Applied Research on English Language

V. 6 N. 4 2017
pp: 411-434
http://are.ui.ac.ir
DOI: http://dx.doi.org/10.22108/are.2017.106097.1170

Score Generalizability of Writing Assessment: The Effect of Rater's Gender

Rahil Sheibani [1]*, Alireza Ahmadi [2]

[1] *Ph.D. Candidate, Department of Foreign Languages, Qeshm Island Branch, Islamic Azad University, Qeshm, Iran*

[2] *Associate Professor of TEFL, Department of Foreign Languages and Linguistics, Shiraz University, Shiraz, Iran*

Received: 2017/08/25 **Accepted: 2017/11/06**

Abstract: The score reliability of language performance tests has attracted increasing interest. Classical Test Theory cannot examine multiple sources of measurement error. Generalizability theory extends Classical Test Theory to provide a practical framework to identify and estimate multiple factors contributing to the total variance of measurement. Generalizability theory by using analysis of variance divides variances into their corresponding sources, and finds their interactions. This study used generalizability theory as a theoretical framework to investigate the effect of raters' gender on the assessment of EFL students' writing. Thirty Iranian university students participated in the study. They were asked to write on an independent task and an integrated task. The essays were holistically scored by 14 raters. A rater training session was held prior to scoring the writing samples. The data were analyzed using GENOVA software program. The results indicated that the male raters' scores were as reliable as those of the female raters for both writing tasks. Large rater variance component revealed the low score generalizability in case of using one rater. The implications of the results in the educational assessment are elaborated.

Keywords: Writing Assessment, Independent Task, Integrated Task, Raters' Gender, Generalizability Theory

* **Corresponding Author.**

Authors' Email Address:
[1] **rahil sheibani** (rahil.sheibani@yahoo.com), [2] **Alireza Ahmadi** (ar.ahmadi55@gmail.com),

ISSN (Online): 2322-5343**, ISSN (Print):** 2252-0198

Introduction

Writing fluently and expressively is a difficult ability to acquire for all language users. It is a complex intellectual activity in which writers need to boost their mental power, improve their analytical abilities, and make precise and valid distinctions (National Commission on Writing in America's Schools and Colleges, 2003). It is not a usual activity, meaning that all physically and mentally normal people learn to speak a language while they have to be taught how to write. As Nunan (1996) declares, all children, except those with physiological disabilities, can understand and speak their mother tongue while all of them cannot read and fewer learn to write fluently and legibly.

Writing is considered the most complex and challenging skill since it involves "originating and creating a unique verbal product that is graphically recorded" (Emig, 1997, p. 8) and requires writers to try consciously and practice in composing, developing, and analyzing ideas. As a complex activity, it is obvious that L2 learners usually face problems when developing their writing (Evans, Hartshorn, McCollum & Wolfersberger, 2010). Recently, good writing has become a growing pedagogic demand which is "critical to academic and professional success and leads to good grades, admission into college, exit from college, a good job, and upward professional movement" (Ghalib & Al-Hattami, 2015, p. 225). Consequently, writing continues to be one of the most challenging skills for learners to perform and expectedly for teachers to assess. The purposes of writing ability are not limited to awarding degrees, certifying proficiency, testing qualification for a profession, placing students into appropriate program, and allowing subjects to exit a program; results may have important consequences which can significantly influence a test taker's life.

Using writing assignments in a course is both "resource and labor intensive" for instructors, and the grading of written assessments is prone to subjectivity and uneven variability (Anatol & Hariharan, 2009; Bell, 1980). ESL/EFL research has shown that writing assessment is a complex and challenging issue (Barkaoui, 2008; Huang, 2011; Huang & Foote, 2010; Huang & Han, 2013; Sakyi, 2000) since different factors may affect writing. The sources of error affecting the reliability of written compositions include the student, the scoring method, raters' professional background, gender, experience, rating scales, the physical environment, the design of the items, and the test itself and even the methods and amount of training (Barkaoui, 2008; Brown, 2010; Cumming, Kantor & Powers, 2001; Huang, 2007, 2009, 2011;

Huang & Han, 2013; Mousavi, 2007; Shohamy, Gordon & Kraemer, 1992; Weigle, 1994, 1999, 2002).

Since written assignments become an essential component of higher education, it is imperative that faculty, teachers, and instructors try to minimize deficiencies in grading essays and to maximize the probability of attaining highly reliable scores. Consequently, language productive skill assessment requires to be both reliable and valid.

The present trend of students' ability and performance assessment in educational systems, in general, and language education, in particular, is that students take part in a single test at the end of a course and their final grades on this examination will classify the student as qualified to pass the course or not. In Bachman's (1990) words, "if we could obtain measures for an individual under all the different conditions specified in the universe of possible measures, his average score on these measures might be considered the best indicator of his ability"(p. 191).

Among various issues of second-language assessment, reliability has been of paramount importance to both language testers and teachers. Reliability is a fundamental concept in language testing; reliable results are critical for a test to serve its intended purpose (Fulcher, 1999). Jones (1979) considers consistency of scores as an influential factor in reliability of performance tests. McNamara (1996) also claims that, since human raters are subjective in assessing a particular test, the issue of reliability is so complicated. Hamp-Lyons (1990) considers rater's attitudes and conceptions as influential factors in reliability of rating.

McNamara (1996) states that due to the presence of raters, traditional assessment in language assessment context is different from performance assessment. In traditional assessment scores are elicited mainly based on the interaction of test-takers and the task while in performance assessment, rater related variables directly affect the results. Consequently, in addition to the increasing predominance of writing performance assessment in both large-scale and classroom assessment, the raters and what they do during rating have attracted researchers' attention.

Weigle (2002) maintains that rater variability is demonstrated in various ways including severity effect, halo effect, central consistency effect and bias effect. Landy and Farr (1983) define severity effect as raters' differences in rating in terms of their overall severity. Halo effect implies that raters may be unable to make distinction between conceptually distinct constructs as such they may instead rate on the basis of their general impression. Central tendency is the next rater variable which occurs when raters avoid the extreme ratings and have

a tendency for the midpoint of the scale. Finally, biased raters tend to rate unusually harshly or leniently on one aspect of the rating situation (Schaefer, 2008).

Another source of variability is rater background which includes factors such as rater's native language, experience, and training (Barkaoui, 2011; Johnson & Lim, 2009; Lumley, 2002). Since the rater is the key factor in the process of rating, improving raters' subjectivity leads to the improvement of both assessment validity and the raters' stance in assessment.

Raters' gender is also among the effective factors which may positively or negatively affect test takers' score. Some studies have inspected the probable effect of raters' gender in the rating of oral interviews. Mostly, these studies indicated some effect of gender on scores although the effect was not always similar. Some of them revealed that male interviewers scored higher than the female ones (e.g., Lock, 1984; Porter 1991a; 1991b) while others revealed female interviewers awarded higher ratings (e.g., Porter & Shen, 1991; O'Sullivan, 2000).

Generalizability Theory

Generalizability (G)- theory "is a statistical theory for evaluating the dependability (or reliability) of behavioral measurements" (Shavelson, R. J., & Webb, 2005, P. 599). It is useful in identifying and measuring the magnitude of the potentially important sources of measurement error. Furthermore, it identifies multiple influencing factors on measurement variance and classified variances into different sources including systematic variance among the measurement objects, multiple sources of error, and their interactions; it can also pinpoint the extent to which each variable affects the true score (Shavelson & Webb, 1991). G-theory is advantageous in separately estimating multiple sources of error in a single analysis. G-theory helps researchers to specify the number of required occasions, tasks, raters, and administrations to obtain dependable scores. It can provide coefficient which indicates the level of dependability and a generalizability coefficient which is similar to reliability coefficient in Classical Test Theory (CTT) (Brennan, 2001a; Shavelson & Webb, 1991).

G-theory, as a "psychometric theory" (Brennan, 2001a), is potentially very useful in many areas of research suffering from inconsistency of measurement. In particular, this theory is a suitable approach to deal with the multifacetedness of the assessment of writing ability from written compositions and the different sources of 'measurement error' in assessment including rater, topic, and task facets. To identify different sources of measurement error, G-theory connects the two main psychometric parameters, reliability and validity, and estimates the variation in scores due to each person, each facet, and their combinations.

Stability and the reliability of writing assessment may be influenced by various factors including the attitude towards rating and raters' consistency, the physical environment, the design of the items, marking rubrics, rater training, topics students write about, the discourse mode, and the test time limit (Schoonen, 2005).

G-theory Studies on Writing Assessment

Recently, several studies utilized G-theory to inspect the reliability and validity of EFL/ESL writing scores and to explore the relative effect of different facets (raters, tasks, rating scale, etc.) on the accuracy of writing scores.

Swartz et al. (1999) investigated the reliability of holistic and analytic writing scores and raters' effect and the type of decision (absolute versus relative) on writing score reliability. The results revealed that decreasing the number of raters leads to the reduction of reliability coefficients of writing scores especially in case of absolute decisions rather than relative decisions.

Some studies have focused on comparing the accuracy and validity of ESL and Native English (NE) writing scores (e.g., Huang & Foote, 2010; Huang, 2008, 2012). Huang (2008) using G-theory to investigate rating variability and reliability of scores given to ESL and Native English essays. He conducted a series of generalizability (G)- studies and decision (D)-studies to explore the differences in score variation between ESL and NE writing samples. The results showed that the writing scores assigned to NE essays were more consistent than those awarded to ESL essays.

In a similar study, Huang and Foote (2010) tried to determine score variations between ESL and NE writing sample in a small-scale assessment context. The findings indicated scores assigned to ESL and NE papers were different in consistency and precision. The obtained results highlighted the necessity of fairness in writing scoring.

In another research, Huang (2012) assessed the accuracy and validity of ESL students' writing scores in provincial English examinations. The findings showed ESL students' writing scores had significantly lower generalizability (G)-coefficients compared to NE students. Besides, ESL students' scores had significantly lower convergent validity in one year and lower discriminant validity in all three years than that of NE students. In sum, this study revealed that these two groups received significantly different scores in accuracy and construct validity which provides evidence about presence of rating bias in assessing ESL students' writing.

Other researchers (Brown, 1991; Shi, 2001) turned their focus from NE and ESL students' writing scores to NE and non-native (NNE) raters. Brown's study (1991) looked into the raters' academic background while Shi (2001) worked on the language background of raters: native English-speaking and nonnative English speaking. Brown investigated the criteria ESL (those who teach ESL students) and English instructors (who teach native English students) used to score essays. The results indicated that these groups had no significant differences in how they rated the writing samples. For example, both groups of raters considered the content as an important positive feature, and syntax as an important negative feature.

In a similar study, Shi (2001) investigated rating behavior of NES and NNS raters in Korea. The findings confirmed NNS raters' inferiority in measuring components of language. Moreover, the NNS raters were stricter toward grammar, sentence structure, and organization, while the NES raters were more severe in scoring content and overall scores. Furthermore, analysis of the raters' responses revealed that for NNS raters, content and grammar were the most difficult features to score while NES raters considered content as the most difficult feature.

Research has focused on raters' effect on score generalizability in writing assessment. Utilizing G-theory, Lane and Sabers (1989) studied the reliability of ratings for a single writing prompt. They randomly selected fifteen essays from grades three and eight. Eight raters with different professional backgrounds rated the samples on different dimensions including ideas, development and organization, sentence structure, and mechanics. The findings revealed that an increase in the number of raters from one to four can lead to an increase in the generalizability coefficient. Besides, the contribution of the rater effect was slightly smaller than the examinees' effects and the interaction between persons and tasks.

Other researchers have focused on rater training effect on score consistency (Shohamy, Gordon, & Kraemer, 1992; Weigle, 1994, 1998). For instance, Weigle (1998) found that rater training decreased rater severity and inconsistency. Although, she remarks that rater training cannot change the raters into becoming very similar or the same. Shohamy et al. (1992) also confirmed the effect of rater training session on consistency of raters.

Han (2013) used G-theory to explore the impact of rater training and scoring methods on the writing scores coming from classroom-based assessment. The results revealed that by training raters, the holistic scoring could produce as consistent and reliable scores as analytic scoring.

The findings from studies on score generalizability in writing are inconclusive, possibly because of the diversity of research contexts and a wide variety of factors (Gebril, 2009). Also the findings of research on the effect of gender on scoring writing are inconsistent. As such, more recent thinking about the issue of raters' gender in the case of performance assessment, for example in rating IELTS writing seems to be required. Male and female raters may assess writing differently; therefore, this study explored the interactions between raters' gender and task as an area that has received scanty research in the past. G-theory enjoys a more powerful theoretical framework. Furthermore, it has been increasingly employed in studies conducted on L2 writing assessment. As such, it was used as the framework of this study.

Research Questions

Specifically, the following research questions guided this research:

1. What are sources of score variation contributing to the scores assigned by male and female raters to writing papers?

2. Does the score reliability improve by increasing the number of male and female raters from one to four?

Method

Participants

Thirty Iranian university students participated in this study. They were majoring in English Language and Literature in the English department at Shiraz University, Iran. These essay writers were both male (46.67%) and female (53.33%) and ranged in age from 21 to 25. They were all non-native speakers of English and came from the same native language background: Persian. At the time of the study, they had taken the class on advanced writing skills and had the competence to be able to write an essay. Since the purpose of this study was exploring the performance of raters, examining the exact level of participants' proficiency level was not required.

Raters and Rating

Raters included 14 (7 male and 7 female) experienced writing teachers with 15 to 30 years of teaching experience and considerable experience in scoring exams. All these raters were EFL teachers who were required to rate writing samples as part of their teaching activity. They came from the same linguistic background and were proficient non-native speakers of English.

A rater training session was held to explain the rating purposes and procedures and to increase rating accuracy and rater agreement; through this session common rating problems were discussed to avoid rating bias. In this session, the raters were given a packet including two writing rubrics, writing prompts and 60 writing samples. Fourteen raters scored 30 EFL essays based on integrated rubric and another 30 essays based on an independent rubric. Each rater scored the essays independently only once.

Two rubrics were employed to rate the writing samples: one for the independent task and one for the integrated task. Both rubrics include a scale ranging from 0 to 9 with 9 as the highest score an examinee can obtain (British Council, IDP: IELTS Australia and University of Cambridge ESOL Examinations, 2005). The response to the integrated and independent tasks for both the Academic and General Training Modules was scored based on the following dimensions:

- Task Achievement (for integrated task) Task Response (for independent task)
- Coherence and Cohesion
- Lexical Resource
- Grammatical Range and Accuracy.

Instruments

The scripts consisted of 60 essays written by the 30 essay writers (each essay writer completed 2 tasks) in response to an independent task and an integrated, timed writing task.

Two writing tasks were selected from IELTS prompts. One of the selected prompts was used to represent the independent category and the other prompt was used with the integrated task (McCarter, 2002, see the appendix). Required instructions related to how essay writers should use the source text were presented in integrated task, a table which summarized the required information was presented as the source text.

The samples were collected on two occasions of mid-term and final exam. On each occasion, the assignment and the instructions were standard for all essay writers. They had to spend 20 minutes for Task 1 and 40 minutes for Task 2.

Data Collection Procedure

A group of 30 students completed the integrated and independent tasks as their mid-term and final exams, respectively. In writing task 1, the integrated task, the essay writers were asked to write a report for a university lecturer describing the information presented in a table. They had to write in an academic or semi-formal/neutral style and spend no more than 20 minutes on this

task. Besides, they were asked to write at least 150 words. In writing task 2, the independent task, the essay writers were given a topic to write about. They were asked to provide a full and relevant response. They had 40 minutes to write at least 250 words on this task. The samples were typed by the researcher, and the essay writers' names were replaced by numbers as ID and written on their sample to protect the anonymity of the essay.

Data Analysis Procedure

G-theory analysis was used to examine the effect of tasks, raters, and gender. Sixty samples of writing from 30 TEFL students were obtained in the condition mentioned above. Each sample was scored by 14 male and female trained raters. So, the study used a nested design for raters' gender (p × t × r: g). In this design, the essay writers were considered as crossed by raters. The number of essay writers and raters were constant. Therefore, this study also extended the principles of G-theory to balanced designs.

All analyses were done with the GENOVA program, version 3.1, for the Windows computer (Brennan, 2001b). The study considered both relative and absolute decisions which helped the researcher to rank the essay writers relative to each other and to make decisions about their standing in relation to a standard of performance.

Results

Descriptive statistics is provided in Table 1, including number (N), mean, and Std. Deviation for the total scores assigned to the integrated and independent tasks by raters.

Table 1. Descriptive Statistics for the Integrated and Independent Writing Tasks

		Integrated task		Independent task	
Raters	**gender**	**Mean**	**Std. Deviation**	**Mean**	**Std. Deviation**
Rater 1	male	5.0500	1.41025	6.4500	1.51059
Rater 2	male	6.9000	1.21343	6.2667	1.25762
Rater 3	male	5.1667	.74664	5.7500	.71619
Rater 4	male	6.0667	1.31131	5.4000	1.03724
Rater 5	male	5.3833	1.36257	5.3167	1.51705
Rater 6	male	5.5167	1.00416	5.5667	1.21580
Rater 7	male	6.5833	.47495	6.5333	.52413
Rater 8	female	8.4500	.66111	8.7167	.40860
Rater 9	female	7.6333	1.06620	7.7333	1.28475
Rater 10	female	7.8667	1.13664	7.1333	1.33864
Rater 11	female	6.5000	1.52564	7.9667	1.06620
Rater 12	female	5.1667	1.42232	6.0833	1.05114
Rater 13	female	6.1333	2.12916	5.6667	1.76817
Rater 14	female	4.2000	.84690	5.6667	1.66782
Total		**6.1869**	**1.68850**	**6.4464**	**1.58682**

According to Table 1, the essay writers' performance is similar in both tasks. However, the independent task received higher mean score (mean = 6. 4464). The highest and lowest mean scores on the integrated task are related to Rater 8 and Rater14, while the highest and lowest mean scores on the independent task are related to Rater 8 and Rater 4, respectively.

As for the variability, the lowest variability on the integrated task was related to Rater7 (SD = .474) and the lowest variability on the independent task was related to Rater 8 (SD = .408).

Table 2 reports the mean scores assigned by male and female raters, for both independent and integrated tasks.

Table 2. Descriptive Statistics for Male and Female Raters' Mean Scores

	Integrated task		Independent task	
Gender	**Mean**	**Std. Deviation**	**Mean**	**Std. Deviation**
Male	5.809524	1.074759	5.907143	1.111231
Female	6.564762	1.255424	6.995238	1.226474

As Table 2 signifies, the female raters assigned higher scores (mean= 6. 7798) than the male raters (mean=5. 8536).

G-study Variance for Persons, Tasks and Raters

The following table presents the variance components for both male and female raters. There are seven different variance components based on the study design, including: persons (P), raters (R), tasks (T), person-by-rater (PR), person-by-task (PT), rater-by-task (RT), and the triple interaction of persons, tasks, and raters (PTR).

Table 3. Variance Components for Male and Female Raters

Effects	**Male**		**Female**	
	σ^2	%	σ^2	%
P (Persons)	0.2909004	17.36456	0.2445676	6.73396
R (Raters)	0.1252189	7.47463	1.5200356	41.8528
T (Tasks)	0.00	0	0.0330172	0.991
PR (Person × Rater)	0.2604953	15.5496	0.1313930	3.617795
PT(Person × Task)	0.0113369	0.6767	0.1176177	3.2385
RT (Task × Rater)	0.2249316	13.4267	0.3488054	9.604
PRT (Person × Task × Rater)	0.7623700	45.5	1.2364168	34.0436

The G-study results showed that variance component of universe score in female group, e.g., the raters' variance (41.85%) was large relative to the variation due to person (6.73%) and tasks (0.91%). It is possible that each rater differently scored persons in each task. While for male raters, the variance due to essay writers' performance was relatively larger (17.36%) than rater variance (7.47%) and the task variance was 0 suggesting that each essay writer's performance was scored differently by male raters, where each mean is the overall number of

essay writers in the population and two tasks in the process. These findings indicated the generalizability of scoring is substantially influenced by the raters' gender.

The estimated variance of the task mean scores for male raters was 0, and for the female raters was less than 1, which suggest that there was not any difference in difficulty for the tasks according to scores assigned by male raters and this variance was not remarkable in female group.

The second largest variance component is related to the persons (P) for male and PRT for female raters. The third largest variance component is PR for male raters and P for female raters. However, the variance of persons (P) for the male raters is larger than for female raters. This finding indicates that more variability exists among the essay writers with respect to their writing proficiency scores assigned by male raters. The component of person-by-rater (PR) effect is the third largest variance for male raters and the fourth one for the female raters. This relatively large value suggests that the rank ordering of essay writers differs by raters to a considerable degree (Brennan et al., 1995).

Interpretation of the variance component from the interactions is more complex; the largest variance component among the seven values is the triple interaction of persons, raters and tasks (PTR) for male raters (45.5%) while it is the second largest variance for female raters (34.04 %). This finding indicates that the essay writers were rank-ordered across different tasks by rater pairs differently.

Increasing the Number of Male and Female Raters from One to Four

This section centers on D-studies which employ different combinations of male and female raters. Different D-studies including different numbers of male and female raters will be suggested in the following section to identify how increasing the number of male and female raters with two task types (integrated and independent tasks) affects measurement precision. Because, in practice, in testing situation 3 or 4 raters are used, that is why D-studies in the present study are not used more than 4 raters. D-study uses variance components produced in the G-study to design a measurement procedure that is of lower the measurement error (Bachman, 2004). From this analysis, the absolute error variance, the relative error variance, the generalizability coefficient, and the phi coefficient were calculated to provide a clear picture of score reliability across different D-studies.

Absolute Error Variance

G-theory can make a distinction between error variances coming from absolute decisions and relative decisions. Absolute decisions are decisions about the absolute level of performance and

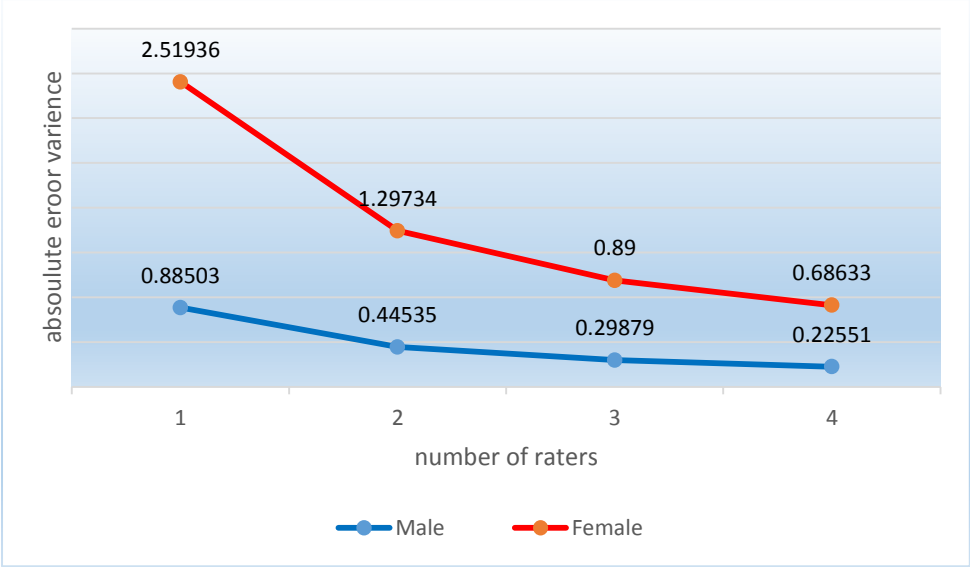
relative decisions concern individual differences (Shavelson & Webb, 1991; Strube, 2002). Absolute error variance is particularly used in criterion-referenced interpretations. The focus in such contexts is on making decisions about whether a student can perform at a prespecified level regarding of how other test takers performed (Brennan, 2001a; Shavelson & Webb, 1991). "It reflects both information about the rank ordering of students and any differences in the average scores" (Shavelson & Webb, 1991, as mentioned in Alkharusi, 2012, p. 192). All sources except for the object of measurement are considered a source of error in case of absolute decisions (Strube, 2002).

Table 4. Absolute Error Variance

Raters	Male	Female	Amount of decrease	
1	0.88503	2.51936	**Male**	**Female**
2	0.44535	1.29734	0.43968	1.22202
3	0.29879	0.89000	0.14656	0.40734
4	0.22551	0.68633	0.07328	0.20367

It is also clear from Table 4 that increasing the number of raters leads to absolute error variance reduction. The most notable decrease happened when the number of raters increased from one rater to two raters. As indicated in the above table, when $n_r = 1$ and $n_t = 2$ the absolute error variance of the male rater was 0.88503 and that of the female raters was 2.51936. However, when the number of raters increased from one to two for both male and female raters, keeping the number of tasks constant ($n_t = 2$), the absolute error estimate was reduced to 0.44535 for male raters and to 1.29734 for female raters. This means a substantial reduction in error coming from just increasing the number of raters from one to two raters.

Figure 1. Absolute Error Variance



Relative Error Variance

The relative error variance is useful when the primary concern of researchers is in making norm-referenced decisions that involve the rank ordering of individuals.

Table 5. Relative Error Variance

Raters	Male	Female	Amount of decrease	
1	0.64735	0.80841	**Male**	**Female**
2	0.32651	0.43361	0.33084	0.3748
3	0.21956	0.30868	0.10695	0.12493
4	0.16609	0.24621	0.05347	0.06247

The findings indicated that the relative error variance for female raters is higher than the relative error variance for male raters. As Table 5 signifies, the relative error variance of having one male rater is 0.64735 and it is 0.80841 for one female rater. Furthermore, the results indicate that the relative error variance of the female raters is consistently greater than that of the male raters across all 4 D-studies.

It should be added that increasing the number of male and female raters from one to two considerably reduced the amount of relative error variance.

Figure 2. Relative error variance

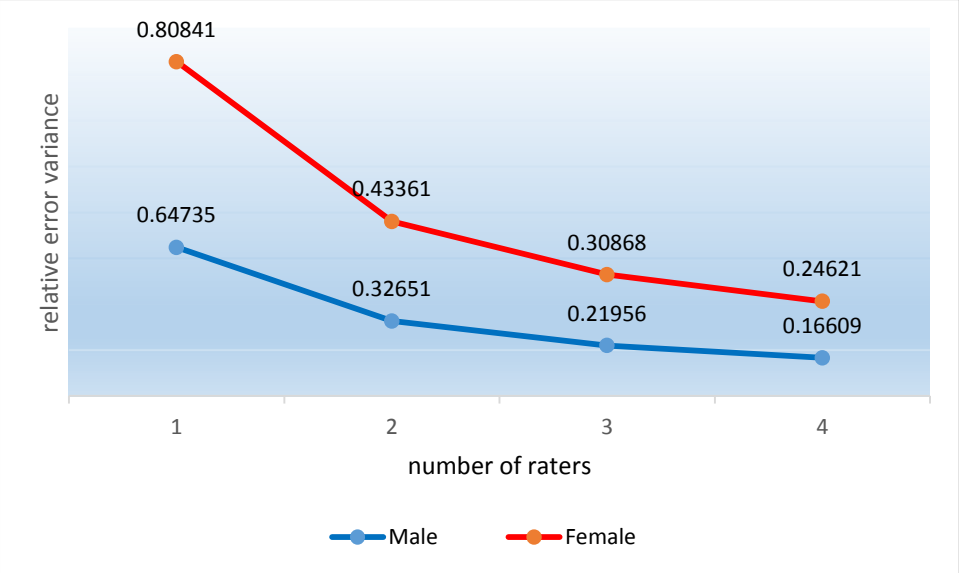


Figure 2 reveals that increasing the number of raters from two to three and from three to four caused the relative error variance to decrease, but this increase was not as considerable as having two raters. The most substantial reduction in error variance happened in case of one male or one female rater and two tasks, 0.33 for male raters and 0.37 for female raters.

Generalizability Coefficient

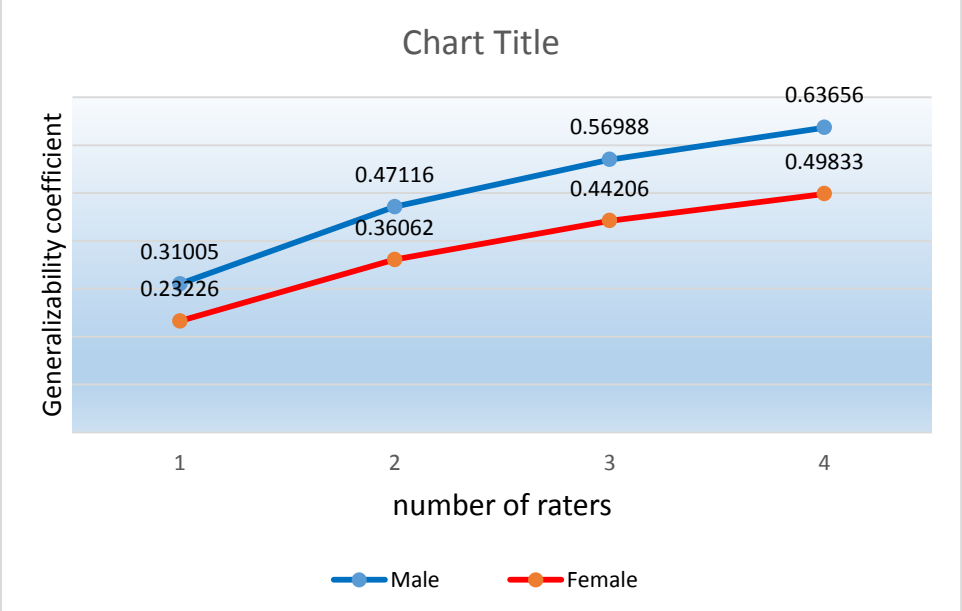
Generalizability coefficient is "the extent to which one can generalize across levels of the facet without misrepresenting the data" (DeVellis, 2003, p. 56). It is the indicator of the dependability of a measurement procedure; this index is equivalent to reliability coefficient in the CTT (Brennan, 2001a). It signifies the ratio of universe score variance to both the universe score variance and the relative error variance. It ranges from 0 to 1; the higher the values the more dependable measurement procedures are. Values approaching 1 indicate that the scores of interest can be differentiated with a high degree of accuracy despite the random fluctuations of the measurement conditions (Allal & Cardinet, 1997; Shavelson & Webb, 1991; Strube, 2002).

Table 6. Generalizability Coefficient

Raters	Male	Female	Amount of increase	
1	0.31005	0.23226	**Male**	**Female**
2	0.47116	0.36062	0.16111	0.12836
3	0.56988	0.44206	0.09872	0.08144
4	0.63656	0.49833	0.6668	0.05627

The results of the present study indicate that generalizability coefficients of the male raters were relatively higher in all the D-studies than the female raters. According to Table 6, the generalizability coefficient for four male raters and two tasks was 0.63656 while it was 0.49833 for four female raters and two tasks. These values decreased to 0.31005 for one male rater and two tasks and to 0.23226 for one female rater and two tasks.

Figure 3. Generalizability Coefficient



As depicted in Figure 3, the most considerable improvement in the generalizability coefficient is observed in increasing the number of raters from one to two holding the number of tasks constant (n_t = 2). The generalizability coefficient of the male raters increased 0.16111 when the number of raters increased from one to two, this increase was 0.12836 for the female raters.

Phi Coefficient

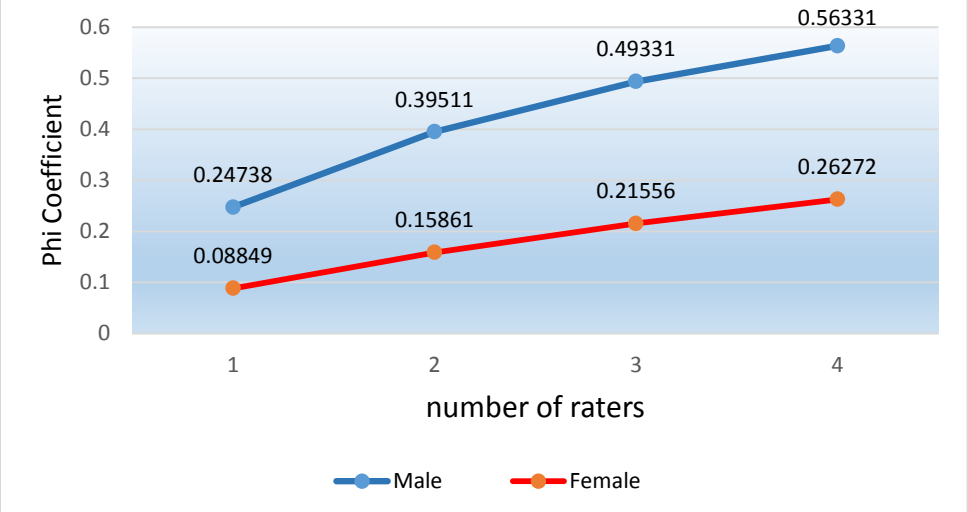
The phi coefficient or the dependability coefficient is "the ratio of universe score variance to itself plus absolute error variance" (Brennan, 2001a, p. 272).

Table 7. Phi Coefficient

Raters	Male	Female	Amount of increase	
1	0.24738	0.08849	**Male**	**Female**
2	0.39511	0.15861	0.14773	0.07012
3	0.49331	0.21556	0.0982	0.05695
4	0.56331	0.26272	0.07	0.04716

Based on Table 7, the male raters have relatively higher phi coefficients in all the D-studies compared to female raters. For instance, the phi coefficient for one male rater and two tasks was 0.24738, while it was 0.08849 for one female rater and two tasks. The phi coefficient was low for both genders, it may be due to low score generalizability of performance assessment such as writing task (Gebril, 2009).

Figure 4. Phi Coefficient



Increasing the number of raters from one to four leads to increasing phi coefficient. As Figure 4 shows, the highest phi coefficient is related to four raters and two tasks in each gender (0.56331 for male raters and 0.26272 for female raters). Besides, the highest increase in the phi coefficient (from 0.14773 for male raters and 0.07012 for female raters) happened in increasing the number of raters from one to two and keeping the number of tasks constant ($n_t = 2$).

Discussion and Conclusion

The purposes of the current study were to assess the reliability of writing assessment when taking into account the facets of tasks, raters, and raters' gender and to find the effect of sequentially increasing the number of male and female raters.

The results revealed that the mean score of the independent task was slightly higher than that of the integrated task and the mean scores of the female raters were higher than those of the male raters. The purpose was not to compare both task types; besides, it cannot be stated that the participants outperformed on independent task. Although some studies have compared the score awarded to the independent and integrated writing tasks, to the best of the researchers' knowledge, no study has provided evidence on the relationship between the assigned scores of these two writing tasks. For example, Lewkowicz (1994) reported no significant difference in the scores of independent and integrated writing tasks. In another study, Gebril (2006) used two different scoring rubrics to compare the performance of EFL students on independent and integrated writing tasks and reported a high correlation between the two sets of scores. However, opposite results have also been reported; for example, performing Pearson coefficient analysis on the scores of independent and integrated writing tasks, Delaney (2008) found that the independent scores were not significantly correlated with those of integrated writing task.

The descriptive statistics indicated that the integrated task had a higher variance compared to the independent task. This difference may be because of the nature of integrated tasks which put more burden on raters' shoulders. In rating integrated task the raters were required to judge about using the source text.

The G-study variance components indicated that, in this study the true score variance (P) is less than the PTR variance which may be due to essay writers' same level of English proficiency and writing ability. The obtained results are in line with Gebril's (2009) study. As pinpointed in methodology section, the essay writers were selected from a homogeneous group; meaning that due to the writing courses they had passed, they were also at the same level of writing ability.

The findings also revealed that the person-by-task (PT) variance component was very low. The relatively low value suggests that either the rank ordering of task difficulty is the same for the various examinees, or that the rank ordering of examinees was the same by task to a substantial degree. Besides, Rater effect (R) was also high in both male and female groups which is accounted for the first highest variability among the female raters and the third highest variability among male raters. This findings lends support to the fact that the rater facet contributes to score variability. This high variance could be because of several factors, such as the number of raters contributing in the present study.

The task (T) effect variability does not account for any of the variance in male group and the lowest variability in female group. This result indicates that both tasks had equal difficulty for essay writers.

As Figure 1 and Figure 2 show, both the relative and absolute error coefficients decreased substantially by increasing the number of male and female raters. Based on the results, when the number of male and female raters increased from one to two, the error decreased to a large extent. This change was by far the most substantial one in both rater's genders. This result lends support to other generalizability studies (Brennan et al., 1995; Gebril, 2009; Lee & Kantor, 2005; Lane & Sabers, 1989; Schoonen, 2005; Swartz et al. 1999). Although increasing the number of raters to more than two brought about in smaller error coefficients, the improvement from one rater to two raters for both male and female raters is the most substantial one.

The univariate analysis, as depicted in Figure 3 and Figure 4, provide evidence that the male raters are much more consistent and much more reliable than the female raters in both the norm-referenced and criterion-referenced score interpretation contexts, meaning that, in both of these contexts, the male raters were much more consistent and reliable than the female raters. Follow up interviews or implementation of think-aloud protocols seem to be required to add additional details about what male and female raters considered during rating writing samples.

In sum, the current study attempted to investigate the score generalizability of independent and integrated writing tasks rated by male and female raters. Results showed that both male and female raters yielded reliable scores. However, caution is warranted when interpreting the writing assignment results. The findings provide support for the fact that under proper conditions and by employing appropriate study designs, very high levels of reliability can be attained in grading writing samples.

Implications of the Study

The present research aimed to scrutinize the effects of raters' gender on the scoring variability and reliability of IELTS different writing tasks. The results can provide a number of important implications.

One of the important implications is that the raters could participate in both independent and integrated writing tasks scoring regardless of their gender since they could rate the task with the same score reliability.

Furthermore, the present research findings specified that the number of raters has an important role in the score generalizability of writing ability. The most significant increase in the reliability of scores resulted from increasing the number of raters from one to two. However, the decision to increase the number of raters is determined by the available resources; the ideal situation is to participate the more experienced raters and holding rater training sessions rather than just increasing the number of raters.

Limitations of the Study

Some limitations of the current study should be considered when interpreting the results. First, the estimates computed for this study may be useful to persons working with Iranian EFL students of similar characteristics. Since the writing tasks were written by L2 essay writers whose first language is Persian, the results may not be generalizable to other language groups. Besides; the raters had also the same first language as the test takers; it is useful to participate native English language raters with L2 writing scoring experience.

An additional limitation of this study is introduced by the small number of conditions sampled within each facet. The sampling error of the estimates of variance components is large when few conditions are used in the estimate (McMaster & Espin, 2007). As the number of degrees of freedom increases, the accuracy of the estimate increases, too. A qualitative measure of writing ability was not included. This study only utilized IELTS holistic rubric for both independent and integrated tasks. Finally, only two writing tasks were used in this study, and increasing tasks number from one to two was not scrutinized; may be increasing the number of tasks besides increasing the number of raters provides a clearer picture of the different variance components.

Acknowledgement

The authors would like to show their gratitude to Dr Robert Brennan, director of the center for advanced studies in measurement and assessment (CASMA) in the college of education of the

University of Iowa, for technical assistance with data analysis and for providing insight and expertise that greatly assisted the research.

References

Alkharusi, H. (2012). Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education*, *2*(1), 184-196.

Allal, L., & Cardinet, J. (1997). Generalizability theory. In J.P. Keeres (Ed.), *Educational research, methodology, and measurement: An international handbook* (2nd, pp. 737-741). Cambridge, United Kindom: Cambridge University.

Anatol, T., & Hariharan, S. (2009). Reliability of the Evaluation of Students' Answers to Essay-type Questions. *West Indian Medical Journal, 58*(1), 13-16.

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay Rating processes and outcomes*, (Unpublished Doctoral Dissertation). Canada: University of Toronto.

Barkaoui, K. (2011). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, *44(1)*, 31-57.

Bell, R. C. (1980). Problems in Improving the Reliability of Essay Marks. *Assessment & Evaluation in Higher Education, 5*(3), 254-263.

Brennan, R. L. (2001a). *Generalizability Theory*. New York: Springler-Verlag New York, Inc.

Brennan, R. L. (2001b). *mGENOVA* (Version 2.1) [Computer software and manual]. Iowa City, IA: University of Iowa.

Brennan, R. L., Goa, X., & Colton, D. (1995). Generalizability analyses of work keys listening and writing tests. *Educational and Psychological Measurement*, *55*(2), 157–176.

Brown, G. T (2010). The validity of examination essays in higher education: issues and responses. *Higher Education Quarterly*, *64*(3), 276-291.

Brown, J. D. (1991). Do English and ESL faculties rate writing samples differently? *TESOL Quarterly*, *25*(4), 587-603.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework (TOEFL Monograph Series, Report No: 22).* Princeton, NJ: Educational Testing Service.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, 7, 140-150.

DeVellis, R. (2003). *Scale development: Theory and applications* (2nd ed.) thousand Oaks, CA:Sage.

Emig, J. (1997). Writing as a mode of learning. In Villanueva, V. (Ed.), *Cross talk in composition theory*. Urbana, IL: National Council of Teachers of English.

Evans, N. W., Hartshorn, K. J., McCollum, R. M., & Wolfersberger, M. (2010). Contextualizing corrective feedback in second language writing pedagogy. *Language Teaching Research*, 14(4), 445–463.

Fulcher, G. (1999). Assessment in English for Academic Purposes: Putting content validity in its place. *Applied Linguistics*, 20 (2), 221-236.

Gebril, A. (2006). *Independent and Integrated academic writing tasks*: A study in generalizability and test method, (Unpublished doctoral dissertation). The University of Iowa, Iowa City.

Gebril, A. (2009). *Score generalizability in writing assessment: The interface between applied linguistics and psychometrics research.* Saarbrücken, Germany: VDM Verlag Dr. Müller.

Ghalib T., K & Al-Hattami, A., A. (2015). Holistic versus analytic evaluation of EFL writing: A Case Study, *English Language Teaching, 8*(7), 225-236.

Hamp-Lyons, L. (1990). Second language writing: assessment issues. In Kroll Barbara, (Ed). *Second Language Writing: Issues and Options*. New York: Macmillan

Han, T. (2013). *The impact of rating methods and rater training on the variability and reliability of EFL students' classroom-based writing assessments in Turkish universities: An Investigation of Problems and Solutions.* (Unpublished Doctoral Dissertation). Turkey: Atatürk University.

Huang, J. (2007). *Examining the fairness of rating ESL students' writing on large scale assessments*, (Unpublished Doctoral Dissertation). Canada: Queen's University.

Huang, J. (2008). How Accurate Are ESL Students' Holistic Writing Scores on Large-Scale Assessments?—A Generalizability Theory Approach. *Assessing Writing*, *13*(3), 201-218. http://dx.doi.org/10.1016/j.asw.2008.10.002

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. http://dx.doi.org/10.5054/tj.2011.269751

Huang, J. (2012). Using Generalizability Theory to Examine the Accuracy and Validity of Large-Scale ESL Writing Assessment. *Assessing Writing, 17,* 123-139. http://dx.doi.org/10.1016/j.asw.2011.12.003

Huang, J., & Foote, C.J. (2010). Grading Between Lines: What Really Impacts Professors' Holistic Evaluation of ESL Graduate Student Writing? *Language Assessment Quarterly, 7*(3), 219-233.

Huang, J., & Han, T. (2013). Holistic or analytic – A dilemma for Professors to score EFL essays? *Leadership and Policy Quarterly, 2*(1), 1-18.

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

Jones, L. (1979). *Notions in English*. Cambridge: Cambridge University Press.

Landy, F.J., Farr, J.L. (1983). T*he measurement of work performance: Methods, theory, and applications.* New York: Academic Press.

Lane, S., & Sabers, D. (1989). Use of generalizability theory for estimating the dependability of a scoring system for sample essays. *Applied Measurement in Education,* 2(3), 195–205.

Lee, Y., & Kantor, R. (2005). *Dependability of new ESL writing test scores: Tasks and alternative rating schemes* (TOEFL Monograph Series no. 1). Princeton, NJ: ETS.

Lewkowicz, J. (1994). *Writing from sources: does source material help or hinder students 'performance?* Paper presented at the Annual International Language in Education Conference, Hong Kong. [ERIC Document Reproduction Service No. ED386050].

Lock, C. (1984). *The influence of the interviewer on student performance in tests of foreign language oral/aural skills.* Unpublished master's theses, University of Reading, Berkshire, England

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276.

McCarter, S. (2002). *Academic writing practice for IELTS*. London: IntelliGene.

McMaster, K., & Espin, C. (2007). Technical features of curriculum-based measurement in writing: A literature review. The Journal of Special Education, 41(2), 68-84.

McNamara, T. F. (1996). *Measuring second language performance*. New York: Longman.

Mousavi, S.A. (2007). *Computer Package for the Assessment of Oral Proficiency of Adult ESL learners: Implications for Score Comparability*. (Unpublished PhD Thesis) Griffith University.

National Commission on Writing for America's Families, Schools, and Colleges. (2003). *The Neglected "R": The need for a writing revolution*. Retrieved from http://www.host-collegeboard.com/advocacy/writing/

Nunan, D. (1996). Towards autonomous learning: some theoretical, empirical and practical issues. In R. Pemberton, S.L. Edward, W.W.F. Or, and H.D. Pierson (Eds.), *Taking Control: Autonomy in Language Learning*. Hong Kong: Hong Kong University Press. 13-26.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *Journal of System, 28*, 373-386.

Porter, D. (1991a). Affective factors in language testing. In C. J. Alderson & B. North (Eds.), *Language testing in the 1990s* (pp. 32-40). London: Modern English Publications.

Porter, D. (1991b). Affective factors in the assessment of oral interaction: Gender and status. In S. Arniva (Ed.), *Current development in language testing* (pp. 92-102). Singapore: SEAMEO Regional Language Center

Porter, D., & Shen. H. (1991). Sex, status, and style in the interview. *Dolphin, 21*, 117-128.

Sakyi, A. A. (2000). Validation of holistic scoring 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.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing, 25*(4), 465-493.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, 22(1), 1–30.

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

Shavelson, R. J., & Webb, N. M. (2005). Generalizability theory. In J. L. Green, G. Camilli, & P. B. Elmore (Eds.), *Complementary methods for research in education* (3rd ed.). Washington, DC: AERA.

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, 18(3), 303-325.

Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of rater background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 28–33.

Strube, M. J. (2002). Reliability and generalizability theory. In L.G. grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (23-66). Washington, DC: American Psychological Association.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., DeKruif, R. E. L., Reed, M. et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59*(3), 492–506.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197–223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15*(2), 263–287.

Weigle, S. C. (1999). Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches. *Assessing Writing*, *6*, 145-178.

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

Appendix: Writing tasks

Integrated writing task

You should spend about 20 minutes on this task.

The table below shows the percentage of the rooms occupied in six hotels during May to September between 1985 and 2000. The table also indicates the star rating of each hotel.

Write a report for a university lecturer describing the information shown below.

You should write at least 150 words.

Hotel	Star	1985	1990	1995	2000
Hotel Concorde	*****	90	90	30	66
Hamilton's	*****	100	100	95	70
The Tower	****	57	85	55	85
Hotel Olivia	***	90	85	89	95
Hampton's	***	100	100	90	100
The Continental	***	79	83	70	80

Independent writing task

Present a written argument to an educated reader with no specialist knowledge of the following topic:

Nowadays in countries like Russia some people try to find their matches for marriage through the internet. While some of these relationships have been reported to have happy endings, traditional marriages are more dependent and stable.

Which opinion do you agree with?

You should write at least 250 words.

Use your own idea, knowledge and experience and support your arguments with examples and relevant evidence.