JOURNAL OF ENGLISH STUDIES
Islamic Azad University, Science & Research Branch, 1(4), 59-68**,** Fall 2011

The Impact of Rating Methods and Task Types on EFL Learners' Writing Scores

Mahnaz Saeidi*
Islamic Azad University, Tabriz Branch, Tabriz, Iran

Shokouh Rashvand Semiyari
Islamic Azad University, Tabriz Branch, Tabriz, Iran

Received: June 20, 2012 **Accepted: February 21, 2013**

ABSTRACT: The difficulty of assessing the writing skill is well known. Different testing facets seem to affect the result of assessing the writing skill. In addition to the writer's ability, the topic of the writing task, and methods of rating may contribute to the writer's score. In this study, 50 EFL learners wrote four different types of writing tasks (convincing, describing, instructing, and explaining), and these tasks were scored by three raters using two scoring methods: holistic and analytic. The Repeated-Measures ANOVA and variance components analyses of these ratings illustrated that the rating methods and their interaction significantly affected EFL learners' writing scores.

Keywords: G-theory, holistic and analytic rating method, writing score, task type.

The assessment of writing proficiency (as well as speaking) forms the most common type of performance-based testing. In the real world, the writer has to generate the relevant information, formulate the message linguistically, and take into account the characteristics of the readers. In fact, in testing situations, test developers should present test-takers with authentic tasks to enable them to interpret and create authentic language. However, test taker's writing ability cannot be easily assessed; the assessment of writing proficiency has always been problematic. As Cooper (1984) proposes, raters often do not agree with themselves and score the same text differently. At the same time, we realize that the scores derived from test performance need to be both reliable and valid. That is, the inferences about language ability that we make from the scores need to be valid (Bachman, Lynch, & Mason, 1995). Ackerman and Smith (1988) assert that more restrictive writing tasks are less valid as they activate fewer writing skills than authentic writing tasks.

As Schoonen (2005) asserts, different sources of variance contribute to the variance of the scores in writing assessment. Possible sources of variance according to Schoonen (2005) are the topic the student writes about (e.g., prescribed or self-chosen), the discourse mode, the type of the text or genre that is required (e.g., description, exposition, narrative or argumentation), the time limits imposed, the writing mode (e.g., paper-and-pencil or text processor), the conditions of testing, rater inconsistency, scoring procedure (e.g., holistic or analytic), and traits to be scored (content or form).

Identifying as many potential sources contributing to variation in scores as possible is the main goal of G-theory. This theory, which was developed by Cronbach, Gleser, Nanda, and Rajaratnam (1972), as an extension of the classical testing theory, describes ways to estimate the size of different sources of (error) variance in multifaceted measurements. As Suen (1990) explains, "G-theory

* Corresponding author`s email address: m_saeidi@iaut.ac.ir

provides testers with a conceptual framework to assess multiple sources of variation, or measurement error, within the context of a given testing situation" (pp.41-42). For instance, behavioral measurement that often focuses on individuals' performance across a set of test items or tasks is scored by two or more judges. Thus, conceptually, individuals, items/tasks, and judges can contribute to variation among the scores (Shavelson & Webb, 1991).

There are a number of studies that employ G-theory. Bachman, Davidson, and Milanovic (1996) investigated rater agreement by means of generalizability analysis. The results indicated that the overall level of rater agreement was very high, and raters were more consistent in rating method than ability. McNamara and Lumley (1997) studied the effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. The results of the analysis revealed the effects of interlocutor variability and audiotape quality on ratings. In fact, the candidate's score was clearly the outcome of an interaction of variables, only one of which was the candidate's ability.

In this study, our focus is on the assessment of writing, but similar points can be made for the assessment of speaking (Lee, Kantor, & Mallaun, 2002). Lumley and O' Sullivan (2005) studied the effect of test-taker gender, audience and topic on task performance in tape-mediated speaking assessment. The results showed small effects for some, but not all of the hypothesized interactions. There were also some differences in performance by male and female candidates on the relevant tasks. Lee (2006) applied G-theory procedures to examine the impact of the number of tasks and raters per speech sample and of subsection lengths on the dependability of speaking scores. The finding in the univariate analyses showed that it would be more efficient to increase the number of tasks rather than the number of ratings per speech sample to maximize the score dependability. The multivariate G-theory analyses also revealed that (1) the universe (or true) scores among the task-type subsections were highly correlated and that (2) slightly larger gains in composite score reliability would result from increasing the number of listening–speaking tasks for the fixed section lengths. Schoonen (2005) investigated generalizability with an application of structural equation modeling with respect to writing scores. This study found that the specific assignment introduces a fair amount of variance in scores (Lee, Kantor, & Mollaun, 2002). Taking into account that most studies in this realm have focused on the application of G-theory for the purpose of exploring score reliability and because of the lack of sufficient studies in EFL contexts, the researchers aimed at investigating the effect of rating method and task type on Iranian EFL learners' writing scores based on the generalizability of these scores within the context of G-theory. In other words, the researchers intended to study the effect of raters and writing tasks on the generalizability of writing scores and they tried to determine how these effects may depend on the way rating takes place (i.e., holistically or analytically). To this end, the following research questions were raised:

1. Do rating methods have any effects on EFL learners' writing scores?
2. Do task types have any effects on EFL learners' writing scores?
3. Is there any interaction between rating methods and task types and their effect on the writing scores?

Method

Participants

This study included 50 participants. They were English language learners who were studying English as their field of study; that is, they were studying English Language Translation at Islamic Azad University, Shahre-Qods Branch. These students were both male and female. Their age range was between 19 and 22. At the time of this study, these students were taking an advanced writing course.

Instrument

This study used four writing tasks which are listed in Table 1. In these tasks, the students had to describe how to proceed from a certain starting point to a certain goal. The topics (as indicated in Table 1) were chosen due to the fact that similar topics had been already covered in the students' textbook. They were; therefore, familiar with these types of topics.

Table 1. *The Four Writing Tasks: Task Name, Communicative Act and a Description of the Task*

Task name	Communicative act	Description
A	To convince	Write an article for your favorite newspaper to convince readers to adopt your opinion about the elimination of exams from the educational system
B	To describe	Describe to one of your close friends what happened during the birthday party she had missed due to a severe headache.
C	To instruct	Write the instructions for a T.V. cooking program on how to bake chocolate cake.
D	To explain	Write a letter to a teacher in which you explain why you're always late for the class.

Procedure

The participants were supposed to write four essays on communicative functions as Table 1 indicates. Forty five minutes were devoted to each essay. All texts were scored twice, once holistically, based upon general impression on language use, using the revised scale for the British Council's ELTS test (Hughes, 1989) and once analytically, focusing on content, organization, vocabulary, language use, and mechanics (Meisuo, 2000). There were three raters. The first rater was one of the researchers of the present study. She was teaching advanced writing course. The two other raters were selected from among the teachers who were teaching the same course at the same semester. All raters; therefore, had both educational and professional backgrounds in applied linguistics.

Design

In a descriptive-analytic study two variables contributed to this study: task types (to convince, to describe, to instruct, to explain) assigned to students to write about and two rating methods (holistic vs. analytic) based on how their essays were corrected.

Results

The researchers aimed at investigating whether method of rating, type of task and their interaction had any significant effect on the performance of the participants on the writing test? To this end, two steps of data analyses, Repeated-Measures ANOVA and variance components, were conducted:

Repeated-Measures ANOVA

As Table 2 indicates, the F-observed value for the effect of the method of rating, according to Repeated-Measures ANOVA, is 45.38. This amount of value, 4.03, is higher than the critical value of F at 1 and 49 degrees of freedom. It should be noted that the SPSS produced four F-values, the first of which, Pillai's Trace, is the most reliable one even under violation of assumptions and/or unbalanced groups.

Table 2. *Repeated-Measures ANOVA: Method of Rating and Type of Task and their Interaction*

Effect		Value	F	Hypothesis df	Error df	Sig.	Partial Eta Squared
METHOD	Pillai's Trace	.481	45.382	1	49	.000	.481
	Wilks' Lambda	.519	45.382	1	49	.000	.481
	Hotelling's Trace	.926	45.382	1	49	.000	.481
	Roy's Largest Root	.926	45.382	1	49	.000	.481
TASK	Pillai's Trace	.503	15.868	3	47	.000	.503
	Wilks' Lambda	.497	15.868	3	47	.000	.503
	Hotelling's Trace	1.013	15.868	3	47	.000	.503
	Roy's Largest Root	1.013	15.868	3	47	.000	.503
METHOD * TASK	Pillai's Trace	.289	6.369	3	47	.001	.289
	Wilks' Lambda	.711	6.369	3	47	.001	.289
	Hotelling's Trace	.407	6.369	3	47	.001	.289
	Roy's Largest Root	.407	6.369	3	47	.001	.289

Two other statistic support the significant F-value; the probability of .001 which is lower than the significance of .05 and the effect size (partial eta squared) of .48, which is higher than .14. Based on the criteria developed by Cohen (as cited in Cohen & Brooke, 2004) an effect size of .14 or higher is considered strong. The statistically significant F-value indicates that there are significant differences among the mean scores of the two methods of rating. It is worth mentioning that intra/inter-rater reliability has also been calculated as below (Tables 3 to 6 below).

The intra-rater reliability index for rater one is .75 (P = .000 < .05). This result suggests that there is a significant agreement between the ratings of the first rater by using two methods of rating (i.e., analytical and holistic).

Table 3. *Intra-Rater Reliability: Rater 1*

		95% Confidence Interval		F Test with True Value 0			
	IntraclassCorrelation[a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.280[b]	.181	.408	4.107	49	343	.000
Average Measures	.757[c]	.639	.847	4.107	49	343	.000

The intra-rater reliability index for rater two is .66 (P = .000 < .05). This result suggests that there is a significant agreement between the ratings of the second rater by using two methods of rating (i.e., analytic and holistic).

Table 4. *Intra-Rater Reliability: Rater 2*

		95% Confidence Interval		F Test with True Value 0			
	IntraclassCorrelation[a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.196[b]	.110	.315	2.949	49	343	.000
Average Measures	.661[c]	.498	.786	2.949	49	343	.000

The intra-rater reliability index for rater three is .50 (P = .000 < .05). This result suggests that there is a significant agreement between the ratings of the third rater by using two methods of rating (i.e., analytic and holistic).

Table 5. *Intra-Rater Reliability: Rater 3*

		95% Confidence Interval		F Test with True Value 0			
	IntraclassCorrelation[a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.111[b]	.042	.214	1.999	49	343	.000
Average Measures	.500[c]	.259	.685	1.999	49	343	.000

The inter-rater reliability index for the three raters is .88 (P = .000 < .05). These results suggest that there is a significant agreement between the ratings of the three raters.

Table 6. *Inter-Rater Reliability: Three Raters*

		95% Confidence Interval		F Test with True Value 0			
	IntraclassCorrelation[a]	Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.237[b]	.168	.338	8.471	49	1127	.000
Average Measures	.882[c]	.829	.924	8.471	49	1127	.000

Meanwhile, as displayed in Table 7, the students' mean scores under a holistic method of rating, 16.25, is higher than the analytic mean score of 15.29. Thus, it can be concluded that the method of rating (holistic vs. analytic) has the significant impact on the mean scores of the students on the writing scores.

Table 7. *Overall Mean Scores of the Participants on Analytic and Holistic Measures of Writing*

			95% Confidence Interval	
METHOD	Mean	Std. Error	Lower Bound	Upper Bound
ANALYTIC	15.290	.156	14.977	15.603
HOLISTIC	16.253	.122	16.009	16.498

The F-observed value for the effect of the type of task is 15.86 (Table 2). This amount of F-value, 2.80, is higher than the critical value of F at 3 and 47 degrees of freedom. Two other statistical procedures support the significant F-value, the probability of .000 which is lower than the significance of .05 and the effect size (partial eta squared) of .50 which is higher than .14. As mentioned before, based on the criteria developed by Cohen (as cited in Cohen & Brooke, 2004) an effect size of .14 or higher is considered strong. The statistically significant F-value indicates that there are significant differences among the mean scores of the types of tasks.

As displayed in Table 8, the students' mean scores on the second task, to describe, is the highest followed by to convince, to instruct and to explain. It can be concluded that the second null-hypothesis as type of task does not have any significant effect on the mean scores of the students on the writing test can be rejected.

Table 8. *Descriptive Statistics for Types of Tasks*

			95% Confidence Interval	
TASK	Mean	Std. Error	Lower Bound	Upper Bound
TO CONVINCE	15.720	.226	15.265	16.175
TO DESCRIBE	16.427	.182	16.061	16.793
TO INSTRUCT	15.617	.131	15.354	15.880
TO EXPLAIN	15.323	.110	15.103	15.544

Although the F-value of 15.86 indicates significant difference between the mean scores of the four tasks, the post-hoc Scheffe's test should be run to locate the exact place of difference between the means.

The following points can be observed in table 9.

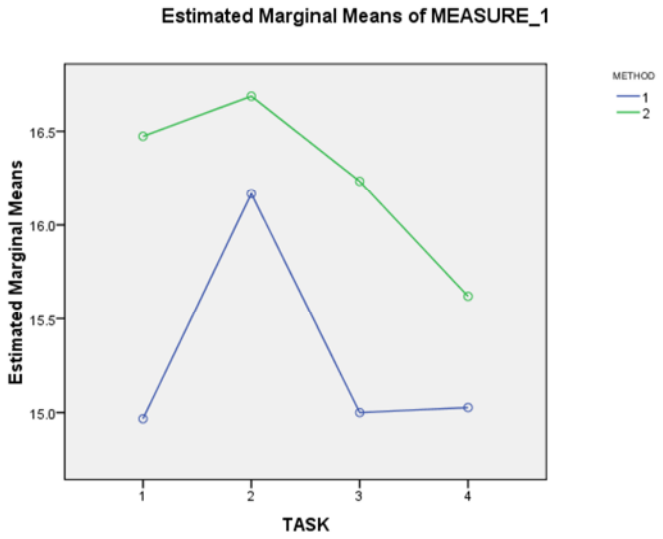
1. There was a significant difference between the students' mean score on describing task and convincing task. The students performed better on the describing task with a mean score of 16.42.
2. There was a significant difference between the students' mean score on describing task and instructing task. The students performed better on the describing task with a mean score of 16.42.
3. There was a significant difference between the students' mean score on describing task and explaining task. The students performed better on the describing task with a mean score of 16.42.

Table 9. *Post-Hoc Scheffe's Tests: Types of Tasks*

					95% Confidence Interval for Difference[a]	
(I) TASK	(J) TASK	Mean Difference (I-J)	Std. Error	Sig.[a]	Lower Bound	Upper Bound
To convince	To describe	-.707*	.159	.000	-1.143	-.270
	To instruct	.103	.234	1.000	-.539	.746
	To explain	.397	.231	.557	-.240	1.033
To describe	To instruct	.810*	.168	.000	.347	1.273
	To explain	1.103*	.195	.000	.567	1.639
To instruct	To explian	.293	.155	.385	-.132	.719

*. The mean difference is significant at the .05 level.

The F-observed value for interaction between methods of rating and the type of task is 6.39 (Table 2). This amount of F-value, 2.8, is higher than the critical value of F at 3 and 47 degrees of freedom. Thus, there is an interaction between rating methods and task types on the writing score. As displayed in Figure 1, the analytic method shows a rapid rise and fall of means, while the holistic method shows a steady and decreasing trend.



Method 1: Analytic; Method 2: Holistic
Task 1: To convince; Task 2: To describe; Task 3: To instruct; Task 4: To explain

Figure 1. *Interaction between method of rating and type of task*

Variance components

G-study; Persons by Raters nested within Tasks, or (p × r: t)

In order to examine the persons, raters, and tasks variance components, G-study with (p × r: t) design was conducted. In such cases, an analysis of variance (ANOVA) procedure is run for persons by raters nested within tasks design. The result was a two facet G-study with raters and tasks as the facets. Raters and Tasks were used as facets in these designs because they are often important sources of variance in scores. The ANOVA results are shown in Table 10.

Table 10. *Variance Components and 4 G –studies for the (P ×r: t) Design*

Source	Total	Task A	Task B	Task C	Task D
P	0.3412213	0.03252779	0.041235	0.03357921	0.03754127
T	0.00213257	0.0042379	0.00321192	0.00391257	0.00386629
r:t	0.01334521	0.00722541	0.00635479	0.00546672	0.00495264
Pt	0.00657929	0.00749219	0.00665427	0.00572442	0.00616272
Pr:t	0.16824499	0.13137221	0.14191372	0.13925971	0.13006571
$\sigma^2(\tau)$	0.03352578	0.03216197	0.04125721	0.03662725	0.03600219
$\sigma^2(\delta)$	0.00345777	0.00394452	0.00265729	0.00312041	0.00335741
$E\sigma^2(x)$	0.03667912	0.02919543	0.03251972	0.03002921	0.03421579
$Ep^2(\delta)$	0.85743255	0.86523541	0.83925571	0.81232554	0.82919457

Table 11. *Variance Components and G –study for the (P ×r: t) for Holistic Method of Rating*

Variance Source	Component	Percentage
P	0.034392312	**14.74**
T	0.00501372	**2.14**
r:t	0.00793214	**3.40**
pt	0.00061192	**0.26**
Pr:t	0.17151294	**73.53**
$\sigma^2(\tau)$	0.06132412	
$\sigma^2(\delta)$	0.00321594	
E$\sigma^2(x)$	0.05200372	
E$p^2(\delta)$	0.94320721	

By examining the relative magnitudes of the variance components in Table 11, the relative importance of each facet and interaction to the total test score variance in each design can be seen. The relative magnitude is made easier to compare because the percentage of each variance component (relative to the sum of the variance components) is also given. In interpreting these results, it should be noted that raters are nested within tasks in the universe defined here, and therefore it is not meaningful to think about the effect of r:t as confounding of separate effects. No separate r effect exists because a score cannot be interpreted independently of the scale according to which it is scored. As displayed in Table 11, following points are worth noting:

The relatively large variance component (.034392312, or 14.74%) due to persons shows that the participants performed differently in writing. The relatively small variance component due to tasks (.00501372, or 2.14%) indicates that the tasks are of about the same difficulty. The larger variance component due to raters (.00793214, or 3.40%) indicates that, to some extent, raters vary in their ratings. The relatively small variance component due to the persons by tasks interaction (.00061192, or 0.26%) shows that, to a small degree, persons' relative performance differed across tasks. Obviously, the greatest share of variance was due to the variance component for the persons by raters interaction (.17151294, or 73.53%) which indicates that, to a large extent, persons' performance differed across raters. In other words, the participants' proficiency differed considerably across raters and somehow across tasks.

Table 12. *Variance Components and G –study for the (P ×r: t) for Analytic Method of Rating*

Variance Source	Component	Percentage
P	0.038321924	**14.72**
T	0.0051321792	**1.97**
r:t	0.009125321	**3.50**
pt	0.000421725	**0.16**
Pr:t	0.19254321	**73.99**
$\sigma^2(\tau)$	0.04255272	
$\sigma^2(\delta)$	0.003512726	
E$\sigma^2(x)$	0.049527262	
E$p^2(\delta)$	0.97002591	

The relatively large variance component (0.038321924, or 14.72%) due to persons in table 12 shows that the participants performed differently in writing. The relatively small variance component due to tasks (.0051321792, or 1.97%) indicates that the tasks are of about the same difficulty. The larger variance component due to raters (0.009125321, or 3.50%) indicates that, to some extent, raters vary in their ratings. The relatively small variance component due to the persons by tasks interaction (0.000421725, or 0.16%) shows that, to a small degree, persons' relative performance differed across tasks. Obviously, the greatest share of variance was due to the variance component for the persons by raters interaction (0.19254321, or 73.99%) which indicates that, to a large extent, persons' performance differed across raters.

Discussion

The results illustrate that the writing scores are substantially affected by rating methods and task types. Our findings are in line with the results of earlier writing performance assessment, which indicated that the scoring method and specific assignment or task type introduces a particular variance in scores (Lee et al., 2002; Schoonen 2005).

As McCutchen (1986) stated, task specific variance is induced by the topic of the text and is varied according to the amount of topic knowledge and the degree of interest in or familiarity with the topic and the rhetorical context. Langer (1984) demonstrated that knowledge cannot be considered one-dimensional. She found that the amount of knowledge needs to be differentiated from the degree of mental organization of knowledge; the latter is more helpful in writing argumentative texts and the former in writing expository texts.

Some explanations have been suggested concerning task type effects. One explanation is that discourse mode or text type influences the writing performance. For example, it is suggested that the task specific variance relates to the degree (or lack) of specification of the rhetorical context in the writing assignment (Brossell, 1983; Huot, 1990). However, in this study the purpose of writing was stated and the necessary information was given for all four tasks to the students. Factors which might have contributed as sources of errors, in this study, can be topic familiarity, test takers' interest, and test-wiseness.

In order to arrive at a more conclusive interpretation of task-specific variance and writing ability scores, studies with better experimental designs should be conducted. The tasks should be carefully designed with single facets being changed in each successive task. It should then be investigated how these changes affect the constructs that are being assessed. This might provide us with some insights into the interaction between tasks and constructs (Bachman, 2002).

This study is limited as two scoring procedures were calculated separately. It should be noted that a combination of analytic-holistic approach in one piece of writing and analysis of other traits including language use, grammar, spelling, content and organization might be a worthwhile investigation. An improvement in the scoring guides could be the use of Likert scales for scoring elements in the text. For example, 'How well is element X included in the text?' with a five-point scale for the rater ranging from '1 Not included' to '5 adequately included'. This kind of scoring could make the analytic scoring less rigid, and therefore give its raters more opportunities to weight their judgments.

The participants in this study were EFL students. Both their age and their language status might affect the external validity. These students were still at a relatively early stage of grammatical development. Their writing ability might be less stable, and therefore further research could be carried out to deal with advanced EFL learners.

In short we can conclude that the writing scores are substantially affected by facets of the writing assessment. The facets that featured in this study (task type and rating method) exerted a large influence on the score variance. Thus, we recommend that the generalizability of scores should be well established in every assessment.

References

Ackerman, T.A., & Smith, P.L. (1988). A comparison of the information provided by essay, multiple choice and free-response writing tests. *Applied Psychological Measurement, 12* (2), 117-128.

Bachman, L.F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19* (4), 453-476.

Bachman, L.F., Davidson, F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13* (2)*,* 125-150.

Bachman, L.F., Lynch, B.K., & Mason, M. (1995). Investigating variability in tasks and rater judgments in a performance test of foreign language speaking. *Language Testing, 12* (2), 238-257.

Brossell, G. (1983). Rhetorical specification in essay examination topics. *College English, 15* (4)*,* 165-173.

Brown, T.T., Levine, M.D., & White, K.P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytical scoring methods. *Educational and Psychological Measurement, 59* (4), 492-506.

Cohen, B. H., & Brooke L. R. (2004). *Essentials of statistics for the social and behavioral sciences.* New York: John Wiley & Sons, Inc.

Cooper, P.L. (1984). *The assessment of writing ability: A review of research.* NJ: Educational Testing Service.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: theory of generalizability scores and profiles. *Research Report,12,* 84-12.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Huot, B. (1990). The literature of direct writing assessment: major concerns and prevailing trends. *Review of Educational Research, 60* (2)*,* 237-63.

Langer, J.A. (1984). The effects of available information on responses to school writing tasks. *Research in the Teaching of English, 18* (1), 27-44.

Lee, Y.W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing*, *23* (2), 131–166.

Lee, Y.W., Kantor, R., & Mollaun, P. (2002). *Score reliability as an essential prerequisite for validating new writing and speaking tasks for TOEFL*. Salt Lake City: UT.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*. *22* (4), 415–437.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25*, 431-44.

McNamara, T. F., & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in offshore assessments of speaking skills in occupational settings. *Language Testing, 14*(1), 140 – 156.

Meisuo, Z. (2000). Cohesive features in the expository writing of undergraduates in two Chinese universities. Retrieved January 4, 2011 from http://rel.sagepub.com/ content/31/1/61

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: Sage publications.

Suen, H. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erbaum

Authors:

Mahnaz Saeidi is an assistant professor of English language at Islamic Azad University, Tabriz Branch. She is a Ph.D. holder in Applied Linguistics (TEFL). She is one of the editorial board members of The Journal of Applied Linguistics, and member of the Research Committee at the University. In 2007, 2008, 2009, 2010, and 2011 she won an award for being the best researcher. She has published several articles and books and participated in a number of inter/national conferences. Her major research interests are multiple intelligences, focus on form, feedback, and assessment.

Shokouh Rashvand Semiyari is a Ph.D. student of TEFL at Islamic Azad University, Tabriz Branch. She has published and participated in a number of inter/national conferences. Her major research interest is assessment.