Factors Influencing Teachers' Quality Judgments of Students' Written Work

Adem S. TURANLI[1]

Abstract

This study aimed at identifying the relationships between teachers' quality judgments about student written work and some related variables. 48 teachers working for a language school at tertiary level constituted the participants. The researcher himself wrote two essays and set up a design that would provide the required data. The study indicated that assessment strategy and the adjacency of essays assessed are likely to affect teachers' judgments of students' written work. Similar conclusions were found when the contribution of assessment strategy is controlled. The perceived quality of the essay (assigned scores) are correlated with the amount of the writing.

Keywords: *holistic evaluation, analytic evaluation, writing in a foreign language, reliability in assessment*

[1]**Yrd. Doç. Dr., Erciyes Üniversitesi, Eğitim Fakültesi** turanli@erciyes.edu.tr

Öğrencilerin Yazılı Çalışmalarına İlişkin Öğretmenlerin Değerlendirmelerini Etkileyen Etmenler

Özet

Bu araştırma, öğrencilerin yazılı çalışmalarına ilişkin öğretmenlerin değerlendirmeleriyle ilintili bazı değişkenler arasındaki ilişkileri irdelemeyi amaçlamıştır. Çalışma grubunu, üniversite düzeyinde bir dil okulunda çalışmakta olan 48 öğretmen oluşturmuştur. Araştırmacının kendisi, araştırma sorularına koşut olarak, iki makale (yazı-essay) yazmış ve gerekli verileri temin edecek araştırma desenini oluşturmuştur. Çalışma, ölçme yaklaşımı ve daha iyi olan yazının önce ya da sonra okunmasının, öğrencilerin yazılı çalışmalarına ilişkin öğretmenlerin değerlendirmelerini etkileyebileceğini göstermiştir. Ölçme yaklaşımının etkisi denetlenince de benzer sonuçlar bulunmuştur. Ayrıca, yazının algılanan niteliği (verilen notlar); okunan metnin uzunluğuyla ilişkili bulunmuştur.

Anahtar sözcükler: *bütüncül değerlendirme, çözümleyici değerlendirme, yabancı dilde yazma, ölçme güvenirliği*

1. Introduction

In order to discover students' abilities in productive skills, open-ended questions are more favored because they force students to produce, which helps teachers with their judgments about the examinee's language skills (Genesee & Upshur, 1996). Teachers generally prefer traditional evaluation methods in identifying students' achievements and many teachers complain that they cannot use alternative strategies because classes are too crowded, they lack time either for preparation or for evaluation (Gelbal & Kelecioğlu, 2007). Weigle (2002) defines reliability as "a consistency of measurement across different characteristics or facets of a testing situation. The measurement errors of any kind deriving from the factors other than the answer itself decrease the reliability (Tan, 2005). Identifying the factors likely to affect teachers' judgments of the quality of student work may create awareness and sensitivity about the sources of potential errors. However, it is a complex process to identify the criteria and the standards to be used in assessing student work in the target language in terms of its authenticity and academic efficiency (Kroll, 1990). Various contextual factors may influence the assessors' mood at the time of scoring, causes significant systematic differences among the assigned scores (Tekin, 1984). Wragg (2001) claims that even the teachers using double blind assessment procedure tend to approximate the score to the average and, thereby, decrease the reliability of the examination.

As an alternative to traditional holistic assessment, analytic assessment was developed to eliminate systematic errors in assessment. In holistic judgment, different assessors may focus on different aspects of written work although in analytic scoring, focusing on several aspects of the work separately may be time consuming and often difficult (Davies et al., 1999). Therefore, holistic assessment is a more natural process than analytic one (White, 1995). Nakamura (2002) claims that asking the raters to assess analytically will not ensure that they will really score analytically, since they may tend to adjust analytic scores to match their holistic impressions (Weigle, 2002). Especially in holistic assessment, some external features such as the gender of the examiner or the examinee may affect assessors (Goddard-Spear, 1983; Gipps & Murphy, 1994).

Several other studies have also dealt with other factors which may contaminate true scores such as the length or quantity of the essay produced predicts holistic scores (Grobe, 1981; Breland & Jones, 1984), syntactic maturity, usage, mechanics and vocabulary (Grobe, 1981), the context position of an essay in the order of assessment (Hales & Tokar, 1975; Hughes et al., 1980; Daly & Dickson-Markman, 1982), the degree to which the essay content fulfills the expectations (Tedick

& Mathison, 1995), the first impressions (Vaughan, 1992) and assessors' expectations (Stock & Robinson, 1987). Likewise, the teacher's prejudices about the student may contaminate their judgments (Wragg, 2001). It is difficult to bring assessors to a complete agreement about their quality judgments, even if a temporary agreement on common standards can be reached through training (Weigle, 2002). Weigle suggests investigating how some variables interact to influence assessors' judgments rather than dealing with individual variables. Accordingly, this study was conducted to identify how assessors' quality judgments are related to some contextual variables. The variables which are included in the study are assessment strategy, location of mistakes, the length of the language course allegedly completed, the adjacency of essays assessed (whether a better essay is assessed earlier or later), the length (amount) of the text considered before scoring. Therefore, the research questions are:

1. Do the assessors' quality judgments interact with their assessment strategy, the location of grammatical or spelling mistakes, the length of the language course allegedly completed, the adjacency of essays assessed (whether a better essay is assessed earlier or later), the length (amount) of the text considered before scoring?

2. When the assessment strategy is controlled, do the assessors' quality judgments interact with the abovementioned variables?

2. Method

2.1. Context

This study aimed to identify the factors that interact with assessors' judgments in ELT writing. In order to answer this research question, the scores assigned by the assessors (teachers) for two essays were used as dependent variables while assessment strategy, location of mistakes, the length of the language course completed, the adjacency of essays assessed (whether a better essay is assessed earlier or later), the length of the text considered before scoring were included as independent variables. The study was carried out in a school at tertiary level, where as a regulation, students who have gained a place in it attend an eight-month Basic English program through upper-intermediate level before they start their majors. Students follow an intensive program for eight months and were taught by 58 non-native English teachers, 48 of whom constituted the participants. The independent variables were included, mainly because the school administrators and program planners believed that these factors contaminated scores.

2.2. Data gathering

First, the researcher invited all the teachers working at the school to take part in the study, 52 of whom accepted to participate. Afterwards, the researcher gave them the two essays and asked them to assess the essays as they were instructed on the cover page of the essays. The essays had been prepared by the researcher and specifically equipped with some features so as to help answer the research questions. Within the same day, the researcher gathered the essays with the ascribed scores. However, during the data entry, it was discovered that four teachers had not followed the instructions properly. These teachers, who would assess the essays analytically, had assigned holistic scores. Consequently, they were eliminated, and the data from the 48 teachers were used for the analysis.

2.3. Instrumentation

Since it was not possible to find a context that could naturally present enough data in line of the research questions, the researcher purposefully prepared the essays. Initially, the researcher prepared two essays, with some mistakes purposefully included (see Appendix). The essays were intended to be at different levels (one better than the other in terms of the vocabulary and structures used). Despite the mistakes which were purposefully placed in both essays, they were named as E-Essay (elementary) and I-Essay (intermediate). E-Essay was a description of a boy's life and his

family, and I-Essay was about the importance of learning a foreign language. Ten of the teachers working for the school were asked to judge their appropriateness for the stated levels. They all commented that students of the stated levels could also write similar essays with similar syntactical and spelling mistakes. However, this was of limited concern whether since existence of the difference itself in level was not considered as a factor. Additionally, the average scores calculated showed that the essays were of different levels. The mean score for E-Essay was calculated to be 64,5 while the mean for I-Essay was 73,0. By paired-samples t-test, the difference was found statistically significant ($t=4,37$; $p<0,01$). Also, the big standard deviations in the scores imply that the teachers had highly differing quality judgments.

For each essay the researcher created two versions with slight differences (Version A and Version B in terms of the mistakes purposefully located). Each version had a unique purpose so as to provide data for a different research question. First, in order to investigate how the length of the text and the assigned scores interacted, each essay was divided into four equal parts in terms of the number of the words. The first part was named 'text 1', the first and second parts were called 'text 2', the first, second and third parts were names as 'text 3', and finally the whole essay was called 'text 4'.

Table 1

Distribution of Assessors According to Independent Variables

Essays	Versions	Course Length (hours)	Assess-ment Strategy	N
E-Essay (Elementary)	Version A (mMistakes)	150	Holistic	6
			Analytic	6
	Increasing numbers of mistakes	300	Holistic	7
			Analytic	5
	Version B (Mmistakes)	150	Holistic	7
			Analytic	5
	Decreasing numbers of mistakes	300	Holistic	6
			Analytic	6
I-Essay (Intermediate)	Version A (mMistakes)	150	Holistic	6
			Analytic	6
	Increasing numbers of mistakes	300	Holistic	7
			Analytic	5
	Version B (Mmistakes)	150	Holistic	7
			Analytic	5
	Decreasing numbers of mistakes	300	Holistic	6
			Analytic	6

Dispersion of mistakes: For each essay, two versions were constructed which were the same in essence but different in terms of the dispersion of the mistakes: in Version A, 'texts' 1 through 4 had increasing numbers of mistakes while in Version B, 'texts' had decreasing numbers of them (See Appendix). Namely, in Version A, 'texts' 1 through 4 had increasing numbers of spelling and syntactical mistakes (1, 3, 3 and 5 respectively) and in Version B, they had 5, 3, 3 and 1 mistakes and Versions A and B were were also named as mMistakes and Mmistakes respectively for simplicity and clarity)

Length of the course allegedly completed: In order to identify how the length of the allegedly-completed course and assessors' judgments interact, each of the abovementioned versions was affixed a note of 'The student has written this essay after the completion of a course of

150 (or 300) hours'. It was because 150- and 300-hour courses were referred to as 'elementary' and 'pre-intermediate' levels respectively at the school where the study was conducted.

Assessment strategy: In order to test how the assessment strategy and assessors' judgments interact, half of the teachers were asked to assess the essays holistically while the other half were to assess them analytically. In order to avoid ambiguity and confusion, the analytic and holistic assessment procedures were used in the same way they were carried out for the examinations at the school.

The holistic assessors were asked to consider the percentages of the allotted time to the writing components in the course (*given below*) while assigning a score and assign just one score to each 'text'. On the other hand, the analytic rubric was quite simple and the assessors would ascribe separate scores between 0 and 100 to each of the three components for every 'text'. Next, these scores were multiplied by the corresponding weights and the results were added to find a composite score for each 'text'.

Organization and mechanics	: %25
Vocabulary and language use	: %40
Coverage of content	: %35

Adjacency of essays (order of assessment): In order to explore how assessors' judgments interact with the level of the previously adjacent paper, the two essays (E-Essay and I-Essay) were clipped together so that half of the assessors started with E-Essay while the rest graded I-Essay first.

Length of the text: This study also aimed at identifying how the length of the 'text' interacts with assessors' judgments. Therefore, the scores assigned to 'texts' 2 and 4 were compared. 'Text' 2 covered the first half of the essay and 'text' 4 constituted the whole of it. 'Texts' 1 and 3 were not added into comparisons, just to keep analyses simpler.

Table 1 displays the distribution of the 48 assessors (teachers). Although at the phase of planning, an equal distribution had been planned, four teachers did not assess the essays properly. At the phase of assessment, the assessors were asked to assign scores for 'texts' 1 through 4 and not to do any changes backwards in the previous grades that they had already assigned, even if they thought they needed to.

3. Findings

In order to test whether there were differences between the subgroups of each independent variable, independent samples t-test was used as a statistical tool. The interactions between the pairs of independent variables were descriptively investigated, cautiously discussed, and generalizations were avoided. The mean scores and standard deviations were used for the analyses and the findings are presented below.

Location of mistakes: Readers' first impressions are commonly believed to be shaped right after starting to read an essay and not to change afterwards. In order to test this, the teachers assessed either Version A or Version B of either E-Essay or I-Essay. The assessors had not been informed about the frequency or the location of the mistakes. The difference between the two versions for either essay was insignificant ($t=0,14$; $p>0,05$ for E-Essay and $t=1,45$; $p>0,05$ I-Essay). However, the high dispersions of the scores (from 12,4 to 14,7) imply that assessors have different tendencies towards mistakes.

Length of the course completed: This study also aimed to identify whether the scores changed according to the length of the course presumably done. The mean scores for E-Essay were $\bar{x}$=70,0 and =59,0 after the completion of 150 and 300 hour courses, respectively. This difference was found to be statistically significant ($t=3,06$; $p<0,01$). However, this was not the case for I-Essay ($t=0,30$; $p>0,05$); the means for I-Essay was found to be =73,6 and =72,4 after the alleged

completion of the 150 and 300 hour courses, respectively.

Table 2

Scores According to Independent Variables Included (Means and SDs)

Features Compared	Means and SDs for E-Essay		Means and SDs for I-Essay		
	$\bar{x}$	SD	$\bar{x}$	SD	N
Overall					
	64,5	13,5	73,0	14,0	48
	t (46) = 4,37; p<=0,05				
Assessment Strategy					
Holistic	67,9	14,7	76,5	13,7	26
Analytic	60,5	10,8	68,9	13,5	22
	t (46) = 1,94; p>0,05		t (46) = 1,95; p>0,05		
Location of Mistakes					
A (mMistakes)	64,3	12,4	75,9	14,1	24
B (Mmistakes)	64,8	14,7	70,1	13,6	24
	t (46) = 0,14; p>0,05		t (46) = 1,45; p>0,05		
Course Length					
150 hours	70,0	12,7	73.6	15.2	24
300 hours	59,0	12,1	72,4	13,0	24
	t (46) = 3,06; p<0,05		t (46) = 0,30; p>0,05		
Adjacency of Essays (Order of Assessment)					
E-Essay First	66,2	9,5	74,1	12,1	25
I-Essay First	62,7	16,8	71,9	16,0	23
	t (46) = 0,92; p>0,05		t (46) = 0,54; p>0,05		
Length of the Text					
Text 2 (Half)	60,3	15,1	71,2	15,0	48
Text 4 (Whole)	64,5	13,5	73,0	14,0	48
	t (47) = 2,43; p<0,05		t (47) = 0,98; p>0,05		

Adjacency of essays (order of assessment): In order to answer whether teachers are affected by the quality of the previously assessed paper, half of the teachers assessed E-Essay first and I-Essay second; the other half assessed I-Essay first. For E-Essay, the mean was calculated as x=66,2 when it was assessed first (before I-Essay) and =62,7 when it was scored after I-Essay. This difference of 3,5 was not found statistically significant (t=0,92; p>0,05). The case was quite similar for I-Essay; when E-Essay was assessed first and I-Essay second, the mean of I-Essay was calculated to be somewhat higher, although not statistically significant (t=0,54; p>0,05). In other words, the means of both E-Essay and I-Essay were higher when E-Essay was assessed first than when it was assessed second. Another point worth consideration was big standard deviations in the scores of both E-Essay (SD=16,8) and I-Essay (SD=16,0) when I-Essay was assessed first. On the other hand, the standard deviation of E-Essay was calculated the smallest (SD=9,5) of all. This finding indicates that when an essay of a lower level is assessed after one of a higher level, assessors' judgments may be contaminated by the quality of the previous one and also, raters have more difficulty complying with the assessment criteria in scoring the higher level text.

Length of texts: In order to answer whether assessors' quality judgments about an essay chang while reading through it, each essay was divided into four equal parts, and for simplicity the scores assigned for 'text' 2 (half of the essay) and 'text' 4 were compared. Since the assessors had already been informed about the length of the complete essays, their quality judgments were expected not to change only due to the length of the 'texts'. However, the means of 'text' 2 (=60,3) and 'text' 4 (=64,5) significantly differed for E-Essay ($t=2,43$; $p<0,05$). However, the difference between 'text' 2 (=71,2) and 'text' 4 (=73,0) for I-Essay was smaller and statistically insignificant. This means that the length of the text influences assessors' judgments for students' work within elementary level and this may arise from assessors' tendency to be affected by students' efforts considered as 'good' work at earlier levels.

Interactions: This study also aimed at investigating the interactions between the assessment strategy (holistic or analytic assessment) and the other variables included in the study. In order to avoid overgeneralization which may derive from the limitations of the study, the interactions were investigated descriptively and accordingly, a further in-depth study was believed to be needed for statistical comparisons.

Table 3
Interactions of Variables Included: A Comparative Exploration

	Assessment Strategy	E-Essay		I-Essay	
		$\bar{x}$	SD	$\bar{x}$	SD
		Location of Mistakes			
Version A	H	66,2	13,7	73,1	11,1
(mMistakes)	A	62,0	10,7	66,6	15,9
Version B	H	69,6	16,0	80,0	15,5
(Mmistakes)	A	59,1	11,2	71,1	11,0
		Adjacency of Essays (Order of Assessment)			
E-Essay	H	69,6	8,2	78,2	9,6
Assessed First	A	63,6	10,0	70,9	13,3
I-Essay	H	66,7	18,3	75,3	16,3
Assessed First	A	55,1	10,7	65,4	14,1
		Course Length			
150 hours	H	76,5	9,7	80,0	16,8
	A	62,3	11,6	74,3	11,2
300 hours	H	59,2	14,0	73,1	9,1
	A	58,8	10,2	63,5	14,0

Assessment strategy and location of more mistakes: When they were sorted out according to the assessment strategy and the location of more of the mistakes for either essay, there did not appear to be big differences between the groups However, high dispersions were observed in the scores. First, the range of the means for E-Essay (10,5) was smaller than that of the means for I-Essay (13,4), indicating that for I-Essay, the interaction between the assessment strategy and the location of the mistakes in the essay was bigger, even though not statistically significant. The mean of the analytic scores for Version A of I-Essay (x=66,6) was much smaller than that (=80,0) of holistic scores for Version B. Although a similar pattern was observed for E-Essay, the lowest mean (=59,1) was calculated for analytic scores for Version B.

Assessment strategy and adjacency of essays (order of assessment): The analysis revealed that when the scores were grouped according to the assessment strategy and the assessment order

(whether E-Essay or I-Essay was assessed first), there appeared to be some noteworthy points related to the means and standard deviations. In both cases the teachers assigned higher holistic scores to both essays than when they assessed them analytically. Second, while holistic scores showed a difference of around 3 for both cases, the differences were bigger for the analytic scores. For the analytic scores, the differences were 8,5 for E-Essay and 5,5 I-Essay, and in both of the cases the scores were calculated to be higher when E-Essay was assessed first. Also, when the means were considered as a whole, for E-Essay, the means ranged from =55,1 (analytic scores when the essay was assessed after I-Essay) to =69,6 (holistic scores when it was assessed first). Similarly, for I-Essay, the means ranged from =65,4 (analytic scores when it was assessed before E-Essay) to 78,2 (analytic scores when E-Essay was assessed first). These differences in means (14,4 for E-Essay and 12,8 for I-Essay) and standard deviations (from 8,2 to 18,3 for E-Essay and from 9,6 to 16,3 for I-Essay) show that scores spread much according to the assessment strategy and the adjacency of assessment.

Assessment strategy and length of course completed: When the scores were grouped according to the assessment strategy and the length of the course allegedly completed, considerable patterns were observed. Unlike after the completion of a 150 hour course, in which holistic scores were higher than their counterparts, the holistic and analytic scores for E-Essay were fairly close to each other (a difference of 0,4) after the completion of 300 hour course. However, for E-Essay again, the difference between holistic and analytic scores were very big (with a difference of 14,2) after the completion of 150 hour course. Besides, for I-Essay, for both courses, analytic scores were calculated smaller than their holistic counterparts (a difference of 6,7 for the 150 course and 9,4 for the 300 hour course). Also, the distributions of the mean scores for the 150 and 300 hour courses looked different.

Correlations: The scores assigned for E-Essay were found significantly correlated with those for I-Essay ($r=0,52$; $p<0,01$), meaning that the teachers assessed both essays either more positively or more negatively. A teacher who assigned a higher score for one essay also assigned a higher score for the other. This indicates that teachers are inclined to assign either higher or lower scores to students' work regardless of its quality. Furthermore, with the assessment strategy being controlled, the correlation was found statistically insignificant between the analytic scores for the essays ($r=0,32$; $p>0,05$) but statistically significant for holistic ones ($r=0,59$; $p<0,00$). Given that the assessors who participated in the study had to assess both essays within a short time, the analytic assessment strategy might have helped them eliminate the subjectivity arising from the adjacency of papers.

4. Discussion And Implications

It is much harder to achieve reliability in writing tests than any other examinations. It results from the multitude of subjective elements in what is assessed (Çetin & Kelecioğlu, 2004) This study provided insights about the factors that shape assessors' quality judgments of student written work in ELT settings despite the limitation that the researcher himself prepared the essays.

First, the scores show a big dispersion when the task requires students to write an essay for the development of productive skills. Therefore, teachers are to consider the cons and pros that open-ended questions might involve. Holistic and analytic scores show considerable differences in favor of holistic ones. Nevertheless, when the length of the course taken is used as a controlling variable, the differences get bigger in favor of holistic scores. Also, no difference was found between the holistic and analytic scores ascribed after the completion of the longer course, but holistic scores were higher for the shorter course. So it can be concluded that assessors' analytic and holistic judgments are differently shaped by the length of the course that students do. This may lead to deceptive conclusions about the effectiveness of the program and, consequently, may delude decision makers. Therefore, the institutional priorities and the curriculum implemented should have a key role determining the appropriate assessment strategy. Since both holistic and

analytic assessments have cons and pros, teachers should be flexible in their preferences. Mainly, holistic scales are found faster and more efficient, and analytic scales are somewhat more reliable; however, analytic assessment will provide better feedback for students, because students are informed about their respective strengths and weaknesses in different aspects of writing (Bauer, 1981; Weigle, 2007).

When other effects are ignored, assessors' judgments do not seem to be affected by the location where mistakes abound in an essay. However, when there appear more mistakes at the beginning of the essay, assessors assign lower analytic scores. Analytic assessment may limit assessors' subjectivity by being more rigid in criteria and may prevent assessors from revising their first impressions, even if the quality of the text gets better or worse. Therefore, assessors must read the examination paper thoroughly before they assign a score and a second reading may help to decrease assessment errors.

When a student paper is assessed after a better one, the score assigned for the second one tends to decrease, and assessors' judgments disperse more than when the better work is evaluated as second. Besides, analytic scores considerably decrease when the lower-level student work is assessed first. This indicates that with a big difference between the essays in quality, the judgments about the earlier paper may shape the judgments about the following; the quality of the first paper is unconsciously taken as a point of reference and the next paper is compared to that, which may or may not be in favor of the second paper, which is in line with Hales & Tokar (1975), Hughes et al. (1980), Daly & Dickson-Markman (1982). They maintain that when a medium quality essay is read after a number of high quality essays, it will likely receive a low score or it will be given a high score if it is assessed after a number of high quality essays. In order to avoid this tendency, after first reading, teachers may need to sort out the papers according to their first impressions of the quality and then read them thoroughly and more carefully.

Although assessors consider the length of the course taken by the student whose essay is assessed, the scores are not proportional with the lengths of the courses. The assessors expect proportionally much more from the students who have finished a shorter course than those who have finished twice as long as the other. When students take up learning a foreign language, they often have unrealistically high expectations. This is also how some teachers think; they may have too high expectations from their students. The small difference, even though statistically significant, between the scores of the two groups of students who have finished courses of different lengths implies that teachers' expectations become more realistic as students reach higher levels. Since unrealistically high goals which are set in order to motivate students may lead to disappointment, both teachers and students should build realistic expectations at the very beginning of the course in order to avoid frustration.

The study shows that the scores gradually increase in accordance with the length of student work, when the other variables are controlled. This shows that teachers' judgments are affected by the amount of text on the paper regardless of its quality. Therefore, teachers should review their expectations before they start assessing the papers and eliminate the factors leading to systematic measurement errors. It was also observed that the more mistakes assessors observe in the successive parts of the essay while reading through the paper, the lower scores they assign. However, the first and final judgments for the essay with fewer mistakes in the second part do not show any considerable change. This indicates that assessors' first impressions at the beginning of the student work hardly change. Thus, it can be said that any work that builds positive first impressions ends positively, even if its quality gets poorer. The teacher should be helped to overcome some of the problems related to reliability; teaching them how to use rubrics will pay off (Çetin & Kelecioğlu, 2004). Alderson (2001) suggests that examiners should understand the principles behind the rating scales with which they work.

Despite these findings and their implications, the results should be carefully dealt with

because of the limitations of the study. First, it was carried out with teachers at the tertiary level. Second, the number of the teachers was limited to 48, which made the means vulnerable to measurement errors. Therefore, the findings related to the analyses of significance need to be cautiously accepted. Another limitation was with the number of essays used; two purposefully invented essays were used to depict teachers' attitudes in assessment. Even if the number of the essays (two) may be assumed to be sufficient for the aim of this study, further studies with more deliberate designs (both qualitative and quantitative) may provide further information on how assessors are influenced by various factors.

References

Alderson, J. C.; Clapham, C. & Wall, D. (2001). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Bauer, B. A. (1981). A study of the reliabilities and cost-effeciencies of three methods of assessment for writing ability (ERIC Document Reproduction Service No. ED 216 357).

Breland, H. M. & Jones, R. J. (1984). Perception of writing skills. *Written Communication* 1 (1), 101-19.Daly, J. A. & Dickson-Markman, F. (1982). Contrast effects in evaluating effects, *Journal of Educational Measurement*, 19 (4) 309-16.

Çetin, B & Kelecioğlu, H. (2004). The relation between scores predicted from structured features of essay and scores based on scoring key and overall impression in essay type examinations. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education)*, 26, (2004), 19-26.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Gelbal, S. & Kelecioğlu, H. (2007). Teachers' proficiency perceptions of the measurement and evaluation techniques and the problems they confront. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education),* 33, 135-145.

Genesee, F. & Upshur, J. A. (1996). *Classroom-based evaluation in second language education.* Cambridge: Cambridge University Press.

Gipps, C. & Murphy, P. (1994). *A fair test? Assessment, achievement and equity*. Milton Keynes, Open University Press.

Goddard-Spear, M. (1983). *Sex bias in science teachers' ratings of work.* Contribution to the second GASAT Conference, Oslo, Norway.

Grobe, (1981). Syntactic maturity, mechanics, and vocabulary as predictors of writing quality. *Research in the Teaching of English,* 15, 75-85.

Hales, L. W. & Tokar, E. (1975). The effects of quality of preceding responses on the grades assigned to subsequent responses to an essay question. *Journal of Educational Measurement*, 12, 115-117.

Hughes, D. E; Keiling, B. & Tuck, B. F. (1980). The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement*, 17, 131-135.

Kroll, B. (1990). *Second language writing: Research insights from the classroom*. Cambridge: Cambridge University Press.

Nakamura, Y. (2002). A comparison of holistic and analytic scoring methods in the assessment of writing. Retrieved from the Web 16.1.2007. .

Stock, P. L. & Robinson, J. L., 1987. Taking on testing: Teachers as researchers. *English Education*, 19, 93–121.

Tan, Ş. (2006). *Öğretimi planlama ve değerlendirme*, Ankara: Pegem Yayıncılık.

Tedick, D. J. & Mathison, M. A. (1995). Holistic scoring in ESL writing assessment: What does an analysis of rhetorical features reveal? In: *Academic writing in a second language: Essays on research and pedagogy*. Norwood: Ablex Publishing Corporation, 205–230.

Tekin, H. (1984). *Eğitimde ölçme ve değerlendirme*. Ankara: Has-soy Matbaası.

Vaughan, C. (1992). Holistic assessment: What goes on in the rater's mind? In: HampLyons, L., Editor, *Assessing second language writing in academic contexts*, Norwood: Ablex, NJ, 111–126.

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

Weigle, S. C. (2007). Teaching writing teachersaboutassessment. *Journal of Second Language Writing*, 16 (3), 194-209.

White, E. M. (1995). An apologia for the timed impromptu essay test. *College Composition and Communication*, 46 (1)*,* 30-45.

Wragg, E. C. (2001). *Assessment and learning in the secondary school*. Florence, KY, USA: Routledge.