Journal of Translation and Languages مجلة الترجمة واللغات
ISSN : 1112-3974

Scoring academic writing from subjectivity to objectivity: A scale for evaluating students' written product

Abu Shawish Jaber
University of Al-Quds - Palestine
jabushawish@qou.edu

Abstract: *Due to the drawbacks of adopting any of the common scoring scales independently and in order to make the scoring process less subjective and more objective, the researcher thought of developing a rubric for evaluating students' written product. Some of the well-known scoring scales are time consuming, others mainly depend on the scorer's impression which is not always accurate. The researcher here made a hybrid of both the holistic and analytic scoring scales and developed it to suit academic writing. The rubric was refereed by a group of specialists who assured its validity. Content validity and internal consistency were also calculated. To test the suitability of this scoring rubric for the purpose it was designed for, a stratified random sample of 30 essays was selected from a corpus of N=120 essays written by Palestinian tertiary level students majoring in English in the academic year 2009/ 2010. Pearson correlation and Alpha scale were used in order to assure the reliability of the scale. The Scoring scale follows a taxonomy in which errors experienced in the subjects' writings are divided into three major categories: errors related to conventions, content development and style. Each category includes a number of subcategories; for instance, conventions include writing mechanics and grammar and vocabulary common errors; each of which is scored out of 25. In addition, content development includes cohesion and coherence, which are scored out of 20 marks and 15 marks respectively and finally style and cogency are marked out of 15 marks only.*
Keywords: *Academic writing, scoring scale, written product.*

ملخص كان الدافع وراء تطوير هذا المقياس لتقويم النتاج الكتابي لدى الطلبة هو أن تصبح عملية التصحيح موضوعية أكثر نظرا للسلبيات الكثيرة لإتباع أي من مقاييس التصحيح المعروفة. فبعض هذه المقاييس يبدد الوقت، والبعض الأخر يعتمد في الأساس على انطباع المصحح الذي لا يكون دوما دقيقا. قام الباحث في تصميمه لهذا المقياس بالمزج بين المقياس التحليلي والمقياس الكلي وطور فيهما لما يتناسب مع الإنشاء الأكاديمي. تم عرض هذا المقياس على مجموعة من المختصين الذين أكدوا صلاحيته، وقد تم التأكد من صدق المحتوى والاتساق الداخلي لمكونات المقياس. ولقياس ثبات المقياس الجديد اختيرت عينة عشوائية طبقية مكونة من 30 مقالة أكاديمية كتبها طلاب المستوى الثالث الجامعي المتخصصين في اللغة الانجليزية في العام الدراسي 2010/2009. وقد استخدم أيضا مقياس "الفا" ومعامل ارتباط "بيرسون" للتأكد من ثبات هذا المقياس. يتبع هذا المقياس طريقة مختلفة حيث أن أخطاء الطلبة في الإنشاء يتم تقسيمها إلى 3 أنواع رئيسة: منها ما يتعلق بأسس الكتابة وتلك المتعلقة بتطوير المحتوى وما تتعلق بالأسلوبيات، وتم تقسيم كل منها إلى عدة أفرع. بالنسبة لأسس الكتابة تم تقسيمها إلى الهجاء من ناحية والقواعد والمفردات من ناحية أخرى ويتم تصحيح كل منها من 25 علامة، بالإضافة إلى انه تم تقسيم

تطوير المحتوى إلى الترابط القواعدي وترابط الأفكار وتم تخصيص 20 علامة للأولى و15 علامة للثانية، كما يتم تصحيح الأسلوبيات من 15 علامة.

الكلمات المفتاحية: الإنشاء الأكاديمي، مقاييس التصحيح، النتاج الكتابي.

1. Introduction

Assessment or evaluation? These two terms are most often used interchangeably. However, evaluation sometimes refers to assigning a score to a direct writing product based on predefined criteria. It is distinguished from assessment in that the scoring for the latter focuses more on feedback and alternative evaluative techniques in the process of learning (Bacha, 2001:371). In this study the term evaluation will refer to assigning a score to a direct writing product.

Then, evaluating the writing of students of English as a foreign language poses a number of problems amongst which is objectivity of evaluation. Given, scoring writing mainly depends on the rater's impression and hence it is subjective. Nevertheless, coloring the scoring process with some kind of objectivity seems not totally impossible. "The marking of writing tests will always be at least somewhat subjective, but the use of descriptors for each level of the marking scheme can at least help make the marking consistent" (Kitao and Kitao 1996:7). Therefore, in order to make the scoring process of any written work, more valid and reliable scoring scales or rubrics should be used. The demand for such valid and reliable methods of assessing second and foreign language writing has grown in significance in the preceding few years.

In adopting scoring instruments with clearly identifiable criteria for evaluating English as a Foreign Language EFL academic writing, the paramount guiding principle is obviously the purpose of this study. Evaluating academic writing in EFL/ESL programs has been mainly for diagnostic, developmental or promotional purposes (Weir, 1993, 1990). Thus, in order for these programs to obtain valid results upon which to base decisions, the choice of evaluation instrument to be adopted becomes significant. Although Gamaroff (2000) states that language testing is not in an ‘‘abyss of ignorance’’ (Alderson, 1983 cited in Gamaroff, 2000), the choice of the ‘right’ essay writing evaluation criteria in many EFL/ESL programs remains problematic as often those chosen are inappropriate for the purpose. This is most crucial when decisions concerning student promotion at the end of the semester to the next English course have to be made mainly based on essay writing scores. It is then important that teachers are aware of the potential of the evaluation criteria being adopted.

2. Rationale and Purpose

The impressionistic holistic scoring method is the paramount one adopted in scoring writing in the Palestinian Academic institutions as it is the case in most of Arab countries. In evaluating essay writing either analytically or holistically, teachers have had to address a number of concerns that affect the assigning of a final

score to a writing product. Some of these concerns have included the need to attain valid and reliable scores, set relevant tasks, give sufficient writing time, set clear essay prompts, and choose appropriate rhetorical modes (Braddock et al.,1963; Cooper and Odell, 1977).

Due to the drawbacks of adopting any of the common scoring scales independently and in order to make the scoring process less subjective and more objective, the researcher thought of developing a rubric which is proper for evaluating his students' written product. Some of the well-known scoring scales are time consuming, others mainly depend on the scorer's impression which is not always accurate. The researcher here made a hybrid of both the holistic and analytic scoring scales and developed it to suit academic writing. The rubric was refereed by a group of specialists who suggested some modifications until it occurred in its present form.

3. Problem Statement

Various analytical and/or holistic scoring schemes have been suggested for assessing the different types of academic writing i.e. argumentative, expository, narrative writing. However, few have been empirically validated and fewer proved reliability and practicality to assign a score for the different parts of the written work. To the researcher's best knowledge none of the scales adopted measures the writer's competency which is sometimes necessary, particularly in competitions.

Scoring processes and rubrics are generally concerned with assigning the total mark for a piece of writing; they rarely give the student feedback on the strengths and weaknesses his work has. Raters do not tend to examine each part of writing in isolation; most often, they deal with a piece of writing as a whole. Accordingly, the present study attempted to address the following research question:

To what extent is the proposed scoring scale reliable and valid to evaluate students' writings?

4. Literature Review

Assessing students' writing is one of the most difficult and time-consuming activities for assessing their competency or achievement. The need for creating scoring rubrics, designing guidelines increases with time. Multiple raters need to be trained, and then the students' writings need to be scored, typically by multiple raters. With different people evaluating different essays, interrater reliability becomes an additional concern in the writing assessment process. Even with training, differences in the background, training, and experience of the raters can lead to noticeable but important differences in grading.

In this respect, Abu Shawish (2009: 93) states " … almost the scores recorded by rater 2 were similar to those granted by rater 1. However, there were discrepancies in some papers where the final scores recorded by the two raters were very far from one another. Consequently, a third rater was consulted to solve the discrepancy in

the scores recorded for those few papers. The third rater reconsidered and revised the scoring of these few papers and found that they were improperly scored due to the raters' fatigue. A sort of compromise was agreed on for those papers. That is the three raters sat and negotiated the matter together. The final scores for these few papers were modified and finally recorded."

Besides, the scoring method adopted affects the reliability and nature of the work to be scored. In addition to the choice of the scoring rubric which suits the topic, time allotted to writing and applying the scoring rubric by raters are of importance to the evaluation of writing. East (2009) deals with a method which has a central place in many testing contexts all over the world, i.e. the timed writing test. East revealed that reliability of this test method - timed writing test- is heavily influenced by the scoring procedures, including the rating scale to be used and the success with which raters can apply the scale. According to East, reliability is crucial because important decisions and inferences about test takers are often made on the basis of test scores and determining the reliability of the scoring procedure frequently involves examining the consistency with which raters assign scores.

Furthermore, the scoring process should enjoy appropriateness and accuracy in order to result in reliable outcomes. Brown (2004) believes that accuracy in the scoring of writing is critical if standardized tasks are to be used in a national assessment scheme. He identifies three approaches to establishing accuracy i.e., consensus, consistency, and measurement.

5. Related Studies

Beyreli and Ari (2009) conducted a study whose purpose was to determine whether there was concordance among raters in the assessment of the writing performance using analytic rubric. In addition, it examined the different factors that may affect the assessment process. The analytic rubric used in the study consists of three sections and ten properties: External structure (format, spelling and punctuation), language and expression (vocabulary, sentences, paragraphs, and expression), organization (title, introduction, story, and conclusion). The basis of Beyreli and Ari's (2009) study is composed of narrative texts written by 200 students studying at the sixth and seventh grades of schools located on the Anatolian side of Istanbul. Texts were assessed in accordance with the analytic rubric by six raters. It was determined that the concordance among raters was sufficient according to the results of the assessment.

From another angle, Rezaei and Lovoron (2010) carried out an experimental project investigating the reliability and validity of rubrics in assessment of students' written responses to a social science "writing prompt". The participants were asked to grade one of the two samples of writing assuming it was written by a graduate student. Both samples were prepared by the writers. The first sample was well written in terms of sentence structure, spelling, grammar, and punctuation; however, the writer did not fully answer the question. The second sample fully answered each part

of the question, but included multiple errors in structure, spelling, grammar and punctuation. In the first experiment, the first sample was assessed by participants once without a rubric and once with a rubric. In the second experiment, the second sample was assessed by participants once without a rubric and once with a rubric. The results showed that raters were significantly influenced by mechanical characteristics of students' writing rather than the content even when they used a rubric. The study results also indicated that using rubrics may not improve the reliability or validity of assessment if raters are not well trained on how to design and employ them effectively.

In addition, Attali and Powers (2009) created a developmental writing scale on the basis of automatically computed indicators of writing fluency, word choice, and conventions of standard written English for timed essay-writing performance. In a large-scale data collection effort that involved a national sample of more than 12,000 students from 4th, 6th, 8th, 10th, and 12 th grade, students wrote (in 30-min sessions) up to four essays in two modes of writing on topics selected from a pool of 20 topics. Scale scores were created by combining essay indicators in a standard way to compute essay scores that shared the same scoring standards across essay prompts and student grade levels. A number of studies were conducted to examine the validity of scale scores. Cross classified random effects modeling of scores confirmed that the particular prompts on which essays are written have little effect on scores. The reliability of scores was found to be higher compared to previous reliability estimates of human essay scores. A human scoring experiment confirmed that the developmental sensitivity of scale scores and human scores was similar. A longitudinal study confirmed the expected gains in scores over one-year period.

6. Methodology

To test the suitability of this scoring rubric for the purpose it was designed for, a stratified random sample of 30 essays was selected from a corpus of N=120 essays written by Palestinian tertiary level students majoring in English in the academic year 2007/ 2008. The essays have been done by the students for the purpose of assessing the level of achievement of those students after completing two writing courses in three of the Palestinian national universities in Gaza Strip.

6.1. The Scoring Scale

This scoring scale is mainly based on the types and frequencies of errors Palestinian university students experience in their writings. The major goal of developing such a scale is to make the scoring process more objective and less subjective. Furthermore, this scoring scale, unlike the common and well known methods of scoring writing, assesses the overall students' written product and competence in writing out of 100. Here each paragraph should be scored independently out of 100 marks according to the categories included in the writing

scale and the final mark of the whole essay can be calculated depending on the number of paragraphs consisting the essay.

The Scoring scale follows a taxonomy in which errors experienced in the subjects' writings are divided into three major categories: conventions, content development and style. Each category includes a number of subcategories; for instance, conventions include writing mechanics and grammar and vocabulary common errors; each of which is scored out of 25. In addition, content development includes cohesion and coherence, which are scored out of 20 marks and 15 marks respectively and finally style and cogency are marked out of 15 marks only (See Writing Scoring Scale).

The rationale for assigning the scores as shown above for each category is that the researcher followed two common types of scoring in this scale, namely, analytic and holistic scoring methods. Analytic scoring based on objective judgment covers 70 % of the scoring scale's total mark, and the rest is scored through holistic scoring which is based on subjective judgment. The researcher did his best to make the latter less subjective through dividing its subcategories into smaller elements.

6.2. Writing Scoring Scale

CONVENTIONS		**CONTENT DEVELOPMENT**		**STYLE**	**Total Mark**
Mechanics	**Grammar & Vocab.**	**Cohesion**	**Coherence**	**Style & Cogency**	
- faulty capitalization - punctuation errors - indentation errors - comma splice - hyphenation - Wrong spelling - should be one word	- S-V agreement - modifier-head noun agreement - dangling modifier - fragment - fused sentence - shift in person - verb tense - error in verb form	- Error in pronoun reference - unnecessary repetition - the use of conjunction - substitution - ellipsis - recurrence	- faulty logic - inadequate development - transition weak/ missing - paragraph not unified - faulty parallelism - irrelevance	- diction weak/ poor - word choice - informal - (non)standard - passive weakening phrasing - meaning vague/ unclear - paragraph	

- should be two words - spacing error	- wrong part of speech - missing word - structure incomplete / unacceptable - misuse of adjectives & adverbs - errors in case forms - faulty abbreviation - faulty subordination - articles incorrect/ missing - misuse of prepositions	- synonyms - collocation	- paragraphing - reasoning - completeness	development - topic sentence - bad translation - makes no sense, confusing, illegible - variety in sentence structure, length, type, inversion - conclusion - contractions	
/25 marks	**/25 marks**	**/20 marks**	**/15 mark**	**/15 marks**	**/100**

This scale has been developed by the researcher.

7. Scoring the test

The test papers designed for the present study were marked out of 100. The scoring scheme was not an easy task because of the variability of responses since the

test is an open essay type. Raters specialized in teaching writing courses from the three aforementioned national universities helped in the scoring and rescoring process in order to explore how competent the subjects were in handling written discourse in terms of tense use, sentence and paragraph structure, coherence and cohesion, mechanics of writing and developing the main idea of the paragraph using the proposed scoring scale in order to test its suitability for the purpose it was designed to achieve.

For the process of rating the subjects' written product in order to be less subjective, the scoring rubric developed by the researcher and refereed by a group of specialists was adopted. It is important to mention in this respect that the subjects were inquired to write an essay of no lesser than 200 words on the topic. Accordingly, for making each two errors in any of the categories underlying writing mechanics, grammar or cohesion, the subject loses one mark. However, if the same error occurred more than once in the same paper, then it would be counted as one error only and the student loses half a mark for it. Coherence and style and cogency which have been scored holistically where the marking rates the overall proficiency level and which depends on the general impression of the rater, the following mechanism was adopted. Coherence was assigned 15 marks, with the least score of 4 and the highest of 12 depending on the rater's impression towards the writer's adoption of the elements composing coherence. With regard to style and cogency which were assigned 15 % of the total mark, the rater would grant a minimal score of four and a maximal score of 12 depending on the writer's inclusion of the elements of style.

The work of the subjects was first marked by the researcher himself, then checked and rechecked by another rater to make sure that almost every item of the scoring rubric was considered. The final mark was then recorded on each paper. For the purpose of granting the scoring process more reliability, a copy of the subjects' work was given to another rater to score them in the light of the scoring rubric adopted by the researcher, rater 1. The second rater analyzed the students' written product thoroughly and finally recorded the final score on each paper. It is safe to say that almost the scores recorded by rater 2 were similar to those granted by rater 1. However, there were discrepancies in some papers where the final scores recorded by the two raters were very far from one another (See Appendix 1). Consequently, a third rater was consulted to solve the discrepancy in the scores recorded for those few papers. The third rater reconsidered and revised the scoring of these few papers and found that they were improperly scored due to the first or the second raters' fatigue. A sort of compromise was agreed on for those papers. That is the three raters sat and negotiated the matter together. The final scores for these few papers were modified and finally recorded.

8. Validity of the scoring scale

The scoring scale, after having been formulated in its final form, has been presented to a panel of specialists to judge its suitability for the purpose of the present study. They recommended it and appreciated the way it was designed. Some comments and modifications have been proposed and hence taken in the researcher's consideration. Finally, the common elements of writing were included in the scoring scale; however, less frequent ones were disregarded and hence discarded.

Another type of validity, namely content validity was also referred to in order to test the consistency of the scoring scale. Two types of consistency were used i.e. the internal consistency and the structure consistency. To test the internal consistency of the scoring process according to the scoring scale, the researcher distributed the a test-One-question essay test- to a sample of thirty male and female students. Pearson correlation criterion was used to check the internal consistency of scores assigned to each of the test items. Students' scores in the test items were correlated with each other. Table (1) below shows the internal consistency of the scoring scale through the use of Pearson correlation.

Table (1): The internal consistency of the scoring scale

Criteria	Coefficient Correlation
Mechanics 25	0.871
Grammar & vocabulary 25	0.862
Cohesion 20	0.700
Coherence 15	0.622
Style & cogency 15	0.748

R table value at (df=28) and sig. level (0.05) = 0.361
R table value at (df=28) and sig. level (0.01) = 0.463

Table (1) above shows that there is a statistically significant correlation at the levels (0.01) and (0.05) between all scoring scale items' i.e. mechanics, grammar and vocabulary, coherence, cohesion, style and cogency scores and the total mark of the test, which assures that the latter is internally consistent, R at the level (0.01) = 0.463 and at the level (0.05) = 0.361. Table (2) below elucidates the structural consistency of the scoring scale through the use of Pearson correlation. The correlation between the mark of each criterion and that of other criteria in the scoring scale was calculated.

Table (2): The structural consistency of the test

Criteria	Mechanics 25	Grammar 25	Cohesion 20	Coherence 15	Style 10
Mechanics 25	1.000				
Grammar & vocab 25	0.527	1.000			
Cohesion 20	0.375	0.504	1.000		
Coherence 15	0.386	0.416	0.845	1.000	
Style & cogency 15	0.369	0.497	0.676	0.689	1.000

r table value at (df=28) and sig. level (0.05) = 0.361

r table value at (df=28) and sig. level (0.01) = 0.463

Like table (1), table (2) above reveals that there is a statistically significant correlation at the levels (0.05) and 0.01) between all the scoring scale items. This result asserts that it is structurally consistent.

9. Reliability

Reliability is referred to as the stability, accuracy or consistency of the instrument used to achieve the purpose of the study. In other words, once an individual achieved the same or nearly the same scores in the same test when applied more than once, then, it is reliable. (Abu Allam 1998: 418-419).

Reliability coefficient is correlation coefficient either between the subjects' scores in the test in different times, between the different scores assessments in different times or between the test scores of the same group of subjects assessed by a group of specialist raters. However, the assessment of the scoring scale reliability indicates a sort of consistency of scores assigned. Reliability could be assessed in different ways: inter-rater reliability, intra-rater reliability, split-half reliability …etc.

For the purpose of the present study, the first type of reliability, seems to be the most suitable one since more than one rater is involved in assessing the written discourse of the subjects. Alpha scale was also used in order to assure the reliability of the test. The raters, as mentioned above, cannot be expected to give identical scores, so their collective decision would be more reliable than the one taken by a single rater. Therefore, inter-rater reliability has been adopted in the scoring of the test. Table (3) represents the reliability of the scoring scale. It shows the statistical methods namely, Person correlation and T-test Paired sample used to achieve test reliability.

Table 3. Pearson correlation & T. test to explore the differences between the scores assigned by the two raters'

		Mean	Std. Deviation	Pearson Coefficient	Sig. level for correlation		Sig. level for T
R1	40	57.033	12.925	.947	Sig. at 0.01	.792	Not sig.
R2	40	57.633	11.941				

T table value at df (29) and sig. level (0.05) = 2.045
T table value at df (29) and sig. level (0.01) = 2.756

Table (3) above shows that correlation between the marks assigned by the two raters was statistically significant at the level (0.01) whereas T value was statistically insignificant, which indicates that there is no statistically significant difference between the marks assigned by the two raters. The results obtained here are a foolproof that scoring process was reliable. In addition, the researcher used Alpha coefficient scale for the same purpose. He found that Alpha coefficient = (0.649), which is also another evidence for the reliability of the scoring scale.

10. **Conclusion**

The aim of the present study was to find out what general lessons can be drawn for the evaluation of writing. Although holistic scoring may blend together many of the traits assessed separately in analytic scoring, making it relatively easier and reliable, it is not as informative (Bacha 2001: 371–383). For the learning situation as analytic scoring although the study was done on a limited sample, the results indicate that more attention should be given to the language and vocabulary aspects of students' writing and a combination of holistic and analytic evaluation is needed to better evaluate students' writing proficiency at the end of a course of study. In the final analysis, relevant evaluation criteria go hand in hand with the purpose upon which the criteria, benchmark essays and training sessions are based (Pierce, 1991; Elbow, 1999). Most of all perhaps, these initial results confirm the complexity involved in choosing rating scales and delineating criteria for valid and reliable essay evaluation on which to base promotion decisions.

In a nutshell, the statistical results obtained in this study strongly confirm the research question. Thus, the present scoring scale can be reliable to score FL learners writings.

References

[1] Attali, Y. And Powers, D. (2009): Validity of Scores for a Developmental Writing Scale Based on Automated Scoring. Educational & Psychological Measurement. Vol. 69, No. 6, pp 978-993.

[2] Bacha, N. (2001): Writing evaluation: what can analytic versus holistic essay scoring tell us? System, Vol. 29, pp 371–383.

[3] Beyreli, L. and Ari, G. (2009): The Use of Analytic Rubric in the Assessment of Writing Performance -Inter-Rater Concordance Study. Educational Sciences: Theory & Practice, Vol. 9, No. 1, pp 105-125.

[4] Braddock, R. et al., (1963): Research in Written Composition, Illinois, National Council of Teachers of English.

[5] Brown, G. (2004): Accuracy in the Scoring of Writing: Studies of Reliability and Validity Using a New Zealand Writing Assessment System. Assessing Writing. Vol. 9, No. 2, pp105-121.

[6] Clauser, B., Harik, P and Clyman, S. (2000): The Generalizability of Scores for a Performance Assessment Scored with a Computer-Automated Scoring System. Journal of Educational Measurement; Vol. 37, No. 3, pp245-262.

[7] Community College Week (2000): A Program that Scores Essays Electronically. Fall Supplement, Vol. 13, No. 2.

[8] Coniam, D. (2009): Experimenting with A Computer Essay-scoring Program Based on ESL Student Writing Scripts. ReCALL; May2009, Vol. 21, No. 2, pp259-279.

[9] Cooper, C.R., Odell, L., (1977): Evaluating Writing: Describing, Measuring, Judging. National Council of Teachers of English, Urbana, Illinois.

[10] East, M. (2009): Evaluating the Reliability of a detailed Analytic Scoring Rubric for Foreign Language Writing. Assessing Writing, Vol. 14, No. 2, pp88-115.

[11] Elbow, P., (1999): Ranking, Evaluating, and Liking: Sorting out Three Forms of Judgments. In: Straub, R. (Ed.), A Sourcebook for Responding to Student Writing. Hampton Press, Inc, New Jersey, pp 175–196.

[12] Gamaroff, R., (2000): Rater Reliability in Language Assessment: The Bug of all Bears. System 28, No. 1, pp 31–53.

[13] Hamp-Lyons, L., (1990): Second language writing: assessment issues. In: Kroll, B. (Ed.), Second Language Writing: Research Insights for the Classroom. Cambridge University Press, Cambridge, pp. 69–87.

[14] Hamp-Lyons, L. (Ed.), (1991): Assessing Second Language Writing in Academic Contexts. Ablex, Norwood, NJ.

[15] Hamp-Lyons, L., (1995): Rating Nonnative Writing: The Trouble with Holistic Scoring. TESOL Quarterly, Vol. 29, No. 4, pp759–762.

[16] Hayward, M., (1990): Evaluations of Essay Prompts by Nonnative Speakers of English. TESOL Quarterly, Vol. 24, No. 4, pp753–758.

[17] James, C. L. (2006): Scoring Academic Writing. Assessing Writing, Vol. 11, No. 3, pp167-178.

[18] Kitao, S. and Kitao, K. (1996): Testing Writing. Available on ERIC.

[19] Pierce, B.N., (1991): TOEFL Test of Written English (TWE) Scoring Guide. TESOL Quarterly, Vol. 25, No. 1, pp 159–163.

[20] Rezaei, A. and Lovoron, M. (2010): Reliability and Validity of Rubrics for Assessment Through Writing. Assessing Writing, Vol. 15, No. 1, pp18-39.

[21] Weir, C. J. (1990): Communicative Language Testing. New York, Prentice-Hall.
[22] ---------- (1993): Understanding and Developing Language Tests. New York: Prentice-Hall.
[23] Yong-Won, L., Gentile, C. and Kantor, R. (2010): Toward Automated Multi-Trait Scoring of Essays: Investigating Links among Holistic, Analytic, and Text Feature Scores. Applied Linguistics, Vol. 31, No. 3, pp391-417.