



Available online at www.sciencedirect.com

 ScienceDirect

Assessing Writing 13 (2008) 201–218

ASSESSING
WRITING

How accurate are ESL students' holistic writing scores on large-scale assessments?—A generalizability theory approach

Jinyan Huang*

College of Education, Academic Complex, 329E, Niagara University, Niagara, NY 14109, United States

Abstract

Using generalizability theory, this study examined both the rating variability and reliability of ESL students' writing in the provincial English examinations in Canada. Three years' data were used in order to complete the analyses and examine the stability of the results. The major research question that guided this study was: Are there any differences between the rating variability and reliability of the writing scores assigned to ESL students and to Native English (NE) students in the writing components of the provincial examinations across three years? A series of generalizability studies and decision studies was conducted. Results showed that differences in score variation did exist between ESL and NE students when adjudicated scores were used. First, there was a large effect for both language group and person within language-by-task interaction. Second, the unwanted residual variance component was significantly larger for ESL students than for NE students in all three years. Finally, the desired variance associated with the object of measurement was significantly smaller for ESL students than for NE students in one year. Consequently, the observed generalizability coefficient for ESL students was significantly lower than that for NE students in that year. These findings raise a potential question about the fairness of the writing scores assigned to ESL students. © 2008 Published by Elsevier Ltd.

Keywords: Writing assessment; Generalizability theory; Rating variability; Rating reliability; ESL writing

1. Introduction

The assessment of writing has long been considered a problematic area for educational assessment professionals. As stated by Speck and Jones (1998), "there are more problems than

* Tel.: +1 716 286 8259.

E-mail address: jhuang@niagara.edu.

solutions—problems of inter-grader reliability, single-grader consistency, and ultimate accountability for the grades we assign” (p. 17). Variation among and within raters’ rating of students’ writing contributes to measurement error and thus may threaten the fairness of the assessment of writing (Popham, 1990). Due to the different linguistic and cultural backgrounds of English-as-a-second-language (ESL) students, the assessment of their English writing becomes even more problematic (Connor-Linton, 1995; Hamp-Lyons, 1991; Sakyi, 2000). On the one hand, many factors affect ESL students’ writing, including their English proficiency, mother tongue, home culture, and style of written communication (Hinkel, 2003; Yang, 2001). In rating ESL students’ writing, raters may differentially consider these factors, and empirical studies have found differences in rater behavior for ESL writing assessments (Bachman, 2000). A number of studies indicate that rater and task as factors affect the assessment of ESL writing. For example, rater background, mother tongue, previous experience, amount of prior training, and types and difficulty of writing tasks have been found to affect the rating of the written responses of ESL students (Santos, 1988; Weigle, 1999). The impact of these factors leads to questions about the accuracy, precision and ultimately, the fairness of the scores obtained from the ratings of written work produced by ESL students.

Increasingly, writing-proficiency standards are being established for both secondary school and university students in North America regardless of students’ native languages (Johnson, Penny & Gordon, 2000). Within this context, ESL students have to compete with native English (NE) students in writing. Like NE students, ESL students are expected to successfully demonstrate their ability to write English compositions or complete high-stakes essay examinations (Wiggins, 1993). However, research shows that ESL students face considerable challenges passing these institutional or provincial/state competency examinations of writing (Thompson, 1990). Further these difficulties may be due to more than language deficiencies. As an example, rating inconsistency could be one reason for ESL students’ failure and poor performance on these writing examinations. Previous studies have found that raters with different teaching experience assign different scores to the same piece of ESL writing (Hamp-Lyons, 1996; Vaughan, 1991).

It is believed that rating consistency or reliability is essential to sound performance assessment practice, although this presumption has been challenged by some scholars (Moss, 1994). In the context of writing assessment, there may exist unwanted variations in scores due to variations among raters and within raters (Bachman, 1990; Johnson et al., 2000). Both of these variations are problematic as they adversely affect the reliability of the scores assigned to students. Rating reliability indicates the precision of the rating of students’ writing, which is related to fairness for test-takers (Johnson et al., 2000). Therefore, rating reliability should be treated as a cornerstone of sound performance assessment.

Classical test theory (*CTT*) is most commonly used as a theoretical framework for the detection of rater variation and estimating reliability in performance assessment situations. However, generalizability (*G-*) theory (Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is a more powerful approach than *CTT* for the detection of rater variation and estimating reliability (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. While *CTT* provides a single estimate of error, *G-*theory can be used to identify not only multiple sources of error but also the impact of these sources of error on the overall accuracy (Shavelson & Webb, 1991). Through *generalizability* (*G-*) and *decision* (*D-*) studies, researchers can evaluate the relative importance of various sources of measurement error and interpret score reliability from both norm- and criterion-referenced perspectives. Thus *G-*theory provides a comprehensive conceptual framework and methodology for analyzing more than one measurement facet (factor) simultaneously in investigations of

assessment error and score dependability (Brennan, 2001). Therefore, *G*-Theory was used as the theoretical framework of this research.

Using *G*-theory, the purpose of this study was to examine both the rating variability and reliability of large-scale ESL students' writing in the provincial English examinations in one province in Canada. This province was chosen because it has a large-scale English examination that is designed to allow secondary school students to demonstrate that they have met provincial graduation requirements and, among the provinces in Canada, this province has a substantial number of ESL students in K-12 schools (CBIE, 2002). ESL students are those who enter Canadian 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-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 major research question that guided this study was: Are there any differences between the rating variability and reliability of the writing scores assigned to ESL students and to NE students for the provincial English examination in that province across a 3-year period? Within the framework of *G*-theory, the following three sub-questions were investigated:

- (1) What are the independent sources of variation (e.g., person within language group, rating, task, and the interactions among these facets) in the writing scores assigned to ESL students in contrast to NE students across different test administrations?
- (2) Does the reliability (i.e., generalizability coefficients for norm-referenced score interpretations) of the writing scores assigned to ESL students differ from the reliability of the writing scores assigned to NE students across different test administrations?
- (3) If differences are found between the rating reliabilities for ESL and NE students, what is the potential impact of these differences on the rating designs for ESL students in comparison to NE students?

2. Description of the provincial English examination

The provincial English examination assesses Grade 12 students' reading and writing skills. It is intended for students who are developing the thinking skills needed to analyze and interpret short stories, poems, and non-fiction essays, and the writing skills needed for communication at the post-secondary education level. This examination consists of four parts: (a) informational text; (b) interpretation of poetry; (c) interpretation of literary prose; and (d) original composition. Reading and writing are integrated components within these four parts, with reading worth 27% and writing worth 73% of the total examination mark. The writing component consists of three writing tasks: (a) unified and coherent paragraph(s) of approximately 125–150 words about a poem; (b) a multi-paragraph response of approximately 300 words to a literary prose; and (c) a multi-paragraph original composition of approximately 300 words. Thematically, the three writing tasks are not related. The first writing task is worth 15% of a student's total examination mark and each of the second and third is worth 29%. The provincial examination score counts for 40% of the students' final course marks and the school-awarded mark counts for 60% of the final course marks. Given the nature of the provincial examination program, a norm-referenced framework rather than a criterion-referenced framework is predominant.

There are different rating guides for each of the three writing tasks. Each task is rated holistically on an almost identical 6-point scale and rating is done in pairs (usually a more experienced rater is paired with a newer one). A bundle of 30 papers is split between the two raters. Each pair is responsible for rating only the section for which they have been trained (i.e., poetry, prose, or

composition). One rater rates the first 15 papers while the second one rates the second 15 papers, and then they switch. Raters record their scores on a separate piece of paper. After all 30 papers have been rated by both raters, they compare their scores. A difference of one score is allowed (e.g., Rater One gave the response a 4 and Rater Two gave the response a 5). If there is a difference greater than one score (e.g., Rater One gave the response a 5 and Rater Two gave the response a 3) and the raters cannot come to an agreement, the paper is then given to the section head, who reads the paper, listens to the arguments from both raters, and then makes a decision. However, the recorded score must still be a double score. For the above example, the section head would decide if the paper would receive a 3/4, 4/4 or 4/5. For each writing task, the scores from the two raters must be whole numbers and there can only be a maximum of a one score difference between the two scores. The writing task for poetry is worth 12 points (the sum of the two raters' scores), and the other two writing tasks are worth 24 points each (the sum of the two raters' scores multiplied by 2).

3. Method

Existing data from the writing components of the 2002, 2003, and 2004 administrations of the provincial English examination were used for *G*-theory analyses. By using data for three consecutive years, it was possible to replicate the analyses and check the stability of the results. Further, given the nature of the rating design, it was not possible to track the scores assigned by each rater. Under these conditions it is not possible to examine rater effects directly; however, it is possible to treat "rating" as a random facet for *G*-theory analyses and examine the impact of rating on the score reliability. This strategy has been used in a variety of large-scale language performance assessments (Bachman, Lynch, & Mason, 1995; Lee, Kantor, & Mollaun, 2002).

The data set used for analysis was comprised of the writing scores of all ESL students and an equal number of randomly selected NE students who wrote one of the June 2002, June 2003, and June 2004 provincial English examinations. The number of students who wrote the June administrations of the provincial English examination in the three years and the sample sizes selected for data analyses are listed in Table 1.

The provincial English examination is administered five times a year—in November, January, April, June, and August. The first four occasions accommodate the fact that different schools follow a quarter, semester, or full year timetable. The August examination is written by students who failed on one of the first four occasions or wish to obtain a higher score on the examination. The June administration was chosen because the largest numbers of students complete the English examination in June. In the English examination, as previously described, the students were asked to complete three separate writing tasks (paragraph format writing task for poetry, essay format writing task for literary prose, and original composition) and each writing task received scores from two different independent raters. However, the final data file obtained for analyses contained adjudicated scores. Consequently, there could only be a maximum of a one score difference

Table 1
Number of students writing the provincial English examinations: June administration.

	All students	ESL students
June 2002	32,069	323
June 2003	31,887	425
June 2004	29,622	357

between the two scores. If there was a difference greater than one score between the two ratings, a third or adjudicated rating was required for that writing task. Due to data management limitations in that province, it was not possible to track the third rating information. Hence, no information was available regarding the number of students who received adjudicated ratings. Further, data prior to adjudication were not available; a limitation that was not known until the data were received. Rather, the adjudicated scores were used for analyses, resulting in a maximum of a one score point difference between ratings. Although the adjudicated scores may mask a much greater uncertainty among ratings than the original scores, they are actually reported in large-scale assessments. Therefore, for each administration of the examination, the data included student language background with an ESL or NE identifier and the two adjudicated scores awarded each student (on a 1–6 point scale) for each writing task.

Three years' data for the writing component of the provincial examination were obtained in a Microsoft EXCEL file. Using Microsoft ACCESS, it was possible to create a database for the obtained data set. Further, using query techniques, the data were separated into six individual Excel files (two for each administration—ESL and NE) containing each student's language background identifier (ESL or NE) and his or her total scores for each writing task. All ESL students who had scores for all writing tasks were included in the analysis. The six Excel files associated with ESL and NE students were converted into SPSS files for the descriptive analyses. Using the SPSS random selection function, NE students who had scores for each writing task were randomly selected for each administration so that the number of NE students matched the number of ESL students. Lastly, the ESL EXCEL files and the sampled NE SPSS files were converted into MANUAL files for the *G*-theory analyses.

Descriptive analyses for the provincial English examination results were conducted for all ESL students, all NE students, and the sampled NE students. The purposes of conducting the descriptive analyses were to examine both the mean and variance differences between ESL and sampled NE students, and to check the representativeness of the sampled NE students of the whole NE population.

Within a *G*-theory framework, data were then analyzed in the following six stages: *G*-study 1, *G*-study 2, comparative analyses on variance components, calculation of *G*-coefficients, *F*-test of *G*-coefficients, and *D*-study.

3.1. *G*-study 1

A separate $p:l \times t \times r'$ (person within language group-by-task-by-rating) mixed effects *G*-study analysis was conducted for each test administration. *G*-study 1 was a mixed effects design with language group fixed and all other facets random. The purpose of this *G*-study was to obtain the eleven independent sources of variation ($l, p:l, t, r', l \times t, l \times r', p:l \times t, p:l \times r', t \times r', l \times t \times r',$ and $p:l \times t \times r'$) and then identify facets of concern.

3.2. *G*-study 2

Separate $p \times t \times r'$ (person-by-task-by-rating) random effects *G*-study analyses were conducted for ESL students and for NE students. *G*-study 2 was a fully crossed design with all facets random. The main purpose of *G*-study 2 was to obtain information for comparison between ESL and NE students in terms of score variability and reliability. It was expected that some differences would be found between ESL and NE students. With the implementation of *G*-study 2, the seven independent sources of variation ($p, t, r', p \times t, p \times r', t \times r',$ and $p \times t \times r'$) for each language

group were obtained. Once this had been done, significant differences in variance components between language groups could be determined. Finally, using the obtained variance components, *G*-coefficients for each language group were calculated and the significant differences tested. Further, if the *G*-coefficients between language groups were found to be significantly different, each *G*-study could then provide information for the corresponding *D*-study in order to examine the potential impact of these differences on the rating designs for ESL students in comparison to NE students (i.e., the number of independent ratings and writing tasks required for each language group so that the *G*-coefficients of both groups were comparable).

3.3. Comparative analyses on variance components

Based on the *G*-study two results, standard errors of variance components and confidence intervals were calculated to examine whether the variance components were significantly different between the two language groups. Like all other statistics, the variance component estimates are subject to sampling variability. Therefore, standard errors of variance component estimates were required. An estimate of a standard error is sometimes sufficient for investigators to make judgments about the variability of estimated variance components. More often than not, however, investigators want to establish a confidence interval (Brennan, 2001). Standard errors of variance components can be used to construct confidence intervals around the variance components computed using the normal method.

Brennan's (2001) formulas were used to calculate standard errors of variance components. The general formula for the estimation of standard errors of *G*-study variance components is

$$\hat{\sigma} \left[\hat{\sigma}^2 \left(\frac{\alpha}{M} \right) \right] = \sum_{\beta} \frac{2[f(\beta/\alpha)MS(\beta)]}{df(\beta) + 2}, \quad (1)$$

where *M* designates the model (e.g., $p \times t \times r'$), β the index for each of the mean squares that enter into $\hat{\sigma}^2(\alpha/M)$, and $f(\beta/\alpha)$ is the coefficient of $MS(\beta)$ in the linear combination of mean squares that gives $\hat{\sigma}^2(\alpha/M)$. The general formula for the estimation of standard errors of *D*-study variance components is

$$\hat{\sigma} \left[\hat{\sigma}^2 \left(\frac{\alpha}{M'} \right) \right] = \frac{C(\bar{\alpha}/\tau)}{d(\bar{\alpha}/\tau)} \hat{\sigma}^2 \left[\hat{\sigma}^2 \left(\frac{\alpha}{M} \right) \right], \quad (2)$$

where $C(\bar{\alpha}/\tau) = \{ \text{the product of the terms } (1 - \frac{n'}{N}) \text{ for all primary indices in } \bar{\alpha} \text{ except } \tau. \text{ If } \bar{\alpha} = \tau, \text{ then } C(\bar{\alpha}/\tau) = 1 \}$; $d(\bar{\alpha}/\tau) = \{ \text{the product of the } D\text{-study sample sizes } (n') \text{ for all indices in } \bar{\alpha} \text{ except } \tau. \text{ If } \bar{\alpha} = \tau, \text{ then } d(\bar{\alpha}/\tau) = 1 \}$.

Using these standard errors, confidence intervals were constructed:

$$\text{Lower limit} = \hat{\sigma}_{\alpha/M}^2 - 1 - \frac{z}{2} SE_{\alpha/M} \quad (3)$$

$$\text{Upper limit} = \hat{\sigma}_{\alpha/M}^2 + 1 - \frac{z}{2} SE_{\alpha/M} \quad (4)$$

If the confidence interval for a variance component for the ESL sample and the confidence interval for the corresponding variance component for the NE sample did not overlap, then it was concluded that the two variance components differed at the α level of significance. The .05-level of significance was adopted for this purpose.

3.4. Calculation of G-coefficients

The following formula (r' stands for 'rating') was used for the calculation of G -coefficient ($E\rho^2$), which is the reliability for norm-referenced score interpretations (Lee & Mollaun, 2002):

$$E\rho^2 = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_\delta^2} = \frac{\sigma_s^2}{\sigma_s^2 + \sigma_{st}^2/n_t + \sigma_{sr/r'}^2/n_{r/r'} + \sigma_{str/r'}^2/n_t n_{r/r'}} \quad (5)$$

where σ_s^2 is the universe score variance and σ_δ^2 is the relative error variance.

Based on the G -study 2 results, G -coefficients for each language group (ESL versus NE) were calculated for each administration. The purpose of calculating the G -coefficients was to answer the second research sub-question: Does the reliability of the writing scores assigned to ESL students differ from the reliability of the scores assigned to NE students?

3.5. F-test of G-coefficients

Within each administration, G -coefficients ($E\rho^2$) were compared between the two language groups by computing an F -statistic and comparing it with a tabled F -value. Feldt (1969) presented a procedure for assessing the significance of a difference between two independent Cronbach's alpha coefficients. This procedure was based on an earlier derivation (Feldt, 1965; Kristof, 1963) of the sampling distribution of coefficient alpha (subsequently denoted r_α for the sample value and ρ_α for the parameter). Feldt (1965) and Kristof (1963) showed that the ratio $F = (1 - r_\alpha)/(1 - \rho_\alpha)$ is distributed as a central F -variable with degrees of freedom equal to $(N - 1)(k - 1)$ and $(N - 1)$ (N is the number of examinees and k represents the parallel parts of a test instrument). Feldt's (1969) procedure is based on the distribution of a product of two independent F -variables, thus limiting the experimental context to one involving two coefficients (Hakstian & Whalen, 1976). The statistic for conducting the test of $H_0: \rho_1 = \rho_2$ is

$$W = (F_1)(F_2) = \frac{1 - r_2}{1 - r_1}, \quad (6)$$

where F_1 has degrees of freedom equal to $(N_1 - 1)$ and $(N_1 - 1)(k_1 - 1)$ and F_2 has degrees of freedom equal to $(N_2 - 1)(k_2 - 1)$ and $(N_2 - 1)$. Here $df_1 = (N_1 - 1)$, $df_2 = (N_1 - 1)(k_1 - 1)$, $df_3 = (N_2 - 1)(k_2 - 1)$, and $df_4 = (N_2 - 1)$. By solving the following pair of equations, the degrees of freedom for the numerator (v_1) and denominator (v_2) of the desired F distribution may be obtained:

$$A = \frac{df_4}{df_4 - 2} \cdot \frac{df_2}{df_2 - 2} = \frac{v_2}{v_2 - 2}, \quad (7)$$

$$B = \frac{(df_1 + 2)(df_4)^2}{(df_4 - 2)(df_4 - 4)(df_1)} \cdot \frac{(df_3 + 2)(df_2)^2}{(df_2 - 2)(df_2 - 4)(df_3)} = \frac{(v_1 + 2)(v_2)^2}{(v_2 - 2)(v_2 - 4)(v_1)}, \quad (8)$$

The resultant values are

$$v_1 = \frac{2A^2}{2B - AB - A^2}, \quad (9)$$

$$v_2 = \frac{2A}{A - 1}. \quad (10)$$

The Feldt (1969) procedure, therefore, was applied to the investigation of whether the differences between G -coefficients of ESL and NE student groups were significant.

3.6. *D-study*

If the G -coefficients were significantly different between ESL and NE student groups for an administration, a separate $p \times T \times R'$ (person-by-task-by-rating) random effects D -study analysis was conducted for ESL and NE students for the administration. This D -study was a fully crossed design with the task and rating facets random. The purpose of this D -study was to answer the third research sub-question: What is the impact of differences in G -coefficients on the rating designs for ESL students in comparison to NE students?

The computer program GENOVA (Crick & Brennan, 1983) was used for the G - and D -studies. GENOVA is a computer program used to estimate the variance components for the main and interaction effects and their standard errors using the formulas provided above and where the design is balanced. The program also computes the G -coefficients ($E\rho^2$) and dependability coefficients ($\Phi(\lambda)$) for different values of the cut-score λ .

SPSS, Microsoft EXCEL, Microsoft ACCESS, and DistCalc (Lohninger, 2002) were used for both the descriptive analyses and the statistical comparisons. The computer program DistCalc calculates probabilities and critical values for the most important distributions. The purpose of this program is to illustrate the concept of critical values and to replace printed distribution tables. DistCalc provides calculations for the normal distribution, the t distribution, the χ^2 distribution, and the F distribution.

4. Results

4.1. *Descriptive results*

Table 2 provides the descriptive statistics for the ESL and NE student data used in the analyses. As mentioned above, three writing tasks on the provincial English examination were rated holistically by two independent raters on a 6-point scale. Table 2 provides both the mean and standard deviation of the ratings of each writing task assigned by each rater. Based on the results, it appears the sample of NE students and the population of NE students had very similar mean scores and standard deviation across the three years. Hence the sampling of NE students could be considered representative of the data set. Comparing the results of the ESL and NE students, the results show that ESL students had consistently lower performance on all three writing tasks than NE students in all three years. Further, these differences were somewhat large given the 6-point scale used. For example, the ESL scores were between 0.69 and 1.10 score points lower than the NE students. Lastly, with the exception of 2004, the standard deviations for the ESL and NE students were similar across items. In 2004, the standard deviations for the ratings of ESL students were consistently lower than those of the NE students. Further, the NE students scores in 2004 tended to be more varied than found in the other two years.

4.2. *G-study 1*

G -study 1 was a $p:L \times t \times r'$ (person within language group-by-task-by-rating) mixed design with language group fixed and all other facets random. Table 3 presents the results for the three

Table 2
Descriptive statistics.

Year	Task/Rating	ESL		NE		ALL NE	
		Mean	S.D.	Mean	S.D.	Mean	S.D.
2002	T1R1	2.70	.94	3.51	.99	3.52	.98
	T1R2	3.04	.97	3.73	1.02	3.72	1.00
	T2R1	2.56	.97	3.38	1.00	3.40	.98
	T2R2	2.76	.98	3.69	1.03	3.71	1.01
	T3R1	2.76	.88	3.75	.79	3.76	.77
	T3R2	2.94	.82	4.04	.80	4.03	.78
2003	T1R1	2.82	1.13	3.77	1.04	3.69	1.12
	T1R2	2.97	1.12	3.88	1.06	3.89	1.13
	T2R1	2.75	1.06	3.78	1.05	3.78	1.12
	T2R2	2.86	1.06	3.88	1.02	3.80	1.09
	T3R1	3.10	.93	3.94	.97	3.97	1.02
	T3R2	3.16	.98	4.07	.95	4.08	1.01
2004	T1R1	2.70	1.06	3.63	1.24	3.69	1.22
	T1R2	3.07	1.09	3.76	1.26	3.82	1.24
	T2R1	2.73	1.02	3.67	1.16	3.73	1.14
	T2R2	2.96	.98	3.80	1.13	3.87	1.11
	T3R1	3.01	.90	3.82	1.00	3.88	.98
	T3R2	3.16	.96	4.02	1.01	4.08	.98

Note: For each year, NE students were randomly sampled from all NE students.

For 2002, $N(ESL) = N(NE) = 323$, $N(ALL NE) = 28,956$.

For 2003, $N(ESL) = N(NE) = 425$, $N(ALL NE) = 28,462$.

For 2004, $N(ESL) = N(NE) = 357$, $N(ALL NE) = 25,664$.

years. The left column of the table shows the sources of variability and the remaining columns indicate both the magnitude of each variance component and the corresponding percentage of the total variance for the three years. As provided in Table 3, the results demonstrate that person within language group ($p:L$) was the largest variance component for 2002, 2003, and 2004 (31.6%,

Table 3

Variance components for the mixed effects $p:l \times t \times r'$ G-study designs ($N_{language} = 2$, $N_{task} = 3$, and $N_{rating} = 2$).

Source of variability	2002		2003		2004	
	$\hat{\sigma}^2(\alpha)$	%	$\hat{\sigma}^2(\alpha)$	%	$\hat{\sigma}^2(\alpha)$	%
L	0.40	29.91	0.45	28.90	0.36	23.13
$p:L$	0.42	31.57	0.59	38.44	0.62	40.03
t	0.02	1.40	0.01	0.91	0.02	1.05
r'	0.03	2.43	0.01	0.35	0.02	1.27
Lt	0.01	0.62	0.00	0.00	0.00	0.00
Lr'	0.00	0.00	0.00	0.00	0.00	0.05
$tp:L$	0.30	22.60	0.30	19.26	0.40	25.43
$r'p:L$	0.00	0.00	0.00	0.00	0.00	0.00
tr'	0.00	0.00	<0.01	0.12	0.00	0.06
Ltr'	<0.01	0.31	<0.01	0.17	0.00	0.30
$ur'p:L$	0.15	11.17	0.18	11.86	0.14	8.67
Total	1.34	100.00	1.55	100.00	1.56	100.00

38.4%, and 40.0% of the total variance, respectively), suggesting that the writing scores of students within each language group were the single largest source of variance. In other words, within their language group, students' writing scores differed greatly.

Language group (L) yielded the second largest variance component for 2002 and 2003, and the third for 2004 (29.9%, 28.9%, and 23.1% of the total variance, respectively), indicating there was a relatively large difference in writing performances that could be attributed to language group. The person within language-by-task ($tp:L$) interaction yielded the third largest variance component for 2002 and 2003, and the second for 2004 (22.6%, 19.3%, and 25.4% of the total variance, respectively), indicating that students within each language group performed very differently on different tasks. The residual yielded the fourth largest variance component for 2002, 2003, and 2004 (11.2%, 11.9%, and 8.7% of the total variance respectively). The residual contains the variability due to the interaction between ratings, tasks, students within language group along with other unexplained systematic and unsystematic sources of error. As expected, the variance component for ratings and the variance components for the interaction of ratings with the remaining facets in the design were close to zero.

4.3. G -study 2

G -study 2 was a $p \times t \times r'$ (person-by-task-by-rating) random effects G -study conducted separately for ESL and NE students. This G -study provided information for the corresponding D -study in order to examine the potential impact of G -coefficient differences on the rating designs for ESL students in comparison to NE students (i.e., how many independent ratings and writing tasks were needed in order for ESL students to have comparable G -coefficients as NE students). Table 4 presents the results of the G -studies for ESL and NE students for the three years. The upper section of the table displays the results for ESL students and the lower section displays the results for NE students. Each section shows the same information as shown in Table 3. For example, for

Table 4
Variance components for random effects $p \times t \times r'$ G -study designs ($N_{\text{language}} = 2$, $N_{\text{task}} = 3$, and $N_{\text{rating}} = 2$).

Language group	Source of variability	2002		2003		2004	
		$\delta^2(\alpha)$	%	$\delta^2(\alpha)$	%	σ^2	%
ESL	p	0.37	41.22	0.57	50.23	0.40	37.37
	t	0.01	1.35	0.01	1.30	0.02	1.55
	r'	0.03	3.08	0.00	0.34	0.03	2.72
	pt	0.30	32.59	0.30	26.40	0.42	39.77
	pr'	0.00	0.00	0.00	0.00	0.00	0.00
	tr'	<0.01	0.34	0.01	0.55	0.01	0.58
	ptr'	0.19	21.41	0.24	21.19	0.19	18.00
	Total	0.91	100.00	1.14	100.00	1.07	100.00
NE	p	0.47	49.28	0.62	58.48	0.85	64.30
	t	0.03	3.49	0.01	1.10	0.01	1.06
	r'	0.04	3.76	0.01	0.56	0.01	0.86
	pt	0.31	32.38	0.30	27.98	0.37	27.86
	pr'	0.00	0.00	0.00	0.00	0.00	0.00
	tr'	0.00	0.09	0.00	0.00	0.00	0.02
	ptr'	0.11	11.00	0.13	11.88	0.08	5.90
	Total	0.96	100.00	1.05	100.00	1.32	100.00

ESL students in 2002, the magnitude of variance component person (p) was 0.37 and it explained 41.2% of the total variance.

The results for ESL students presented in Table 4 indicate that person (p) yielded the largest variance component for 2002 and 2003, and the second largest for 2004 (41.2%, 50.2%, and 37.4% of the total variance, respectively), suggesting that ESL students differed in their writing scores. The person-by-task interaction (pt) yielded the second largest variance component for 2002 and 2003, and the largest for 2004 (32.6%, 26.4%, and 39.8% of the total variance, respectively), suggesting that the standing of ESL students varied from task to task. The residual yielded the third largest variance component for 2002, 2003, and 2004 (21.4%, 21.2%, and 18.0% of the total variance, respectively). In this case, the residual contains the variability due to the interaction between ratings, tasks, students, and other unexplained systematic and unsystematic sources of error. Again, as expected, the variance component for ratings and the variance components for the interaction of ratings with the remaining facets in the design were close to zero.

The results for NE students presented in Table 4 indicate that person (p) yielded the largest variance component for 2002, 2003, and 2004 (49.3%, 58.5%, and 64.3% of the total variance, respectively). The person-by-task interaction (pt) yielded the second largest variance component (32.4%, 28.0%, and 27.9% of the total variance, respectively). The residual yielded the third largest variance component (11.0%, 11.9%, and 5.9% of the total variance for 2002, 2003, and 2004, respectively). And again, the variance component for ratings and the variance components for the interaction of ratings with the remaining facets in the design were close to zero.

In summary, the results of both G -studies indicate that the greatest source of variation in students' writing performances was due to differences among students' English writing skills as measured by the writing tasks. This suggests that, as intended, the writing tasks did distinguish among students. While the variance due to task was minimal, the variance associated with the interaction of person by task was the second largest source of variance. These latter two findings suggest that, while writing tasks were, on average, comparable in difficulty, they were not uniformly difficult for all students in both language groups. The variance due to ratings was negligible, and the person-by-rating (pr') and task-by-rating (tr') effects were zero for both the ESL and NE students. These results are due to the low variance of the adjudicated scores used for analyses. However, the variance due to language group was large, showing that there was a large difference between the ratings of the writing performances of ESL and NE students. These differences may be due to one or a combination of the following three factors: (a) ESL students might have difficulty understanding the writing tasks; (b) rating bias against ESL students existed; and (c) ESL students were systematically different in writing abilities.

4.4. Comparative analyses on variance components

Based on the G -study 2 results, standard errors and confidence intervals of variance components were calculated to examine whether corresponding variance components were significantly different between language groups (ESL versus NE). Using formulas (1) and (2), estimates of the standard errors of the variance components for each language group were obtained. These standard errors were used to construct confidence intervals around the variance components. Using formulas (3) and (4), confidence intervals were created around the variance components for each language group. The results are presented in Table 5. The left column of the table indicates the G -study 2 variance components and the remaining columns show the 95% confidence intervals (lower limit and upper limit) on these variance components for ESL and NE students for the three years. The symbol "*" indicates significant differences between ESL and NE students where

Table 5
95% Confidence intervals on *G*-study variance components.

VC	2002		2003		2004	
	ESL	NE	ESL	NE	ESL	NE
<i>p</i>	(0.29, 0.45)	(0.38, 0.57)	(0.48, 0.67)	(0.52, 0.72)	(0.31, 0.48)*	(0.70, 0.99)*
<i>t</i>	(-0.01, 0.03)	(-0.02, 0.08)	(-0.01, 0.04)	(-0.01, 0.03)	(-0.01, 0.05)	(-0.01, 0.04)
<i>r'</i>	(-0.02, 0.07)	(-0.02, 0.09)	(-0.01, 0.01)	(0.00, 0.02)	(-0.02, 0.08)	(-0.01, 0.03)
<i>pt</i>	(0.25, 0.34)	(0.27, 0.35)	(0.26, 0.34)	(0.26, 0.33)	(0.37, 0.48)	(0.33, 0.41)
<i>pt'</i>	(-0.01, 0.01)	(-0.01, 0.01)	(-0.01, 0.01)	(-0.01, 0.01)	(-0.01, 0.01)	(0.00, 0.01)
<i>tr'</i>	(0.00, 0.01)	(0.00, 0.003)	(0.00, 0.02)	(0.00, 0.001)	(0.00, 0.02)	(0.00, 0.001)
<i>ptr'</i>	(0.17, 0.22)*	(0.09, 0.12)*	(0.22, 0.26)*	(0.11, 0.14)*	(0.17, 0.21)*	(0.07, 0.09)*

* Significant difference at .05-level.

significance is defined as non-overlapping confidence intervals. For example, for the 2004 June administration, the 95% confidence intervals on person (*p*) were from 0.31 to 0.48 for ESL students and from 0.70 to 0.99 for NE students and did not overlap; hence the variance components of person (*p*) were significantly different between ESL and NE students.

The results show that, for 2002, 2003, and 2004, the 95% confidence intervals on the residual variance components (*ptr'*) did not overlap, indicating that the residual variance components were significantly different between the ESL and NE students. Except for the variance components for person (*p*) in 2004, the 95% confidence intervals on the remaining variance components overlapped, indicating that these variance components were not significantly different between ESL and NE students.

4.5. Calculation of *G*-coefficients

Using formula (5) and the *G*-study 2 variance component results, *G*-coefficients for both ESL and NE students were calculated for each administration. The results are presented in Table 6. As shown, the *G*-coefficients for ESL students for the current three-task-and-two-rating scenario were .74, .80, and .70 for 2002, 2003, and 2004, respectively, while the corresponding *G*-coefficients for NE students were .80, .84, and .86.

4.6. Comparisons of *G*-coefficients

Using formulas (6)–(10), the differences in *G*-coefficients between ESL and NE students were compared to determine if they were statistically different at the .05-level of significance. The

Table 6
Summary of *G*-coefficients.

Year	<i>G</i> -Coefficient	
	ESL	NE
2002	.74	.80
2003	.80	.84
2004	.70	.86

Table 7
F-test results on G-coefficients.

Year	Group	$\hat{\rho}^2$	df	F	F critical value
2002	ESL	.74	215.56 (v_1)	0.79	0.77
	NE	.80	214.76 (v_2)		
2003	ESL	.80	283.13 (v_1)	0.82	0.79
	NE	.84	283.10 (v_2)		
2004	ESL	.70	78.00 (v_1)	0.45*	0.73
	NE	.86	709.97 (v_2)		

* $p < 0.05$.

results are presented in Table 7. For each year, the table shows the G-coefficients ($\hat{\rho}^2$), degrees of freedom (*df*), the F values, and the critical F values. If the F value is below the critical F value, then the two G-coefficients are significantly ($p < .05$) different. For example, for June 2004 administration, the G-coefficient was .70 for ESL and .86 for NE students; these two G-coefficients were significantly ($p < .05$) different because the F value ($F = 0.45$) was below the critical F value ($F \text{ Critical} = 0.73$).

Table 7 shows that the G-coefficients for ESL and NE students in the June 2002 and June 2003 administrations were not significantly different at the .05 significance level. In contrast, the G-coefficients for ESL and NE students in the June 2004 administration were significantly different ($p < .05$). The value of the G-coefficient for NE students was statistically greater than that for ESL students. Given that the G-coefficients were not significantly different for 2002 and 2003, there was no need to alter the rating designs for ESL students in those years. Therefore, a separate D-study analysis was conducted for ESL and NE students for only the June 2004 administration.

4.7. D-study

The D-study was a $p \times T \times R'$ (person-by-task-by-rating) random effects D-study and its purpose was to determine the number of independent ratings and/or writing tasks needed in order for the generalizability of ratings for the ESL students to be equal to the current generalizability of ratings for the NE students for 2004. The G-coefficients for the ESL and NE students for 2004 are summarized in Table 8. The numbers in the three rows indicate the number of tasks (N_t), the number of ratings (N_r), and the G-coefficients ($\hat{\rho}^2$), respectively. For example, the current three-task-and-two-rating design for NE students had a G-coefficient of .86.

For 2004, Table 8 shows that the current three-task-and-two-rating design for ESL students had a G-coefficient of .70. The number of writing tasks would need to be increased to seven and the number of ratings increased to four in order for ESL students to obtain the same value of

Table 8
A comparison of G-coefficients for the June 2004 administration.

	NE		ESL													
N_t	3	3	3	3	4	4	5	5	5	6	6	6	7	7	7	
N_r	2	2	3	4	3	4	3	4	5	3	4	5	3	4	5	
$\hat{\rho}^2$.86	.70	.71	.72	.77	.77	.80	.81	.81	.83	.84	.84	.85	.86	.86	

.86 as NE students. Naturally, such a change could not be implemented but this comparison did illustrate the relative severity of the difference in the *G*-coefficients.

5. Discussion and conclusions

The first research sub-question was about the differences in score variation between ESL and NE students. The results showed that differences in score variation did exist between ESL and NE students when adjudicated scores were used. First, there was a large effect for both language group and person within language-by-task interaction. Previous research has indicated that there is little consistency amongst diverse tasks (Lee et al., 2002). These findings suggest that ESL and NE students had unequal performance across tasks. Together with the descriptive results and the large variance component for language group, the ESL students had consistently lower performance. Second, the residual variance component was significantly ($p < .05$) larger for ESL students than for NE students in all three years. The residual is unwanted variance. This important difference indicated that there was consistently more unexplained variance in ESL students' writing scores than in NE students' writing scores in all three years. Finally, the desired variance associated with the object of measurement was significantly ($p < .05$) smaller for ESL students than for NE students in 2004. These differences likely decreased the rating reliability for ESL students in that year. However, as shown in the descriptive statistics, the standard deviations for the ratings of ESL students were consistently lower than those of the NE students in 2004 and this may have also affected the comparison of ESL and NE samples.

The second and third research sub-questions focused on the differences in the reliability of the writing scores assigned to ESL and NE students and the impact of these differences on the rating designs for ESL students in comparison to NE students. As mentioned above, there was a significant ($p < .05$) difference in the *G*-coefficients for ESL and NE students in 2004 and this difference had considerable impact on the rating designs for ESL students. In that year, even with the scores after adjudication, ESL students had a significantly ($p < .05$) lower *G*-coefficient and it seemed that they would never feasibly have a comparable *G*-coefficient to NE students.

Together, the differences in 2004 in terms of the desired and unwanted variations between the ESL and NE writing scores and the lower *G*-coefficient for ESL students raise a potential question about the fairness of the writing scores assigned to ESL students in this examination. If the ratings of ESL and NE students are not equally reliable, then fairness may become a concern because there should be no significant differences in the rating variability and reliability of scores assigned to ESL and NE students (Johnson et al., 2000). Therefore, reliability is and should remain a paramount concern in quality performance assessment such as writing. Further, the differences in accuracy and precision may be due to factors outside of the writing skills of ESL students. For example, as previously mentioned, ESL students might have difficulty understanding the writing tasks; they might also be different in writing abilities; or rating bias against ESL students might exist. The former two explanations support the argument of construct validity made by Angoff and Sharon (1971) that language matters in a language-laden test that is applied to ESL students. To sum up, the point here is not that the ESL students do worse or better on these examinations, or that they do worse or better depending on interaction with other *G*-theory facets. The point is that the findings for ESL and NE do not compare in terms of reliability, and without such comparability, how can the examination be seen as fair, regardless if its results seem predictable? However, further exploration of these issues is needed. Thus it would be important to conduct further research on these results to determine if the differences were due to fairness issues or actual systematic differences in English writing skills.

The present study was limited in the following two ways. First, as mentioned above, due to limited data availability, the adjudicated rather than the initial raw scores awarded by the two raters were used; their scores could vary by no more than one score point, thereby leading to spuriously high levels of agreement. However, the adjudicated scores were the only data that could be used for the analyses. Further, as previously mentioned, the adjudicated scores reflect the real-world assessment situations; they are actually used to report large-scale assessment results, although they may mask a much greater uncertainty among ratings than the original scores.

Second, the use of “rating” instead of “rater” might have produced an impact on the results of this study. The assumption of treating “rating” instead of “rater” as a facet of analysis is that raters are by and large comparable in terms of severity. But in large-scale writing assessment contexts, this assumption may not be met. Although raters received intensive training prior to marking, differences in rater severity cannot be avoided when large numbers of raters are involved (McMillan, 2000; Welch & Miller, 1995).

In light of the limitations, the following three conclusions were reached. First, given that the testing program has a desire to maintain high inter-rater agreement, adjudication, not unexpectedly, will increase this agreement. Based on the results of the study, adjudicated scores also have a much lower person by rating variance component.

Second, in such high-stakes provincial examinations, the nature of the writing tasks matters. The use of separate unrelated tasks instead of thematically related tasks can lead to performance differences. For example, there was a large person-by-task effect for both ESL and NE students when separate unrelated tasks are used. As found in previous research, varied and unrelated tasks on the writing examination resulted in students’ differential difficulties in responding to these tasks (Lee et al., 2002).

Finally, there is still unexplained variability. While not the largest component, the residual contains the variability due to the interaction between ratings, tasks, students, and other unexplained systematic and unsystematic sources of error. Residual effects can indicate hidden facets (Brennan, 2001). For example, “gender” was not considered in this study, yet research has shown that females outperform males in writing (Henderson, 1999; Willingham & Cole, 1997). The variance of the hidden facets is included in the residual variance, thus leading to a larger residual than when the facet is explicitly considered.

Overall, the study provides initial evidence that the ratings of ESL and NE students’ writing result in differences in terms of consistency and precision. Further examination is required to determine the extent to which these consistency differences affect the fairness and accuracy of the assessment of ESL students’ English writing skills and to find ways to alleviate these differences.

This study was intended to examine the consistency of the rating of ESL students’ writing on large-scale assessments and the potential impact this would have on fairness. Fairness is a priority in the field of educational assessment in Canada (Joint Advisory Committee, 1993). Fairness issues in ESL writing assessments will become increasingly important because of the significant growth in the number of ESL students being educated in Canadian schools. Not unexpectedly, ESL students face considerable challenges passing writing examinations due to their linguistic deficiencies (Blackett, 2002). However, these deficiencies are compounded by cultural factors that are supposedly irrelevant to the writing skills being assessed (Yang, 2001). Further, technical difficulties in ESL students’ writing or differences in writing styles may mask relatively strong organizational and conceptual writing skills (Hinkel, 2003; Yang, 2001). Raters who are unable to differentiate amongst these different skills or who use the rating scales differently when scoring ESL students will create less consistent and reliable assessments of ESL students’ writing.

Similarly, scoring scales that do not provide mechanisms to address these potential differences in technical and conceptual skills will also result in less consistent scores. Therefore, the results of this study have implications from both a policy perspective and a measurement perspective. From a policy perspective, there must be a process in place to address the discrepant ratings of ESL students' writing. From a measurement perspective, attention needs to be paid to ensuring the accuracy and consistency of the scores assigned to ESL students' writing.

Based on the findings of this study, the following four recommendations are proposed in order to minimize the impact of factors that may reduce the consistency of the ratings of ESL students' English writing. The first recommendation is to use adjudication. Adjudication can reduce the discrepancy among the ratings of students' writing. Given the potential for increased discrepancies amongst the writing scores assigned to ESL students, adjudication will provide more consistent scores.

Second, care needs to be paid to the nature of the writing tasks students are asked to respond to in such provincial examinations. The use of unrelated writing tasks results in less consistent student performance across tasks. While important to ensure that students demonstrate the expected breadth of writing skills, this reduced consistency is also a source of error. The types of writing tasks can affect the rating of ESL students' writing (Weigle, 1999). It will be important to determine if such varied tasks differentially and unexpectedly affect the writing performance or scoring of ESL students as opposed to NE students. Such varied tasks might differentially jeopardize the reliability of a composite score based on a set of different tasks for ESL students (Lee et al., 2002).

Third, although the province has strict procedures for rater training, scoring, and reliability reviews, there is currently no differential treatment of ESL writing samples or specific training in the rating of ESL students' writing. Research shows that the scores that raters assign to ESL writing tasks may fluctuate due to many factors; for example, the scoring methods used (holistic versus analytic), the differences in raters' application of scoring criteria, and the differences in raters' linguistic and professional backgrounds (Russikoff, 1995; Sakyi, 2000; Song & Caruso, 1996). Rater training, however, can minimize the differences caused by these factors (Reid & O'Brien, 1981) and modify raters' expectations of good writing by clarifying for the raters both the task demands and writer characteristics (Huot, 1990). Further, rater training is especially effective for inexperienced raters of ESL students' writing to minimize the differences in ratings (Weigle, 1994). One solution is to include ESL papers as exemplar papers for rater training, and for marker agreement or reliability review papers during the monitoring process. It would be worth exploring how rater training could better support the rating of ESL papers. Further, the scoring of ESL papers needs to be carefully monitored to determine if such inconsistencies with initial ratings continue to affect the reliability with the adjudicated rating. Lastly, there is a need to review ESL papers in relation to the rating scales. By doing that, we can try to make raters equally use the rating scales to rate ESL students' writing.

Finally, all rater information on each student's writing samples should be tracked. This information is important for both monitoring ratings and conducting reliability reviews that can be used to inform what changes might be warranted in the initial and final scoring processes to ensure fair assessment of all students.

In total, these recommendations address the issue of divergent ratings of ESL students' writing. They will help promote rating accuracy and consistency, and ensure that ESL papers are rated in the same way as NE papers. Thus the aspects of fairness that require that all examinees are treated fairly during the testing process itself will be addressed (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999; Brown, 1996).

Acknowledgements

I would like to acknowledge my gratitude and register my sincere thanks to Dr. Don A. Klinger, Dr. Nancy L. Hutchinson at Queen's University and Dr. Todd Rogers at the University of Alberta for their valuable advice and guidance as I conducted this study.

References

- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Angoff, W. H., & Sharon, A. T. (1971). A comparison of scores earned on the Test of English as a Foreign Language by native American college students and foreign applicants to U.S. colleges. *TESOL Quarterly*, 5 (2), 129–136.
- Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, 17 (1), 1–42.
- Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgment in a performance test of foreign speaking. *Language Testing*, 12, 239–257.
- Blackett, K. (2002) Ontario schools losing English as a second language programs – despite an increase in immigration. Available online at: www.peopleforaction.com/releases/2003/Oct24.02.
- Brennan, R. L. (1992). *Elements of generalizability theory*. Iowa City, IA: ACT.
- Brennan, R. L. (2001). *Statistics for social science and public policy: Generalizability theory*. New York: Springer-Verlag.
- Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall.
- Canadian Bureau for International Education. (2002, April 15). *International student numbers hit record high, but Canada offers dwindling support for African students*. Retrieved October 28, 2002, from http://www.cbie.ca/news/index_e.cfm?folder=releases&page=rel_2002-04-15_e.
- 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). 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.
- Feldt, L. S. (1965). The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty. *Psychometrika*, 30, 357–370.
- Feldt, L. S. (1969). A test of the hypothesis that Cronbach's alpha or Kuder–Richardson coefficient twenty is the same for two tests. *Psychometrika*, 34, 363–373.
- Hakstian, A. R., & Whalen, T. E. (1976). A k-sample significance test for independent alpha coefficients. *Psychometrika*, 41, 219–231.
- Hamp-Lyons, L. (1991). Issues and directions in assessing second language writing in academic contexts. In: L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 323–329). Norwood, NJ: Ablex.
- Hamp-Lyons, L. (1996). The challenges of second language writing assessment. In: E. White, W. Lutz, & S. Kamusikiri (Eds.), *Assessment of writing: Policies, politics, practice* (pp. 226–240). New York: Modern Language Association.
- Henderson, D. L. (1999). *Investigation of differential item functioning in exit examinations across item format and subject area*. Unpublished doctoral dissertation. Edmonton, AB: University of Alberta.
- Hinkel, E. (2003). Simplicity without elegance: Features of sentences in L1 and L2 academic texts. *TESOL Quarterly*, 37, 275–301.
- Huot, B. A. (1990). Reliability, validity, and holistic rating: What we know and what we need to know. *College Composition and Communication*, 41, 201–213.
- Johnson, R. L., Penny, J., & Gordon, B. (2000). The relation between score resolution methods and interrater reliability: An empirical study of an analytic rating rubric. *Applied Measurement in Education*, 13 (2), 121–138.
- Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Edmonton: AB.
- Kristof, W. (1963). The statistical theory of stepped-up reliability coefficients when a test has been divided into several equivalent parts. *Psychometrika*, 28, 221–238.

- Lee, Y., & Mollaun, P. (2002). *Score dependability of the writing and speaking sections of new TOEFL*. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.
- Lee, Y., Kantor, R., & Mollaun, P. (2002). *Score dependability of the writing and speaking sections of new TOEFL*. Paper presented at the annual meeting of National Council on Measurement in Education, New Orleans, LA.
- Lohninger, H. (2002). Vienna, Austria: Vienna University of Technology.
- MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *Journal of Experimental Education*, 68 (2), 167–190.
- Moss, P. A. (1994). Can there be validity without reliability? *Educational Researcher*, 23, 5–12.
- Popham, W. J. (1990). *Modern educational measurement: A practitioner's perspective* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
- Reid, J., & O'Brien, M. (1981). *The application of holistic grading in an ESL writing program*. Paper presented at the annual convention of Teachers of English to Speakers of Other Languages, Detroit, MI. ERIC Document Reproduction Service no. ED 221 044.
- 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.
- Sakyl, 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: Cambridge University Press.
- Santos, T. (1988). Professors' reactions to the writing of nonnative-speaking students. *TESOL Quarterly*, 22 (1), 69–90.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.
- Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of Educational Measurement*, 30, 215–232.
- 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 (2), 163–182.
- Speck, B. W., & Jones, T. R. (1998). Direction in the grading of writing? In: F. Zak & C. C. Weaver (Eds.), *The theory and practice of grading: Problems and possibilities* (pp. 17–29). Albany: SUNY Press.
- Thompson, R. (1990). Writing-proficiency tests and remediation: Some cultural differences. *TESOL Quarterly*, 24, 99–102.
- Vaughan, C. (1991). Holistic assessment: What goes on in the raters' minds? In: L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 111–126). Norwood, NJ: Ablex.
- Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, 11, 197–223.
- Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing*, 6 (2), 145–178.
- Welch, C., & Miller, T. (1995). Assessing differential item functioning in direct writing assessments: Problems and an example. *Journal of Educational Measurement*, 32, 163–178.
- Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco, CA: Jossey-Bass.
- Willingham, W. W., & Cole, N. S. (1997). Fairness issues in test design and use. In: W. W. Willingham & N. S. Cole (Eds.), *Gender and fair assessment* (pp. 227–346). Hillsdale, NJ: Lawrence Erlbaum.
- Yang, Y. (2001). *Chinese interference in English writing: Cultural and linguistic differences*. ERIC Document Reproduction Service no. ED 461 992.

Jinyan Huang (Ph.D.) is an assistant professor in TESOL and Assessment at the College of Education in Niagara University. His areas of research center on: (a) ESL students' learning challenges and coping strategies; (b) factors that affect ESL students' learning outcomes; and (c) reliability, validity, and fairness issues of ESL assessments.