may affect the reliability of the rating of ESL students' writing (Huang & Foote, 2010; Russikoff, 1995). In this study, the raters provided only a holistic score based on a number of criteria, it may be that "a single criterion focused on writing clarity, the use of jargon, misspellings, and grammatical error was the most influential factor in the overall grade" (Huang & Foote, 2010, p. 230).

Conclusions

In light of the limitations, the following two conclusions were reached. First, rater educational background did, in fact, affect the rating variability and reliability of ESL students' writing. Based on the findings of this study, the TESOL teacher candidates are more consistent and reliable than the non-TESOL candidates in rating ESL essays. This is perhaps because the TESOL teacher candidates have systematically learned and understood the factors affecting ESL students' learning and how they should assess ESL writing in a consistent and reliable manner. Further, this difference between the TESOL and non-TESOL teacher candidates in rating ESL essays could be larger if writing tasks of a variety of genres and essays of different qualities are considered in the design.

Second, there is still large unexplained variability. As mentioned above, the residual contains the variability due to the interaction between raters and papers, and other unexplained systematic and unsystematic sources of error. Large residual effects can indicate hidden facets (Brennan, 2001). The variance of the hidden facets is included in the residual variance, thus leading to a larger residual than when the hidden facets are explicitly considered in the design (Huang & Foote, 2010).

Implications

This study was designed to examine the effects of raters' TESOL-related educational background on the scoring variability and reliability of ESL writing. The results can provide important implications for state/provincial level policy makers, teacher preparation institutions/programs as well as in-service teachers and teacher candidates.

First, with the increasing number of ESL students studying in North American K-12 schools (Huang et al., 2011; Huang et al., 2010), policy makers should adjust current policies to address these students' learning needs at schools. To ensure that ESL students have an equal opportunity for success, it is suggested that the state/provincial level policy makers make it mandatory for teacher preparation institutions/programs to include TESOL course(s) in their curricula. Such a policy can ensure that future teachers are better prepared to meet the learning needs of ESL students North American school systems.

Second, whether or not state/provincial policies are made to require institutions to include TESOL course(s) in their teacher training curricula, it is highly recommended that the institutions require their teacher candidates to take a certain number of TESOL courses. This institutional requirement will surely make the teacher training programs more suitable for today's multicultural classrooms.

Third, like many professionals, in-service teachers are often encouraged to seek professional development opportunities. It is suggested that they make it a priority to attend workshops, take training courses, and attend conferences related to the teaching and assessment of ESL students in order to enhance their knowledge and skills in teaching and assessing ESL students' learning in the classroom.

Finally, this study has implications for teacher candidates as well. It is highly recommended that they take TESOL course(s) as their electives even if they are not required for the completion of their programs. This is because a well-prepared future classroom teacher should have the basic knowledge and skills needed for the teaching and assessment of ESL students in the classroom.

Recommendations for Future Research

This study provides at least three directions for future research in the area of ESL assessment. First, future studies should examine the effects of more factors that are found to affect ESL writing assessment. For example, the rating methods (holistic vs. analytical), the writing tasks (persuasive vs. descriptive), and essay qualities (low, intermediate, high) should all be considered in the investigation. It is believed that these factors jointly affect the assessment of ESL students' writing (Huang, 2008, 2011, 2012; Huang & Foote, 2010; Reid, 1990).

Second, future studies can use qualitative approaches as a complementary method in the investigations. This is because qualitative approaches such as think-aloud protocols and rater interviews can provide valuable information and evidence about the rating processes and products (e.g., Connor-Linton; 1995; Sakyi; 2000; Weigle, 1994). Further, these qualitative procedures can provide more in-depth and valid data for the research (Connor-Linton, 1995).

Finally, future studies should expand this area of research to include large-scale standardized (e.g., state or provincial examinations) ESL writing assessment in their designs so that comparisons can be made between large-scale assessment and small-scale classroom assessment contexts in terms of rating reliability, validity, and fairness issues. These comparisons allow the researchers to better understand the assessment issues and make new contributions to the current body of knowledge in the field of ESL writing assessment.

REFERENCES

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*, 57-74.

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

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

Casanave, C. P., & Hubbard, P. (1992). The writing assignments and writing problems of doctoral students: Faculty perceptions, pedagogical issues, and needed research. *English for Specific Purposes, 11*, 33-49.

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

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? *TESOL Quarterly, 29*(4), 762-765.

Crick, J. E., & Brennan, R. L. (1983). *GENOVA: A general purpose analysis of variance system. Version* 2.1. Iowa City, IA: American College Testing Program.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Davidson, F. (1991). Statistical support for training in ESL composition rating. In L. Hamp-Lyons (Ed.), *Assessing second language writing* (*pp*. 155-165). Norwood, NJ: Ablex.

Gamaroff, R. (2000). Rater reliability in language assessment: The bug of all bears. *System, 28*, 31-53.

Gere, A. R. (1980). Written composition: Toward a theory of evaluation. College English, 42, 44-48.

Hamp-Lyons, L. (1996). The challenges of second language writing assessment. In E. Whit, W. Lutz and S. Kamusikiri (eds.), Assessment of writing: Policies, politics, practice (*pp.* 226-240) New York: Modern Language Association.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach. *Assessing Writing, 13,* 201-218.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies, 5*(1), 1-17.

Huang, J. (2011). Generalizability theory as evidence of concerns about fairness in large-scale ESL writing assessments. *TESOL Journal, 2*(4), 423-443.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17,* 123-139.

Huang, J., Cunningham, J., & Finn, A. (2010). Teacher perceptions of ESOL students' greatest challenges in academic English skills: A K-12 perspective. *International Journal of Applied Educational Studies, 8*(1), 68-80.

Huang, J., & Foote, C. J. (2010). Grading between lines: What really impacts professors' holistic evaluation of ESL graduate student writing? *Language Assessment Quarterly, 7,* 219-223.

Huang, J., Smith, A., & Smith, M. (2011). Teacher perceptions of strategies for improving ESOL students' academic English skills: A K-12 perspective. *The Canadian and International Education Journal, 40*(3), 61-80.

Huot, B. A. (1990). Reliability, validity and holistic scoring: What we know and what we need to know.*College Composition and Communication, 41,* 201-213.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughhey, J. B. (1981).*Testing ESL composition: A practical approach.* Rowley, MS: Newbury House.

Janopoulos, M. (1992). University faculty tolerance of NS and NNS writing errors: A comparison. *Journal of Second Language Writing, 1*(2), 109-121.

Janopoulos, M. (1995). Writing across the curriculum, writing proficiency exams, and the NNS college student. *Journal of Second Language Writing, 4,* 43-50.

Jenkins, S. (2000). Cultural and linguistic miscues: A case study of international teaching assistant and academic faculty miscommunication. *International Journal of Intercultural Relations, 24,* 477-501.

Jennings, M., Fox, J., Graves, B., & Shohamy, E. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. *Language Testing, 16*(4), 426-456.

Johnson, R. L., Penny, J., Gordon, B., Shumate, S. R., & Fisher, S. P. (2005). Resolving score differences in the rating of writing samples: Does discussion improve the accuracy of scores? *Language Assessment Quarterly, 2*(2), 117-146.

Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Edmonton, AB.

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

Mendelsohn, D., & Cumming, A. (1987). Professors' ratings of language use and rhetorical organization in ESL compositions. *TESL Canada Journal, 5,* 9-26.

Popham, W. J. (2011). *Classroom assessment: What teachers need to know?* (6th ed.). Boston, MA: Pearson Education, Inc.

Reid, J. (1990). Responding to different topic types: A quantitative analysis from a contrastive rhetoric perspective. In B.Kroll (Ed.), *Second language writing: Research insights for the classroom* (*pp.* 191-210). Cambridge: Cambridge University Press.

Ruetten, M. K. (1994). Evaluating ESL students' performance on proficiency exams. *Journal of Second Language Writing, 3*, 85-96.

Russikoff, K. A. (1995). *A comparison of writing criteria: Any differences?* Paper presented at the annual meeting of the Teachers of English to Speakers of Other languages, Long Beach, CA.

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

Santos, T. (1988). Professors' reactions to the writing of nonnative-speaking students. *TESOL Quarterly, 22(1)*, 69-90.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement, 30*, 215–232.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A premier*. Newbury Park, CA: Sage.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking and ESL students? *Journal of Second Language Writing, 5*, 163-182.

Statistics Canada. (2011). Population projections: Canada, the provinces and territories. Retrieved October 15, 2011 from Statistics Canada: http://www.statcan.gc.ca/daily-quotidien/100526/dq100526b-eng.html

Sweedler-Brown, C. O. (1993). ESL essay evaluation: The influence of sentence-level and rhetorical features. *Journal of Second Language Writing, 2*(1), 3-17.

Trice, A. G. (2001). *Faculty perceptions of graduate international students: The benefits and challenges.* Paper presented at the 26th Annual Meeting of the Association for the Study of Higher Education, Richmond, VA. (ERIC Document Reproduction Service No. ED 457 816).

U. S. Department of Homeland Security. (2010). 2010 Yearbook of Immigration Statistics. Office of Immigration Statistics. Retrieved October 17, 2011 from U.S. Department of Homeland Security: http://www.dhs.gov/xlibrary/assets/statistics/yearbook/2010/ois_yb_2010.pdf

Vann, R., Meyer, D., & Lorenz, F. (1984). Error gravity: A study of faculty opinion of ESL errors. *TESOL Quarterly, 18*, 427-440.

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

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

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

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197-223.

Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.