Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [H. U. Journal of Education] 28 (1), 369-381 [2013]

# A CASE STUDY ON INSTRUCTORS' PERCEPTIONS OF WRITING EXAM GRADING CRITERIA

# ÖĞRETİM ELEMANLARININ YAZMA SINAVI DEĞERLENDİRME ÖLÇÜTLERİNE İLİŞKİN ALGILARI ÜZERİNE BİR DURUM ÇALIŞMASI

Yeşim TARKAN-YELOĞLU[*], Gölge SEFEROĞLU[**], H. Okan YELOĞLU[***]

**ABSTRACT**: This study was conducted to analyze the instructors' perceptions of the writing exam grading criteria used in the Faculty Academic English within the context of Freshman English courses at a private university in Turkey. Fifty-five instructors were involved in the study. The data were collected via quantitative and qualitative data collection instruments. Close-response items provided quantitative data and the qualitative data were derived from open-response items. The results indicate that the instructors believe the criteria help to establish standard grading across the program. However, they still have some doubts about the way the criteria are applied across the program while assessing students' writing. It is noteworthy that the instructors in this study had different perspectives and approaches while using the criteria in their own settings. Therefore, the results of this study highlight a crucial need for training the raters on how to apply any grading criteria to ensure objectivity in student assessment.

**Keywords:** Academic writing, assessment criteria, English, instructors, perceptions

**ÖZ:** Bu çalışma, Türkiye'de özel bir üniversitede öğretim elemanlarının, İngilizce akademik yazma dersi için kullanılan yazma sınavı değerlendirme ölçütleri hakkındaki görüşlerini ortaya çıkarmayı amaçlamıştır. Araştırmaya 55 öğretim elemanı katılmıştır. Veriler kapalı ve açık uçlu soruların bulunduğu bir anket ile gönüllü öğretim elemanları ile yapılan mülakatlardan elde edilmiştir. Anketteki kapalı uçlu sorulardan nicel veriler, anketteki açık uçlu sorular ile yapılan mülakatlardan ise nitel veriler sağlanmıştır. Sonuçlar, öğretim elemanlarının genel olarak var olan ölçütlerin Akademik İngilizce Programı içinde standart bir değerlendirmeyi sağladığını düşündüklerini göstermektedir. Ancak, ölçmede kullanılan kategorilerin eşit olarak puanlandırılmaması ve ölçütlerin bütün öğretim görevlileri tarafından aynı şekilde kullanılmadığı yönünde kaygılar mevcuttur. Öğretim elemanlarının ölçütleri kullanırken farklı bakış açıları ile hareket ettikleri görülmüştür. Elde edilen bulgulara dayanılarak değerlendirme ölçütlerinin güvenirliğini artırmaya yönelik değerlendiricilerin eğitilmesi gibi çeşitli önerilerde bulunulmuştur.

**Anahtar sözcükler:** Akademik yazım, değerlendirme ölçütleri, İngilizce, öğretim görevlileri, görüşler

## 1. INTRODUCTION

Writing skill is one of the most important components of learning a language since constructing even a single sentence shows how well a student has mastered the target language. It is one of the ways that reflect how much progress students have made in learning the new language because it is a productive skill which requires some deeper processing. The importance of the ability to write effectively has increased more "as tenets of communicative language teaching - that is, teaching language as a system of communication rather than as an object of study – have taken hold in both second-and foreign- language settings" (Weigle 2002, x). As a result, since writing has become more important, there is a greater demand for valid and reliable ways to test writing ability. This is necessary not only for classroom use but also as a predictor of future professional or academic success. In other words, assessing writing plays an important role in every class where students are asked to write. Evaluating students' writing is quite a challenging task for English teachers. Assessment of writing ability is of crucial importance not only for teachers but also for students since many important decisions are made on how well they communicate in writing and such decisions affect students' education and even their lives (William 1996; Brown 1996; White 1994; Bektas ve Sahin 2007; Sahin

[*] Instructor, Bilkent University School of English Language, tyesim@bilkent.edu.tr
[**] Prof. Dr. Middle East Technical University, Faculty of Education, Department of FLE, golge@metu.edu.tr
[***] Assoc. Prof. Dr. Başkent University, Faculty of Administrative Sciences, Department of TKM, okany@baskent.edu.tr

2007; Seferoğlu 2010). As Lumley (2006, 23) highlights "the pursuit of reliability in assessing writing performance became a central concern". Considering the context of the study, it can be claimed that accuracy and the reliability in the assessment of students' writing have utmost importance as every year approximately 63 to 65 instructors assess approximately 1000 to1600 students' writing papers in the Faculty of Academic English in freshman English courses. Depending on whether they pass or fail, they continue their study in their departments. It may be considered as a 'high-stakes' exam within its context as it is "likely to have a major impact on the lives of large numbers of individuals or on large programs". (Coombe, 2007, xix) Moreover, in broader perspective, ensuring that students become competent and fluent writers in EAP is aimed for in the programme.

### 1.1. Literature Review

Assessing students' writing is not an easy task since "examiners are required to make judgments which are more complicated than the 'right – wrong' decisions…" (Alderson et al 1995, 107). Testing students' writing ability in a reliable, valid and fair way is very crucial and the success lies in being able to assess something subjective as objectively as possible.

Testing and assessing writing is challenging due to inherent difficulties. There are certain basic considerations in assessing writing such as task variables, test –taker variables, rater variables, and rating scales (Bachman & Palmer 1996). Assessing writing requires subjective judgments on the part of raters; thus, teachers' perceptions of writing assessment and writing assessment rating scales are important. Coombe (2007, xviii) also states that "..., a subjective test, such as writing an essay, requires scoring based on opinion or personal judgment, so the human element is very important".

As mentioned, another point to be taken into consideration is the rating scale. As Park (2004, 1) confirms, "one of the first decisions to be made in determining a system for directly assessing writing quality is what type of scoring procedure will be used". Although there are some others, three types of scoring procedures have been mainly discussed in the literature: Analytic, holistic and primary - trait (Bachman & Palmer 1996; Weigle 2002; Alderson, Clapham & Wall 1995). Klimova (2011, 391) also confirms that "the most common evaluation methods include holistic and analytic". All of them have advantages as well as disadvantages when they are applied. Considering the facts mentioned above, many researchers claim that no test or composition scoring procedure is perfect. As Perkins (1983) also states, the thing to be done is trying to find the best way for the context one has as no test or scoring procedure is suitable for all purposes. Another point that he makes and which is important to keep in mind is that "Even with guidelines and set criteria, the analytical and holistic scoring schemes can produce unreliable and invalid test information" (666).

As it has been highlighted before, raters have utmost importance while assessing students' papers. As raters use rating scales for assessing writing performance, when designing an effective rating scale, raters' perceptions of writing proficiency and well- worded and comprehensive descriptors that represent the construct of writing ability should be used (Lumley 2002). Knoch (2011, 82) also agrees with Lumley (2002) as he says "raters often seem to struggle when employing these types of scales". Moreover, as Wharton (2003), in her study where she aimed to define appropriate criteria for the assessment of Master's level TESOL assignments claims, group participation in the development of assessment practices is invaluable because it enables everyone to stand by the results. She also invited course participants – teachers with at least 3 years experience- to comment on the usefulness or otherwise of the assessment criteria.

Last but not least, Huang, J. (2012, 124) claims that "the rating methods (holistic versus analytic) used by the raters can change their application of rating criteria in the assessment of ESL writing..." Considering all the literature, this study investigates instructors' perceptions of the freshman English (ENG 101) writing exam grading criteria which is used to assess students' academic writing skills in the final ENG 101 exam in the Faculty Academic English program at a private university in Turkey.

# 2. METHOD

## 2.1. Purpose of the Study

This study has been designed to investigate the instructors' perceptions about the ENG 101 writing exam grading criteria used to assess students' academic writing skills in the final ENG 101 exam in FAE program at a private university. This study will specifically address the following research questions:

1. How do Eng 101 instructors perceive the common ENG 101 writing exam grading criteria in terms of the following dimensions; Overall effectiveness, Categories, Descriptors, Participants' feelings about its application

2. How would instructors mark the paper when a student's paper matched the B band in two categories but merits a C- band in the other two?

3. What do ENG 101 instructors perceive as positive attributes of the common ENG 101 Writing Exam Grading Criteria?

4. What do ENG 101 instructors perceive as negative attributes of common ENG 101 Writing Exam Grading Criteria?

5. What are the participants' suggestions for improving the ENG101 Writing Exam Grading Criteria.?

## 2.2. Background to the Faculty Academic English Program, the Freshman English Course, and the Writing Exam Grading Criteria

The Faculty Academic English Program (FAE) provides English support courses to students in their faculties and schools. The courses offered by the FAE units range from content-based, academic skills courses in the freshman year to graduate writing courses for MA and PhD students. In providing academic skills support to a wide range of students in diverse faculties, instructors in the FAE program work in coordination to design meaningful courses which emphasize high standards of academic writing achievement through challenging materials, active classroom learning, individual tutorial support and extensive feedback on student productions. In addition, in order to meet the needs of specific departments, instructors often work closely with the department staff. The current organization of the post-preparatory programs was established in January 2003 after the teaming up and merging of the First Year English Program with post-preparatory programs in the school of English language. There are currently five FAE units, each with approximately 15 teachers responsible to a head, grouped according to the faculties or schools which they serve.

ENG 101 course, which students have to take as an obligatory course in their first year, aims to introduce students to an academic approach to thinking, reading, speaking and writing in an integrated, meaningful manner so that they are able to apply the skills learnt in their departmental studies. In addition, the ENG 101 course aims to further develop the students' linguistic accuracy and range in English. To this end, there are many objectives to be covered in ENG 101. These objectives are grouped under the headings as academic thinking, reading, discussion /presentation, writing, and linguistic accuracy and document formatting.

In this study the main focus will be on the writing objectives which include academic writing, linguistic accuracy and document formatting.

As stated before, FAE consists of the following units;

Faculty of Engineering and Faculty of Science Unit (FAE-FE / FS)

Faculty of Economics, Administrative and Social Sciences Unit (FAE – FEASS)

Faculty of Humanities and Letters, Faculty of Art, Design and Architecture Unit FAE – FHL / FADA)

Faculty of Business Administration, Faculty of Law Unit (FAE - FBA/FL)

Faculty of Music and Performing Arts, School of Tourism and Hotel Management, and the Vocational Schools of Computer Technology, Office Management, and Tourism and Hotel Services Unit (FAE - VTS/FMPA)

For each unit ENG 101 course objectives are the same. This fact leads to the need for a set of standardized criteria to be used in each unit in order to be fair to students while assessing their progress – in this context academic writing skill is focused on. In the past, each of the five units had different criteria and this situation resulted in inconsistencies in assessing students' performance and this was not something desired for the course ENG 101. To avoid this, the director of FAE felt the need for establishing standard writing criteria across the units. Then, from each unit the writing criteria used for ENG 101 were taken and after many interviews with the heads of the departments and instructors, a new set of criteria was designed. Having finalized the new criteria, the new criteria were launched at the beginning of 2004-2005 academic year.

### 2.3. The Participants

55 instructors out of 64 were involved in the study. Not all the instructors were involved as, during the administration of the questionnaire session, they were teaching summer school and they could not attend the session. One of participants was the head of the FAE program. The other five are the heads of each unit and the rest are the instructors who give ENG courses to the students at the departments. Out of 55 instructors, 24 of them were male and 31 female. Twenty seven instructors were native and 28 non-native. Four of the instructors had a PhD degree whereas 41 a BA degree. Their experience in the FAE program ranged from 4 months to 16 years. Although all 55 instructors seemed to be answering the questionnaire during the administration of the questionnaire, , it was noticed while analyzing the data that 5 of them had only filled in the first section from which demographic data was gathered. This means they did not fill in the rest of the questionnaire stating that they had not used the ENG 101 writing exam grading criteria as they had not taught ENG 101 course since the new criteria were launched. As a result, the data analysis was conducted based on 50 instructors' responses.

### 2.4. The Instruments

The data were collected via quantitative and qualitative data collection instruments. The questionnaire designed provided both qualitative and quantitative data. The interviews held also provided further qualitative data. Fifty instructors were given the questionnaires and 6 of them were interviewed to get their perceptions on the criteria in detail.

In this section, the instruments of the study are described.

#### 2.4.1. The Questionnaire

In order to collect data on FAE instructors' perception of the ENG 101 Writing Exam Grading Criteria, a questionnaire was designed by the researcher considering the categories under which the feedback was planned to be taken. While writing the items, the relevant literature was taken into consideration since questionnaires are widely used and useful instruments for collecting survey information, providing structures, often numerical data, its administration not requiring the presence of the researcher, and often being comparatively straightforward to analyze (Wilson and Mclean 1994, cited in Cohen et al 2000).

The questionnaire had 4 parts. The first part asked for biodata about respondents' background and individual characteristics. The second part was made up of closed-response items using the Likert scale. In this part, the Likert scale was used as it is "generally useful for getting at respondents' views,

judgments, or opinions…" (Brown and Rodgers 2002, 120). In this research a 1 to 4 scale was used (1- Strongly Agree, 2- Agree, 3- Disagree, 4- Strongly Disagree) as the respondents were expected to state their perceptions as positive or negative rather than being noncommittal. Although closed-response items are mostly preferred in questionnaires as "they are quick to complete and straightforward to code and do not discriminate on the basis of how articulate the respondents are (Wilson and McLean, 1994, cited in Cohen et al 2000), a box was added next to each item to enable the participants to write or make extra comments about each statement to express themselves further. In the 3rd and 4th sections, there were open-response items where the participants could express their thoughts and opinions more freely in a detailed way.

Before administering the questionnaire, the items were written keeping some key points in mind such as things to avoid in writing good survey items (Brown and Rodgers 2002; Bailey 1994; Cohen et al 2000). Even though some questions can be seen as overlapping or repetitive, the aim by having such items or sections was to have 'reliability check question pairs' (Bailey 1994, 134). In the second section, the Likert scale was preferred as "rating scales are particularly useful for tapping attitudes, perceptions and opinions of respondents." (Cohen et al 2000, 255) The questionnaire consisted of not only a scale but also open ended questions as "a questionnaire might be tailored even more to respondents by including open-ended questions to which respondents can reply in their own terms and own opinions", (Cohen et al 2000, 255). All the items in the questionnaire and in the interview questions were grouped to get feedback from the instructors under the following categories;

- Overall effectiveness
- Categories
- Bands
- Descriptors
- Match between ENG 101 course writing objectives across the FAE program and the criteria
- Suggestions for improvement

In the questionnaire, the first eight questions were designed to find out the overall effectiveness of the criteria. Items 9 and 10 were to get instructors' opinions about the categories in terms of their weighting and match with the course writing objectives. The next four items aimed to get feedback specifically on descriptors in each category of each band. Finally, the last four items were asked to see how the instructors themselves and others across the program feel about the application of the criteria. In the next section, Section C, a scenario was given to find out how they use the criteria while marking. The aim here was to see if they apply the criteria in the same way or not while marking in the given situation.

Section D aimed to get instructors' positive and negative perceptions on the criteria by asking them to identify the strengths of the criteria as well as the points to reconsider. The last section, Section E, was designed to see what the instructors would suggest to improve the criteria.

Then, as a next step the questionnaire was piloted. This was mainly to increase the reliability, validity and practicality of it (Oppenhaim 1992; Patton 1990; Brown and Rodgers 2002). "A common way to do this is to have someone look at the content and format of the instrument and judge whether or not it is appropriate" (Fraenkel and Wallen 2000, 171). In this research the questionnaire was given randomly to some instructors to have a look and make comments regarding the clarity of the questionnaire items, instructions and layout without actually answering it. As well as the feedback from instructors, two experts from the field of English Language Education were also consulted during this stage. The aim for this was to check face and content validity of the instruments.

Having followed the key points while preparing a questionnaire (i.e. avoiding leading, complex, irritating questions negatives etc.) (Oppenhaim 1992; Brown and Rodgers 2002; Patton 1990; Bailey 1994) and having made the necessary changes, the questionnaire was administered to 55 instructors. After administering the questionnaire and entering the data into the statistical program, the reliability

of the questionnaire was found to be at the Cronbach's alpha level 0, 91 which proves that its reliability is high.

### 2.4.2. The Interviews

As the main aim of this study was to get FAE instructors' perceptions of the ENG 101 Writing Exam Grading Criteria, a survey was carried out. As Brown & Rogers (2002) claim surveys typically take the form of interviews or questionnaires or both. This is why along with the questionnaire, the researcher carried out interviews. In other words, the aim here was triangulation since triangulation is something desirable in the research as viewing the same phenomena from multiple perspectives is possible in this way. (Brown and Rodgers 2002; Bailey 1994; Cohen et al 2000)

In this study, for the interviews, open-ended questions were prepared based on the items in the questionnaire. Later, a few more questions were added having analyzed roughly the common points that the instructors raised in the questionnaire. The aim of preparing the questions beforehand was to establish the reliability of the interviews as "one way of controlling reliability is to have a highly structured interview, with the same format and sequence of words and questions for each respondent" (Silverman 1993; cited by Cohen et al 2000, 121).

The instructors who took part in the interview were volunteers. During the administration of the questionnaire, a piece of sheet was passed around and the instructors who volunteered filled in the chart on the paper by writing their full name, e-mail address and phone number so that the researcher could contact them. In total there were 12 instructors who volunteered but when they were called back, only 6 of them were able to arrange time for the interview. The interviews lasted from 20 to 35 minutes. Interviews were held individually and tape-recorded for future reference.

### 2.4.3. Data Analysis

Quantitative analysis was done for the first and the second sections of the questionnaire using SPSS statistical program. As for the data for the open-response items in the questionnaire and the interview questions answers were subjected to content analysis and common themes was determined in the participants' responses (Miles and Huberman 1994).

Data for this research was gathered through the questionnaires which contained both closed-response items and open-response items. Apart from the questionnaires, interviews were carried out. As the questionnaire had both closed–response items and open-response items, the data analysis for the questionnaire was done both quantitatively and qualitatively. The first two sections of the questionnaire were analyzed statistically using the relevant data analysis program. For the first part of the questionnaire, descriptive statistics of bio-data, frequency analysis and missing data analysis were done. Moreover, for the second part of the questionnaire, the reliability analyses were done. For the open-response questionnaire items and the interview data, descriptive categories, i.e. headings, were developed from the data itself. To do this, all the responses for the questionnaires and the interviews on the sheets were transferred to the computer and under each heading recurring themes were noted down.

## 3. RESULTS

**1. How do ENG 101 instructors perceive the common ENG 101 writing exam grading criteria in terms of the following dimensions; Overall effectiveness, Categories, Descriptors, Participants' feelings about its application?**

The first research question in this study was about how ENG 101 instructors perceive the common ENG 101 writing exam grading criteria in terms of the following dimensions of overall effectiveness, categories, descriptors and participants' feelings about its application.

Regarding all the findings from the questionnaires and the interviews, it may be concluded that most instructors (80 %) were generally satisfied with the new criteria. They believe that the criteria help to have standard grading across the FAE program. However, they still have some doubts about the way that the criteria are applied across the program while assessing students' papers. This fact cannot be denied as some (44 %) instructors, as reported in the examples below, claimed that they have a different approach while using the criteria in their units:

> "We grade differently in our unit. Weighting goes from left to right
>
> in terms of priority, so it depends. If we consider that they are equal, it should be C+ or C" (Respondent 16).
>
> "Because not all bands / categories have equal weight in our unit it would depend on which areas were higher or lower. Also there are specific penalties for such errors as plagiarized passages, no works cited pages etc" (Respondent 28).

Since this is the case in one or more units, this is a serious issue to be resolved. This fact totally contradicts the aim of having such common criteria across the units in the FAE program.

**2. How would instructors mark the paper when a student's paper matched the B band in two categories but merits a C- band in the other two?**

When responses to the second research question which is about the way the instructors mark the papers, are taken into consideration, quite different approaches are adopted. This finding also correlates with what Lumley (2006, 20) states. "... the rating scale is inadequate ...,and that as a result, raters are forced to adopt a range of strategies to help them manage the process.."

When the recurring answers are analyzed, it is seen that a vast majority of the instructors consider the weighting of the categories in a different way and in one way or another they take an average to give a final grade to the paper.

**3 & 4. What do ENG 101 instructors perceive as positive and negative attributes of the common ENG 101 Writing Exam Grading Criteria?**

Although being satisfied with the criteria in general in terms of overall effectiveness, descriptors bands and so on. which can be regarded as positive attributes of the criteria, the difference among instructors in the way they apply the criteria can be considered as a negative attribute of the criteria. These findings match with Lumley's (2002) conclusions after his study to find out what assessment criteria really mean to the raters:

> …although there appears to be some evidence that the raters understand the rating category contents similarly in general terms, there is also evidence that they sometimes apply the contents of the scale in quite different ways. They appear to differ in the emphasis they give to the various components of the scale descriptors (p. 266)

In this study, two data collection techniques a questionnaire and interview were used to find out the instructors' perceptions of the ENG 101 Writing Exam Grading Criteria. The aim of using two different techniques was to have methodological triangulation (Brown and Rogers 2002). When the data from the questionnaires were compared with the data from the interviews, it was seen that they were consistent and parallel to each other. In both, it was found that in general the instructors were happy with the criteria and they were all aware of the rationale behind having common criteria for ENG 101 course writing exam. In both, they stressed the importance of standardization and having a common understanding across the program. In terms of the categories in the questionnaire and in the interviews they stated that they really did not like the idea of having all the categories equally weighted. This may be regarded as the major point to be considered about the criteria. In other words, this may be seen as one of the negative attributes of the criteria. Although the results in both seem to be parallel, there was an interesting point about the descriptors. In the questionnaire, 88% of the instructors stated that the descriptors were easy to understand. However, in the interviews almost all of

them stated that the some descriptors were confusing and needed to be revised. This may be because when they were filling in the questionnaires, they just roughly expressed their perception of the descriptors. On the other hand, during the interviews they had more time to look at the descriptors in detail and so were able to tell more about the quality of the descriptors. Finally, when the participants' feelings about the criteria were focused on, the results matched to a great extent. In both, they stated that they themselves feel confident about using the criteria appropriately but they are not sure about their colleagues since they observed gaps while marking. These findings are summarized in Table 1.

**Table 1. The Distribution of the Responses regarding the Positive and Negative Attributes of the Criteria as Perceived by the Participants**

| | Positive | Negative |
|---|---|---|
| **Overall Feedback For The ENG 101 Writing Exam Grading Criteria** | Effective guide for both students and teachers as it sets out the most important points to consider. (4)<br>Good for standardization across the FAE program.(12)<br>Simple not complicated.(7)<br>Saves time. (3)<br>Should be a model for other tasks / assignments. | Should be more specific and less open to interpretation more simplified.<br>More detailed criteria would be better. |
| **Categories in the ENG 101 Writing Exam Grading Criteria** | | They shouldn't have equal weighting. (14)<br>More emphasis on thesis, topic sentence, development ideas, transitions, conclusion.<br>More points should be allocated for content and organization. (12)<br>The breakdown of each category needs to be revised. The content, organization and language parts should be given a higher percentage in 101. (8) |
| **Bands in the ENG 101 Writing Exam Grading Criteria** | Good description and various levels of proficiency. (3)<br>Enables us to discriminate the borderline pass and fail papers.<br>Good to have C- defined. | I am not satisfied with F band. (5)<br>More discrimination within the F band.<br>Grade by numbers not letters. |
| **Descriptors in the ENG 101 Writing Exam Grading Criteria** | They are satisfactory/ ok. (11)<br>Good in general –still need to be fine tuned.<br>Clear descriptors to fairly evaluate students' products. | Sometimes, the difference among the descriptors is notably slight.<br>C pass is not clear, open to interpretation.<br>Some points could be added.<br>Somehow open to interpretations.<br>The descriptors between categories can be more precise. |
| **Match Between the ENG 101 Course Objectives and the Descriptors in the Criteria** | Reflects the objectives.(13) | There should be more emphasis on introduction, paragraphing and transitions |

## 4. RECOMMENDATIONS

### 4.1.For the Assessment of Writing in General

The instructors can receive training in small groups. Lumley (2002) also highlights the importance of the idea of training. According to him, training plays an important role in influencing raters' behaviors, especially by clarifying rating criteria. When the ENG 101 instructors' suggestions for training are taken into consideration, the literature also supports this idea. Weigle (1994 cited in Lumley 2002) found that rater reliability increased as a result of training and that improved agreement was the result of raters gaining better consensual understanding of the terms and levels represented in the scale. As Weigle (in Coombe et al. 2012, 220) states scholars have looked at the differences between raters with or without specific training. The results have shown that the characteristics of the

raters can have significant effects on their scoring. However, the effects can be minimized through training raters to adhere to these criteria.

Standardization sessions can be held in groups to mark sample papers against the criteria. Group discussions can be held for setting the expectations and making clarification for the raters. In terms of standardization Gottlieb (2012, 75) believes " In essence, the standards themselves are the foundation and source of content validity for the related large-scale test. Language standards also anchor classrooms assessment of students' language development. As the identical set of language standards are the grounding for multiple measures for English learners, educators are becoming more attuned to the value of gathering a body of evidence to create defensible data for student performance within a comprehensive assessment system".

Even papers can be marked by more than one instructor. Coombe (2007, 84) states that "All reputable writing assessment programs use more than one rater to judge essays".

### 4.2. For the Criteria

Not to have the categories in equal weighting. The results show that almost all of them believe that 'content' is the highest priority to be achieved by the students. This means 'content' requires a higher grade or percentage in the criteria. Huang (2012,125) also shared a similar finding. "Unlike ESL faculty raters, the English faculty raters seem to give more weight to overall content and quality of ESL writing than they do to language use... " (Soung & Caurso,1996) Based on the results, it is suggested that 'content' may have the highest weighting , followed by 'organization' and 'language' respectively.

Slight changes need to be made in the wording of the descriptors. As one of the instructors in the interview exemplifies, some of the adjectives used are quite similar to each other.

> "Although descriptors are ok in general very few need to be reworded. For example, here, 'powerfully' and here 'thoughtfully'….The distinction needs to be made clearer…"(Interviewee 1 )

### 4.3. Impact of the study on the Assessment Dimension of the programme

The main and most important aim in designing these criteria was to have a common understanding of the writing objectives across the program and assessing students in the same way with set criteria. Furthermore, there is a need for training the instructors on how to apply the criteria. Almost all the instructors support this idea. Although they had already been given training once, they believe that it was not effective.

Interviewee 1 recommends that "instead of giving training to huge groups of instructors altogether in a hall, as many sample papers as possible should be marked in small groups so that we can come to an agreement" and "the ground rules for the criteria should be set by the trainers but should not be open to discussion". Thus, in the FAE properly designed standardization sessions could be conducted in order to ensure that "raters use the scale appropriately and consistently" (Weigle 2002, 108, Akbıyık et al. 2013; Seferoğlu 2007). For standardization, first, the leader or preferably a team should read through the scripts to find anchor/ benchmark scripts that exemplify the different points on the criteria. In this context, the head of FAE and the heads of the units could come together and decide on the anchor scripts. It would also be helpful to include in the training sets scripts that exemplify certain problematic situations, for example, scripts that do not respond to the task or simply copy the prompt, or scripts that represent the borderline between two critical levels such as pass and fail. It would be important that anchor papers illustrate the nuances of the criteria. Next, other instructors may be asked to use the criteria and the anchor papers to evaluate a sample set of responses. Any discrepancies between the scores that are assigned by the instructors should be discussed. The discussions could be done in groups. However, it should be noted that it is virtually impossible to get a large group of raters to agree on exact scores and that some disagreement is inevitable. In case of extreme disagreement or discrepancy a third rater can be consulted as Coombe (2007) suggests. As well as in groups, the raters may also be asked individually to justify why they

assign that score to the script. Last but not least, raters who consistently rate higher or lower than the rest of the group should be given feedback and perhaps additional training to bring their scores into alignment with the rest of the group (Weigle, 2002).

## 5. DISCUSSION AND CONCLUSION

Based on the responses to the open-response questionnaire items and the interview data, it can be concluded that the instructors believe that the criteria help to have standard grading across the FAE program. However, they still have some doubts about the way that the criteria are applied across the program while assessing students' papers. Although being satisfied with the criteria in terms of overall effectiveness, descriptors bands and etc. which can be regarded as positive attributes of the criteria, different approaches among instructors in the way they apply the criteria can be considered as negative attributes of the criteria. All the participants seem to agree that the main and most important aim in designing these criteria was to have a common understanding of the writing objectives across the program and assessing students in the same way with set criteria. As Lumley (2006, 240) also states "...in addition to the scale, the process relies on training, experience, professionalism and acceptance of the institutional requirements to allow raters to conform in the required manner"

Based on the findings of this study, stakeholders can make the necessary changes to improve the criteria and use it more efficiently. For further research, another study can be conducted to assess the reliability of the criteria since "the two forms of reliability that are typically considered in classroom assessment and in scoring rubric development involve rater (or scorer) reliability. They are interrater and intrarater reliability" (Moskal and Jon 2000, 7). Hence, the interrater reliability and intrarater reliability of the ENG 101 writing exam grading criteria could be studied in another research.

In conclusion, if similar research is to be carried out, some suggestions can also be made. Although the results were quite satisfying and motivating related to the recently launched criteria, more accurate feedback could have been obtained if the instructors were asked to mark same papers under the same conditions and the grades could be compared and discussed. Due to some constraints such as time and human resources, such a study could not be added to support the idea that the ENG 101 writing exam criteria is reliable. Last but not least, in terms of instruments used 'Think- aloud protocols' could have been used as they may have allowed analyses of such mental processes as the sequence of rating, the interpretations the participants make of the scoring categories in the criteria and the difficulties raters face in rating etc.

It is noteworthy that the instructors in this study had different perspectives and approaches while using the criteria in their own settings. Therefore, the results of this study highlight a crucial need for training the raters on how to apply any grading criteria to ensure objectivity in student assessment.

## REFERENCES

Akbıyık, C., Karadüz, A., & Seferoğlu, S. S. (2013). Öğrencilerin internet ortamında kullandıkları yazılı sohbet dili üzerine bir araştırma. *bilig*, 64, 1-22.

Alderson, J. C., Clapman, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: CUP.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bailey, K. D. (1994). *Methods of social research*. New York: The Free Press.

Bektas, E., & Sahin, A. E. (2007). İlköğretim beşinci sınıf öğretmenlerinin soru-yanıt tekniğini kullanım davranışlarının analizi. *Egitim Araştirmalari-Eurasian Journal of Educational Research*, 28, 19-29.

Brown, J. D. (1996). *Testing in language programs*. Upper Saddle River: NJ: Prentice Hall Regents.

Brown, J. D., &Rodgers, S. T. (2002). *Doing second language research*. Oxford: Oxford University Press.

Cohen, L., Manion, L., & Morrison, K. (2000). *Research methods in education*. New York: Routledge Press.

Coombe, C., Folse, K., & Hubley, N. (2007). *A Practical Guide To Assessing English Language Learner*.USA:The University of Michigan Press.

Coombe, C., Davidson, P., O'Sullivan, B., & Stoynoff, S. (2012). *The Cambridge Guide to Second Language Assessment.* USA: Cambridge University Press.

Fraenkel, J. R., & Wallen, N. E. (2000). *How to design and evaluate research in education*. New York: McGrawHill.

Gottlieb, M. (2012). An overview of language standards for elementary and secondary education, In C. Coombe, P. Davidson, B. O'sullivan, S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 74-82). New York, Cambridge University Press.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing*, 17, 123-139.

Klimova, F. B. (2011). Evaluating writing in English as a second language. *Procedia-Social and Behavioral Sciences*, 28, 390-394.

Knoch, U. (2011). Rating Scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16, 81-96.

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

Lumley, T. (2006). Assessing Second Language Writing. The Rater's Perspective. Frankfurt: Peter Lang.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis*. Sage: Thousand Oaks.

Oppenhaim, A.N. (1992). Questionnaire design, interviewing and attitude measurement. New York: Pinter Publishers.

Park, T. (2004). *Scoring procedures for assessing writing*. Retrieved December, 2004, from http://www.tc.columbia.edu/academic/tesol/webjournal/park_Forum.pdf

Patton, M. Q. (1990). *Qualitative evaluation and research methods*. UK: SAGE Publications Ltd.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly*, 17, 651-666.

Sahin, A. E. (2007). İlköğretim bölümü mezunlarının başarılarının mezun oldukları lise türlerine göre karşılaştırılması. *Egitim Araştirmalari-Eurasian Journal of Educational Research*, 29, 113-128.

Seferoğlu, S. S. (2007). Information technologies in teacher education: Teacher candidates' perceived computer self-efficacy. *Proceedings of the 6th WSEAS International Conference on e-Activities*, pp. 374-378. Puerto De La Cruz, Tenerife, Spain.

Seferoğlu, S. S. (2010). Killing two birds with one stone: Establishing professional communication among teachers. *Procedia - Social and Behavioral Sciences*, 9, 547-554.

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

Wharton, S. (2003). Defining a appropriate criteria for the assessment of master's level TESOL assignment. *Assessment and Evaluation in Higher Education*, 28,649-663.

White, E. M. (1994). *Teaching and assessing writing*. San Fransisco: Jossey-Bass Publishers.

William, D. J. (1996). *Assessing writing. Preparing to teaching writing*. Belmond: Lawrence Erlbaum Associates.

## Genişletilmiş Özet

Yabancı dilde yazma becerisi bir dil öğrenmenin en önemli bileşenlerinden biridir. Önemli olmasının nedeni yabancı dilde kurulan tek bir cümle bile o dilin ne kadar iyi öğrenildiğini ve kullanılabildiğini gösterebiliyor olmasıdır. Yazma becerisi üretken bir beceri türü olduğundan, bu becerinin kazanılması, kanıksanması ve kullanılması daha derin ve karmaşık bir aşama içinde gerçekleşebilmektedir. Yabancı dilde bireyin kendini, düşüncelerini ve duygularını düzgün ve anlaşılabilir bir şekilde ifade edebilmesi son derece önem kazanmaktadır. Bu yüzdendir ki, etkin bir şekilde yazabilme, ifadelerin anlam kazanması ve algılanması açısından daha fazla önem arz etmektedir. Bu da beraberinde var olan ve/ya sonradan kazanılan yazma becerisini geçerli ve güvenilir bir biçimde, doğru ölçüm araçlarıyla ölçme ve değerlendirme ihtiyacını doğurmaktadır.

Öte yandan, öğrencilerin yazdıkları ifadeleri değerlendirmek oldukça güç ve zordur. Bunun sebepleri ilgili literatürde hem kuramsal hem de uygulamalı olarak farklı yazarlar tarafından ele alınmış ve sıkça tartışılmıştır. Literatür incelendiğinde ortada buluşulan temel konuların başında öğrencilerin yazma becerilerinin net, kararlı ve iyi bir şekilde değerlendirilebilmesinin sadece öğretmenler için değil öğrenciler içinde çok büyük

önem taşıdığı vurgulanmaktadır. Vurgulanan, yazma becerileriyle ilgili alınan birçok önemli kararlar öğrencilerin o dilde yazarak ne kadar iyi ve doğru iletişim kurduklarına bakarak verilir. Verilen bu kararlar onların aldıkları ve ilerleyen zamanlarda alacakları eğitimlerini hatta kendi yaşamlarını dolaylı veya doğrudan olarak etkiler. Öğrencinin çevreyle olan etkileşimini algılaması, anlatabilmesi, ifade edebilmesi ve beceri olarak yazıya dökebilmesi birbirini takip eden alt süreçlerin bir araya gelmesiyle olan aşamalar bütününü göstermektedir. Durum bu şekilde gerçekleşince, öğrencilerin sahip oldukları veya kazandıkları yazma becerilerini geçerli, güvenilir ve adil bir yolla değerlendirebilmek kolay olmamaktadır. Ayrıca, değerlendirme başarısı son derece öznel olan bir şeyi aynı derecede tarafsız ve nesnel bir şekilde ölçmeye de bağlı olarak gerçekleşmektedir.

Bir öğrencinin yazma becerisini test edebilme ve değerlendirebilme; beraberinde soru oluşturma, sınavı alan bireyler, değerlendirmeyi yapan bireyler, değerlendirme araçları ve değerlendirme ölçütleri gibi birçok değişkeni barındırır. Bu değişkenlerin hepsi aynı anda yapılacak olan değerlendirmelerde göz önünde bulundurulmalıdır. Değişkenlerden biri eksik veya yanlış değerlendirildiğinde diğer değişkenin değerlendirilmesi de yanlı olabilmektedir. Ayrıca, değişkenlerin birbirleriyle olan ilişkileri, bağımlılık dereceleri de yapılan değerlendirmeleri etkilemektedir. Öğrencilerin sahip oldukları farklı derecelerdeki yazma becerisini değerlendirme, değerlendirenler açısından öznel bir yaklaşım doğurabileceğinden, değerlendirmeyi tarafsız, güvenilir ve geçerli bir şekilde yapabilmek için geliştirilen ölçütlerin önemi çok büyüktür. Degerlendirme ölçütler, her bağlamda benzer sonuçlar vermeyebilir. Farklı kültürlerde, ülkelerde hatta eğitim kurumlarında elde edilen sonuçlar değişiklik gösterebilir. Böylesi bir durum çeşitliliği ortaya çıkardığından karşılaştırmalı araştırmaların yapılmasını da mümkün kılabilmektedir. Aynı zamanda, farklı zamanlarda yapılan araştırmalarda aynı ölçütleri kullanıyor olmak yapılacak olan değerlendirmenin her zaman aynı ve tarafsız bir şekilde yapıldığının göstergesi veya garantisi olmayabilir. Diğer yandan, yazma becerileri ile ilgili değerlendirmeyi yapan kişinin kendi fikir ve düşüncesini de değerlendirmeye katması yapılan değerlendirmenin yansızlığını bozabilecek bir seviyeye taşıyabilir. Bu yüzden, yazma becerilerinin değerlendirilmesinde ölçütlerin değerlendirme yapan kişiler tarafından nasıl algılandığı da ayrı bir önem taşımaktadır.

Yukarıda ele alınan tartışmalar doğrultusunda, bu makalenin esas ilgilendiği konu, Türkiye'de faaliyet gösteren ve özel bir üniversitede bulunan Fakülte Akademik İngilizce Geliştirme Birimi'ndeki ENG 101 dersi için kullanılan yazma sınavı değerlendirme ölçütlerinin, ENG 101 dersini veren öğretim elemanları tarafından nasıl ve ne derecede algılandığının tespit edilmesidir. Bu, aynı zamanda çalışmanın kendi araştırma sorusudur. Araştırma sorusunun cevaplanabilmesi için yapılan çalışmaya ilgili birimdeki 55 öğretim elemanı katılmıştır. Bu 55 öğretim elemanına anketler dağıtılarak cevaplanması istenmiştir. Veriler, ankette bulunan sorulara ilişkin olarak hem nitel hem de nicel olarak elde edilmiştir. Ankette yöntemlerin bir arada kullanılabileceği düşüncesinden yola çıkılarak farklı bir yola başvurulmuştur. Tasarlanan ankette hem Likert ölçeğinin kullanıldığı kapalı uçlu sorular, hem de açık uçlu sorular bulunmaktadır. Bu iki tür şekilde soruların kullanılmasıyla birlikte kapalı uçlu sorulardan nicel veriler elde edilmiştir. Nitel veriler ise anketteki açık uçlu sorularla ilgili birim içindeki gönüllü olarak çalışmaya katılan öğretim elemanları ile yapılan yapılandırılmış mülakatlardan elde edilmiştir. Bu nitel veriler daha sonra sıklıklarına göre ayrı olarak da değerlendirilmiştir. Anket çalışmasına katılan ve ilgili bölümde bulunan öğretim elemanlarının verdiği geri bildirimler göz önüne alınarak ölçütlerin genel etkinliği; ölçütlerdeki kategoriler, puan aralıkları, tanımlamalar, ENG 101 ders hedefleri ve ölçütlerdeki tanımlamalar arasındaki uyum başlıkları altında değerlendirilmiştir. Verilerin analizi için bir istatistiksel paket program olan SPSS (Statistical Package for Social Sciences) kullanılmıştır.

Çalışmanın bulgular kısmında yer alan ve sonuçları tartışılan analizlerin ilk aşamasında ölçütlerin güvenilirliği ilgili istatistiksel teknik olan güvenilirlik analizi ile sınanmıştır. İkinci aşamada ise yapılan güvenilirlik analizi sonucu hesaplanan güvenilirlik düzeyi görece yüksek kabul edilen ölçütlerle ilgili olarak ilgili diğer istatistiksel analizler yapılmıştır. Elde edilen sonuçlar, öğretim elemanlarının çoğunun genel olarak ölçütlerden memnun olduğunu ve var olan ölçütlerin Fakülte Akademik İngilizce Geliştirme Birimi programı içinde standart bir değerlendirmeyi sağladığının düşünüldüğünü göstermiştir. Ancak, anketin bütününe bakıldığında karşılaşılan problemlerden biri, değerlendirmede kullanılan kategorilerin eşit olarak puanlandırılmaması ve ölçütlerin bütün öğretim görevlileri tarafından aynı şekilde kullanılmadığı yönündeki güvensizliktir. Bu güvensizlik mülakatlara bağlı olarak incelenmiş ve güvensizliğin nedeninin program genelinde ölçütleri kullanmada farklı öğretim elemanlarının değişik yollar ve tutumlar izlediği olduğu kanısına varılmıştır. Ortaya çıkan başka bir bulguda ise, yapılan tanımlamalarda kullanılan kelimelerin birimde bulunan öğretim elemanlarınca farklı düzeylerde algılanabildiğidir.

Çalışmada, elde edilen nitel verilerin değerlendirilmesi için yapılan yapılandırılmış mülakatlarda, bazı öğretim elemanları kendi bölümlerinde yazma derecelerine değerlendirmede kullanılan bazı kategorilerin ve puan aralıkları aynı ağırlığa sahip olmadığını vurgulamışlar, bunun da değerlendirmeyi bir ölçüde anlamlı derecede farklılaştırdığını açıkça dile getirmişlerdir. Çalışmanın sonunda elde edilen bulgularla ilgili önerilerde

bulunulmuştur. Bunlardan birincisi, öğretim elemanlarının yorum ve değerlendirmeleri göz önünde bulundurularak ölçütlerdeki kategorilerin ağırlığının yeniden gözden geçirilmesi ve düzgün bir şekilde ayarlanmasıdır. İkincisiyse, problemli ya da farklı yorumlanabilecek tanımlamaların yeniden yazılmalarıdır. Böylelikle var olan veya yeni yapılacak olan tanımlamaların öznel yorumlamalara kapalı ve özgül olmaları öngörülmektedir. Üçüncü öneri, kullanılan ölçütlerin etkin bir biçimde kullanılabilmesi için ölçütlerin nasıl kullanılması gerektiğinin gösterileceği bir eğitimin verilmesidir. Bu sayede hem birim hem de bölüm içerisinde ölçütleri uygulamaya yönelik ortak bir anlayış sağlanması mümkün olacaktır. Dördüncüsü ise, örnekleme-anakitle ilişkisinin diğer bir deyişle standardizasyonun sağlanmasıdır. Örneğin, bir kaç öğrenci kâğıdının ortak olarak ölçütler ışığında küçük gruplarca değerlendirilip, yani onları bir örneklem olarak kabul edip, bütün öğretim görevlilerince değerlendirme hakkında fikir birliğine varıldıktan sonra geriye kalan diğer kâğıtların incelemeye alınması ve değerlendirilmeye başlanması önerilmekte, bu şekilde objektif değerlendirmenin sağlanabileceğine inanılmaktadır.

---

## Citation Information:

Tarkan-Yeloğlu, Y. Seferoğlu, G., & Yeloğlu, H. O. (2013). A case study on instructors' perceptions of writing exam grading criteria. *Hacettepe Üniversitesi Eğitim Fakültesi Dergisi [Hacettepe University Journal of Education], 28*(1), 369-381.