Revue SCIENCES HumaineS n°45, Juin2016, Vol B, pp95-106

What Do Consistency Estimates Tell Us about Reliability in Holistic Scoring?

Abstract:
Essay writing assessment is a largely used test in many examination types. Preferred fortheir "validity and authenticity" (Hamp-Lyons, 2003:163), direct writing tests are prevailing in entrance, placement examinations, as well as in continuous assessment. However, when
compared to indirect writing testing, their level of reliability is questioned and their scoring procedures incriminated. True scoring does not exist; errors stem from various sources: theraters, their training, the task (Huot, 1990); rendering essay marking doubtful, and raters'scoring inconsistent. This study reports on a large scale, high-stake writing proficiency test taken by 441 students. The essays were holistically scored on a 7-point scale by 16 raters. ThePearson correlation coefficient was used for assessing the degree of consistency between
raters. The coefficientwas calculated for each pair of judges in the 25 groups of students.
Results show positive correlation, but consistency in relationship has revealed some degree of variability between the paired samples. The range of correlations fell between .16 and .91. with the majority between .50 and .74. These findings raise issues about the factors that threaten consistency of scoring in writing tests.

Dr.SLOUGUI Doudja
Ecole Normale Supérieure de Constantine

ملخص:
تعددت طرق وأساليب الإمتحانات والتقويمات بتعدد وتنوع المواد والميولات الأكاديمية وأحيانا الشخصية للأستاذ الممتحن أو المقوم، وعلى رأس لائحة هذه الأساليب يأتي وبشكل ملفت الإمتحان المقالي بشكليه المفتوح والمغلق. الملاحظ أن الإمتحان المقالي المغلق أكثر شيوعا من نظيره المفتوح لنجاعته وسهولة التأكد من صحة وتطابق محتواه مع الدروس المقدمة، على عكس الإمتحان المفتوح الذي يمنح للطالب الممتحن حيزا أكثر فسحة (للتحليلوالإستدلال مثلا) ما قد يشكل تحديا للأستاذ الممتحن الذي قد يقوم المقالة تقويما ذاتيا غير موضوعي لابتعاد محتواها عن مكونات الدروس ما قد يجعل عملية التقويم برمتها غير مضبوطة ولا موثوقة، كثيرا ما تختلف من أستاذ مصحح لأخر. وللتدقيق في هذا الإعتقادإرتأينا دراسة عينة تتكون من 441 طالب مقسمة إلى 25 فوج تم إخضاعهم لامتحان في مهارات الكتابة المقالية، وتم تكليف 16 أستاذ بعملية التقويم. وبتطبيق معامل "بيرسن" (Pearson) ظهرت جملة من الفوارق في نتائج عمليات التقويم ما إستوجبعلينا النظر في تداعيات هذا النوع من التباين في عملية تقويم الإمتحان المقالي.

Introduction :

Essay writing assessment, commonly known as 'direct' test of writing (Weigle, 2002: 58) is a prevailing tool in many examination types. Authentic and easy to administer, it is widely used in various situations, ranging from coursework assessment to large-scale examinations. Academics argue that these performance-based tests are more valid for assessing language proficiency

(Bachman and Palmer, 1996; Greenberg, 1992; McNamara, 1996; Weigle 2002). They reflect students' understanding of the task, their organizational ability, their thinking skills, and their mastery of the language. Put in other words, they are a good indicator of a student's "communicative language ability" (Douglas, 2000:10). Klapper (2006:264) summarizes their operational value and notes that they "are effective vehicles for assessing higher-level language skills, or cognitive academic language proficiency". Additionally, he points out their power of discriminating between students' abilities and argues, " they allow students space for distinctive treatment of a complex topic and give them scope to show what they can do linguistically, i.e. to employ varied structures and a wide range of vocabulary and idiom" (ibid).

However, one of the problems with direct writing tests is the subjective evaluation and inconsistent scoring process (Huot, 1990; Lumley, 2002). Unlike Multiple Choice and computer-based writing tests, which are machine-scored, direct writing tests bear complex scoring procedure. Their evaluation involves human decisions and inferences. In other words, a rater's decision on examinees' ability is subject to factors that inevitably influence scores (Excks, 2012). "These human beings", according to Hamp-Lyons (2003: 165) "are likely to vary from day to day, from subject to subject, and to have preferences for certain kinds of ideas or structures, or dislike for some choices of words or arguments". Lumley (2002) goes on to argue that even the way raters use rating scales can be quite inconsistent and irrational. As a result, these test scores carry directly influence on an individual's life (Bachman and Palmer, 1996). A pass or a fail judgment very often determines a student's fate. While an accurate judgment would "ensure fairness to students; an inaccurate one would yield spurious results" (Huot, 1996: 556). It is likely to debar some individuals from opportunities and elect others with less appropriate profile. Because of this discriminative power, and because "errors in these decisions are difficult to correct and decisions not easily reversed" (Weigle 2002:41), it is strongly supported that writing tests, to be valid, they must first and foremost be reliable. Huot (op.cit) sums up the idea by stating that: "without a sufficient level of agreement between raters a writing assessment procedure cannot be valid".This article, therefore, reports on a study about the reliability of a holistic essay scoring in a proficiency language test for admission to a post-graduate program. It focuses on the inter-rater reliability as well as on some of the factors that are likely to affect the raters' scoring behavior. The paper will try to answer the following questions:1/ To what extent are the markers consistent and accurate in attributing their scores? and 2/ If any discrepancy in awarding scores, where does the inconsistency come from?

Holistic scoring in essay writing

Scoring essays could be achieved in different ways. However, in large-scale assessments, holistic scoring is the most widely used (Brown, 2009;Weigle,

2002). Unlike analytical scoring, the procedure consists of assessing student's writing as a whole piece of work to which the marker awards a single mark. The score is usually based on a rating scale or rubric that describes the scoring criteria for different levels of competence. Despite its advantages (fast, economical and easy to use), holistic scoring presents some disadvantages. The rater's mark may be a highly subjective one (Wang 2009:41); the scoring rubric broad categories do not differentiate between students sufficiently (Klapper 2006:267). The scores are difficult to interpret (Weigle 2002:114). In other words, the scores are not always reflective of the criteria upon which a piece of work is judged. To be trustworthy, test scores should measure a student's real writing ability independently from any source of influence.

Reliability issues in writing tests

Reliability, in its crudest meaning, refers to the "consistency of scoring across readers, what later became known as interrater reliability" (Huot et al 2007:3). This occurs when markers, for some reason, do not agree on the same scores and award different scores to the same paper. Bachman and Palmer (1996:19) define it as "consistency of measurement". They point out that "a reliable test score will be consistent across different characteristics of the testing situation". Seeking to render test scores consistent and accurate, researchers awarded a growing attention to the reliability issues in writing tests. On the one hand, they developed *Pre- test* procedures, which are considered as essential prerequisites for objective and reliable scoring. These consist of "designing and pre-testing prompts, selecting and training raters, double marking of essays, ensuring the independence of scores, and using a scoring rubric that outline the criteria against which the students' writing are to be judged" (Weigle 2002:59). On the other hand, there are *post-test* ways to estimate accuracy and consistency of a test's score. These include statistical methods and measurement approaches, which estimate the reliability of essay scores and identify the sources that affect it (Brown, 2009).

Estimating reliability of essay scores: Two important approaches of assessing reliability of scores are prevailing in writing assessment: intra and inter reliability. Intra-rater reliability refers to "the tendency of a rater to give the same score to the same script on different occasions" whereas "inter-rater reliability refers to the tendency of different raters to give the same scores to the same scripts". (Weigle 2002: 135).

Measures to ensure reliability: A variety of statistical methods has been developed to estimate the accuracy of essay scores and determine the degree of reliability between raters on scoring essays: consensus estimates; consistency estimates and measurement estimates (Stemler 2004).Whereas consensus estimates measure the proportion of essays that get the same scores; consistency estimates indicate whether the raters are consistent in giving

scores. Measurement estimates indicate the degree to which scores can be attributed to common scoring rather than to error (Brown 2009:44).

Factors affecting reliability: Various studies show that reliability in essay scoring is difficult to achieve (Greenberg, 1992, Brown, 2009, Fei et al 2011, Weigle 2002). Previous research has identified various sources from which error stems. These include the student, the test and the rater, noting that all these factors interact with each other and with the context in which testing takes place.

The purpose of this study is then to estimate the accuracy and consistency of a test scoring procedure. It seeks to measure the extent to which the markers have been consistent in awarding grades. The achieved results will help us throw light on the factors that are likely to have affected the reliability of assessment.

The context of research

This study reports on a large scale, high-stakes writing proficiency test taken by 441 advanced level students. Streamed in 25 groups, the students composed a timed essay, responding to a prompt on an unknown topic. The purpose of the writing task was to measure their ability to use English effectively. The essays were holistically scored on a seven-point scale by 16 raters. At least, two raters scored each essay and judgments were expressed as grades. In discrepancy cases, a third rater judgment was called for. Raters, with varied qualifications, teaching experience and background, were selected among EFL higher education teachers (table 1). Most of them have little or no experience in large-scale marking operations, and none of them received training in holistic scoring.

	Teaching experience in EFL	Teaching writing experience	Experience in large scale examinations
0	-	4	11
≤ 5	4	7	3
Between 5-10	4	3	2
Between 10-15	2	0	-
Between 15-20	4	0	-
≥ 20	2	2	-
N markers	16	16	16

Table1: Demographic profile of the makers

Methodology

The Pearson Product-moment correlation Coefficient computed software [http://www.socscistatistics.com/tests/pearson/Default2.aspx] was used to investigate the consistency of ratings between raters. The correlation coefficient, symbolized by (r), was calculated for each pair of judges in 25

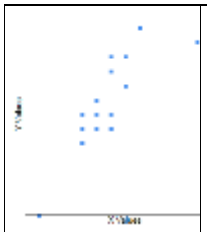
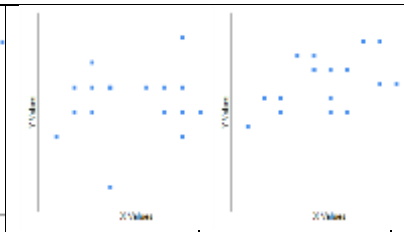
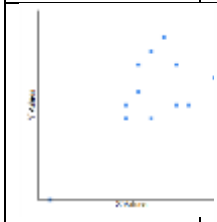
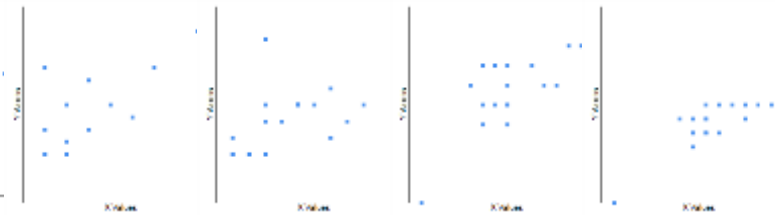
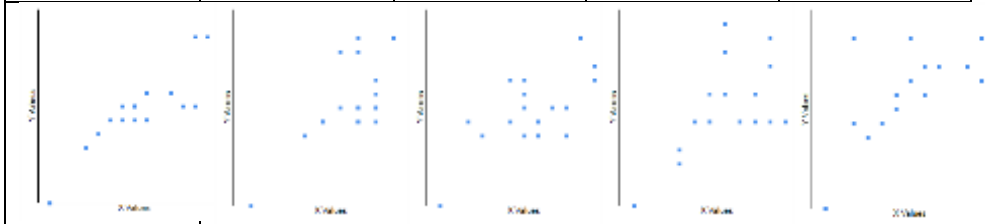
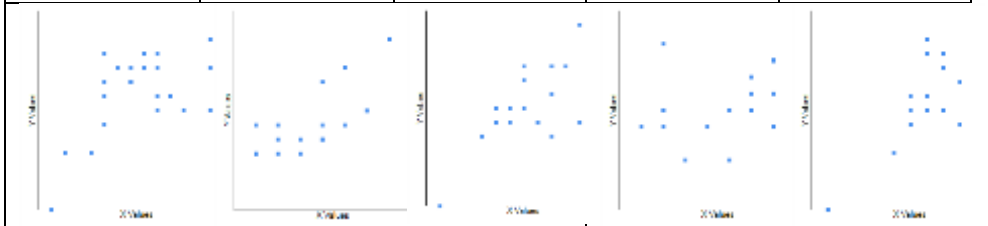
groups of students to measure the degree of relationship between the two scores attributed by the first and second rater. When *r* is close to zero, this indicates little or no correlation between the scores. However, when *r* is close to 1, this indicates a strong relationship between the set of scores.

The reason for choosing a consistency rather than a consensus estimate (percentage agreement or Cohen's Kappa coefficient) is that the raters did not receive any training in using the rating scale. Therefore, that raters come to exact agreement was beyond our expectations. However, the consistency approach is less stringent. Stemler (2004: 6) argues, "it is not really necessary for two judges to share a common meaning of the rating scale, so long as each judge is consistent in classifying the phenomenon according to his or her own definition of the scale".

Results

Each scatterplot below illustrates the relationship that exists between the sets of scores in each group. The variables plotted in the graph and labelled (x) and (y) correspond to the scores attributed by the first and second raters.

When the correlation coefficient is strong, all the data fall closer to a straight line; when the coefficient is moderate, the value decreases and the data points deviate from the straight line. However, if there is no linear relationship between the variables, the data points on the graph will be randomly scattered and approximate a circle.

				
[1] r= +0.7157	[2] r= +0.1607	[3] r= +0.4769	[4] r= +0.8619	[5] r= +0.7827
				
[6] r= +0.6318	[7] r= +0.6931	[8] r= +0.5624	[9] r= +0.7919	[10] r= +0.8316
				
[11] r= +0.9121	[12] r= +0.7291	[13] r= +0.7191	[14] r= +0.6376	[15] r= +0.7409
				
[16] r= +0.5076	[17] r= +0.7723	[18] r= +0.758.	[19] r= +0.4802	[20] r= +0.7209
				
[21] r=	[22] r=	[23] r=	[24] r=	[25] r=

+0.8593	**+0.6808.**	**+0.8629.**	**+0.7229**	**+0.4085**

Fig.1: Summary of the correlation values of (*r*) in the different sets of data

The correlation coefficient *(r)* in the 25 groups of students shows a positive direction (+). In other words, there is a tendency for high scores of one rater (x) to go with high scores of the other rater (y) and the low scores of one tend to pair with the low scores of the other. If, for example, an essay scores high with rater (x), there is greater chance that rater (y) will do the same.

However, the magnitude of scores shows some degree of variability.The range of correlations falls between .16 and .91. with the majority between .50 and .74. Values beyond ≥ .90 are deemed strong, and values equal or below ≤ .35 are considered weak (See table 2).

Markers	N	Value of r	Relationship
Pair 1 M1 and M2	20	+0.7157	Moderate positive correlation
Pair 2 M1 and M2	20	+0.1607	Weak positive correlation
Pair 3 M1 and M2	20	+0.4769	Weak positive correlation
Pair 4 M1 and M2	21	+0.8619	Strong positive correlation
Pair 5 M1 and M2	18	+0.7827	Strong positive correlation
Pair 6 M1 and M2	15	+0.6318	Moderate positive correlation
Pair 7 M1 and M2	19	+0.6931	Moderate positive correlation
Pair 8 M1 and M2	21	+0.5624	Moderate positive correlation
Pair 9 M1 and M2	20	+0.7919	strong positive correlation
Pair 10 M1 and M2	20	+0.8316	strong positive correlation
Pair 11 M1 and M2	20	+0.9121	Very strong positive correlation
Pair 12 M1 and M2	20	+0.7291	Moderate positive correlation
Pair 13 M1 and M2	21	+0.7191	Moderate positive correlation
Pair 14 M1 and M2	20	+0.6376	Moderate positive correlation
Pair 15 M1 and M2	20	+0.7409	moderate positive correlation
Pair 16 M1 and M2	24	+0.5076	moderate positive correlation
Pair 17 M1 and M2	20	+0.7723	Strong positive correlation
Pair 18 M1 and M2	20	+0.758.	Strong positive correlation
Pair 19 M1 and M2	20	+0.4802	weak positive correlation
Pair 20 M1 and M2	20	+0.7209	moderate positive correlation
Pair 21 M1 and M2	13	+0.8593	Strong positive correlation

Pair 22 M1 and M2	15	+0.6808	Moderate positive correlation
Pair 23 M1 and M2	14	+0.8629	Strong positive correlation
Pair 24 M1 and M2	13	+0.7229	Moderate positive correlation
Pair 25 M1 and M2	13	+0.4085	Weak positive correlation

Table 2. Paired Samples Correlations

Discussion

In this paper, we have tried to investigate whether the markers are consistent in scoring essays in a large-scale writing examination. Consistency correlations fell in the range between .16 and .91, meaning that the scoring of paired raters, though positive, exhibits variability, ranging from weak to high correlations.

A high correlation coefficient, as explained earlier, implies that raters attributed high and low scores following a coherent pattern. They, however, did not necessarily award the same scores to the same essays. Thirty-six (36 %) of the paired samples exhibited strong positive correlation. Forty-eight (48%) showed a moderate a correlation and sixteen (16%) displayed a weak correlation.

As an illustration, if we compare the scores attributed by the markers in the paired sample 11 (table 3), which obtained the most significant correlation coefficient [.91]; we will notice that despite their high correlation coefficient, the two markers have not come to a consensus (total agreement). Their mean scores are different. Whereas the mean score of (M1) is 7.4, the mean score of (M2) is 6.7. Yet, one can predict how they apply the rating scale. The two markers are constant in their scoring. Marker (M1) tends to be more lenient than marker (M2).

Group 11	Minimum score	Maximum score	Range	Mean	STD Deviation
M1	3	14	9	7.4	288.8
M2	4	13	9	6.7	202.2

Table 3: paired sample 11 statistical data

Conversely, if we compare the scores attributed by raters in the paired sample 2 (table 4), which obtained a negligible correlation coefficient [.16], we will notice that no relationship exists between the two sets of scores. Markers have attributed scores inconsistently. In other words, what is scored high by one rater is scored low by the other, as these sets of scores, from out data illustrate: { 12,7,12,14,11,13,7,12….} compare with { 5,5,3,5, 4,4,4,4,…}.Unlike, the previous example, the mean values here show a significant discrepancy. Rater's (M2) scoring markedly deviates from the norm.

Group 2	Minimum score	Maximum score	Range	Mean	STD Deviation
M1	5	14	9	9.2	139.2
M2	1	7	6	4.45	28.95

Table 4: paired sample 2 statistical data

Such findings raise questions as to what makes the scores modestly converging or utterly discrepant. Previous research has identified the rater's scoring behavior as the dark side of reliability. This statement of fact is best described in Diederich's words (in Hamp-Lyons2003: 183) who claims, "The score an essay received could depend more on whom the rater was than on any qualities inherent to the text itself".

Differences in raters' scoring have widely been researched and the various studies (e.g. Bachman and Palmer, 1996; Lumley, 2002; McNamara, 1996; Weigle 2002) indicate that the source of variance may arise from a) the criteria, which guide the raters' judgment and b) their individual attributes. Since there is no absolute view on what makes a rater prefer some piece of writing rather than another, raters tend to value certain features; while others may downgrade them. Some give importance to content, but others praise form. Unless qualitativelyresearched, the criteria upon which the markers, in this study, have based their scoring remain difficult to determine. Nonetheless, it is believed that the use of a rating scale might have helped in standardizing the criteria.

The markers' experience and background teaching, however, are assumed to impact scores significantly in this study. Previous studies (Brown: 2009) reported that raters who lack expertise in academic content are less reliable because different disciplines apply different criteria for assessment. In this context, four markers (25%) have never experienced teaching writing, let alone experience in holistic scoring and large-scale examination. Influenced by the criteria developed in their own specific teaching areas, these markers might have imported other standards than those valued by the writing community. Lack of experience in the field or shortish knowledge in composition teaching, is considered as one factor that is likely to have affected objective scoring.

Additionally, comparing senior and junior grading, researchers reported that teaching experience influences scoring as much as content knowledge. Huot (in Weigle 2002:70) argues that although both expert and novice raters were primarily concerned with content, expert raters have developed more coherent strategies. Along these lines, Weigle (1994) supports that inexperienced markers, before training, tended to be both more severe and less consistent in their ratings than the experienced ones. Barkaoui (2010)

counterclaims and argues that experienced raters tend to assign lower scores and to give more importance to linguistic accuracy than the novices do. He, nonetheless, explained that these latter are more concerned with argumentation and their scores exhibited more variability. In general, studies on raters' attributes suggest that a mix of both senior / junior raters, without training, is a problematic issue. If we turn to the present study, we find that seven out of sixteen graders (44%) are novice to the field. Their professional experience is less than five years. Three (19%) have some practice that goes beyond five years and only two (12%) can be said to have a considerable familiarity and knowledge in the field. This association of both experienced/inexperienced is certainly a very fruitful experience in terms of enculturation and preparation of novice writing readers, but its impact on consistency and variability in scoring is, without doubt, significant.

More importantly, background training is reported to be the most influential factor affecting scoring.Shohamy et al (1992) maintain that training is more significant than any other attribute in terms of rater reliability. Weigle (1998) argues that although a rater's training does not eliminate individual trends, it improves reliability largely. Markers' Training is, regrettably, the crucial element that was missing in this reported on writing test. None of the markers has ever been trained in using the rating-scale and scoring large-scale writing examinations. Though very important, there had been no preparation of teachers for such a task.

Conclusion

Essay scoring is one of the most daunting tasks in foreign language teaching. The task is even harder when dealing with large-scale and high-stakes examinations. In order to ensure reliability and accuracy in awarding scores, various precautions that foster consistency between markers need to be taken seriously. First, scoring in large-scale tests should not be assigned to markers who are not acquainted with assessing writing. Second, markers, whose scoring is deemed inconsistent, should be showed how to score correctly. Finally, the most urgent decision is certainly to select and train markers to interpret a scoring rubric and to assign grades objectively.

References

1. Bachman, L.F. and Palmer A.S. (1996) Language Testing in Practice : Designing and Developing Useful Language Tests. Oxford: Oxford University Press
2. Barkaoui K. (2010) Explaining ESL Essay Holistic Scores: A Multilevel Modeling Approach *Language Testing* 27(3)

3. Brown, T.L. Gavin (2009) The Reliability of Essay Scores: The Necessity of Rubrics and Moderation. In Tertiary Assessment and Higher Education Student Outcomes: Policy, Practice and ResearchEds:Luanna H. Meyer; Susan Davidson; Malcolm Rees ; Richard B. Fletcher and Patricia M. Johnston . Wellington, New Zealand : AkoAotearoa
4. Douglas, D. (2000) Assessing Language for Specific Purposes, Cambridge
   Cambridge University Press
5. Excks,T (2012) Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Performance Assessment Quarterly*, 9: 270–292.
6. Fei Wong FookFei, MohdSallehhudinAbd Aziz and ThangSiew Ming (2011).The Practice of ESL Writing Instructors in Assessing Writing Performance.*Procedia Social and Behavioral Sciences* 18 (2011) 1–5
7. Greenberg, L. Karen (1992) validity and reliability: Issues in the direct assessment of writing. *Writing Program Administration* Vol.16 Nos 1-2,
8. Hamp-Lyons,L (2003), Writing teachers as assessors of writing in Exploring The Dynamics of Second Language Writing Barbara Kroll (ed). Cambridge University Press, 162-190
9. Huot, B. (1990). The Literature of Direct Writing Assessment: Major Concerns and Prevailing Trends. *Review of Educational Research*, 60(2), 237–263. Retrieved from http://www.jstor.org/stable/1170611
10. Huot, B. (1996).Toward a New Theory of Writing Assessment *CollegeComposition and Communication*, Vol. 47, No. 4 (Dec., 1996), pp. 549-566 Published by: National Council of Teachers of English Stable URL: http://www.jstor.org/stable/358601Accessed: 30/09/2010 16:23
11. Huot, B and Peggy O'Neill (2007). Assessing Writing: The Introduction A Critical Sourcebook. Bedford/St. Martin's http://casymposium.blogspot.com/2007/10/assessing-writing-introduction.html
12. Klapper (2006) Understanding and Developing good practice : Language Teaching in Higher Education . London: CILT
13. Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *LanguageTesting,* 19*,* 246-276.

14. McNamara. T. F.(1996). Measuring second language performance London; New York: Longman
15. Shohamy, E. Gordon, C. and Kraemer, R. (1992) the effect of raters'background and training on the reliability of direct writing tests. Modern Language Journal 76(4), 513-521
16. Stemler, Steven E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation,* 9(4). http://PAREonline.net/getvn.asp?v=9&n=4 .
17. Wang, P (2009) The Inter-Rater Reliability in Scoring Composition. *English Language Teaching*, 2 (3)
18. Weigle, C, Sarah S.C. (1994). Effects of training on raters of ESL compositions. *LanguageTesting, 11*, 197-223.
19. Weigle, C, Sarah (1998). Using FACETS to model rater training effects
    at: *Language Testing*, 15/2/263 http://ltj.sagepub.com/content/15/2/263
20. Weigle, C, Sarah (2002). Assessing Writing. Cambridge University Press