T.C. KOCAELİ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI
İNGİLİZ DİLİ EĞİTİMİ BİLİM DALI

THE EFFECTS OF WRITING WORKSHOPS ON FEEDBACK PROCESS OF UNIVERSITY INSTRUCTORS

YÜKSEK LİSANS TEZİ

Pınar Ayşe MÜFTÜOĞLU

KOCAELİ 2020

T.C. KOCAELİ ÜNİVERSİTESİ
SOSYAL BİLİMLER ENSTİTÜSÜ
YABANCI DİLLER EĞİTİMİ ANA BİLİM DALI
İNGİLİZ DİLİ EĞİTİMİ BİLİM DALI

THE EFFECTS OF WRITING WORKSHOPS ON FEEDBACK PROCESS OF UNIVERSITY INSTRUCTORS

YÜKSEK LİSANS TEZİ

Pınar Ayşe MÜFTÜOĞLU

Doç. Dr. Doğan YÜKSEL

Tezin Kabul Edildiği Enstitü Yönetim Kurulu Karar ve No: 24062020/14

KOCAELİ 2020

ACKNOWLEDGEMENTS

This thesis becomes a reality with the kind support and help of many individuals. I would like to extend my sincere thanks to all of them. Most importantly, I would like to express my deep gratitude to my supervisor Assoc. Prof. Doğan YÜKSEL for giving me the opportunity to do the research and providing invaluable guidance throughout this research. His vision, knowledge and motivation have deeply inspired me. It was an honour to study under his guidance.

I would love to thank the school and the instructors who enthusiastically supported my thesis and took part in this research. They shared their thoughts, gave me wonderful ideas and my tremendous friends have been supportive and so helpful throughout data collecting process. I am also grateful to Assoc. Prof. Banu İNAN-KARAGÜL, Assoc. Prof. Sibel KAYA, Assoc. Prof. Mehmet Ziya FIRAT for their suggestions, assistance and precious time.

I would especially like to express my love and thanks to my dear family whose love and guidance are with me. Without their support, I do not think that I could overcome the difficulties during this process. Especially, I wish to thank my loving and supportive fiancé Murat ASLAN who was patiently and continuously by my side. He made this process achievable and statistics became comprehensible thanks to his knowledge.

TABLE OF CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

3. METHODOLOGY

CHAPTER 4

4. RESULTS

ÖZET

Açık uçlu sınavların değerlendirilmesi oldukça özneldir. Bu sebeple öğretim görevlileri belirlenmiş kılavuzlardan yararlanmalı ve bunların kullanımı hakkında uygulamalı eğitim almalıdırlar. Rubric kullanım esnasında oluşacak farklı varsayımlar ve her sınav kağıdının kendine özgü olması sebebiyle bazı sorunlar ortaya çıkabilir. Rubric kullanımı, uygulamalı eğitimler sayesinde daha verimli ve standart hale gelir. Bu eğitimler özellikle çok fazla öğrencinin bir arada olduğu hazırlık okullarında belirli standartlara ulaşılması açısından çok önem taşımaktadır.

Bu çalışma açık uçlu sınav soruları okuma üzerine yapılan pratiklerin etkinliğini araştırmak üzere İstanbul'da bir özel üniversitede yapılmıştır. Bu araştırmada Hazırlık okulunda görev yapan 28 öğretim görevlisi ve 421 öğrenciden alınan 842 kâğıt yer almıştır. Veri nicel ve nitel veri toplama yöntemleriyle toplanmıştır. Nitel veriyi uygulama yöneticileriyle yapılan standartlaşma sonrası görüşmeler, nicel veriyi ise uygulamalı eğitim öncesi ve sonrası yapılan sınavları okuyan iki öğretim görevlisi arasındaki puan farkı değişimi sağlamıştır. Önceden okunan ve puanları belirlenmiş kağıtlara öğretim görevlilerinin verdiği notlar da nicel veri olarak toplanmıştır.

İlk olarak workshop öncesi iki öğretim görevlisinin aynı sınav kağıdına verdiği farklı notlar arasındaki farklara ve daha sonra bu farklara uygulamadan sonra tekrar bakılmıştır. Son olarak da uygulamaları düzenleyen üyelerle görüşmeler yapılmıştır. Sonuçlar gösteriyor ki bu uygulama öğretim görevlilerinin açık uçlu sınav okumalarında oldukça etkilidir. Öğretim görevlileri rubric kullanımında daha etkin hale gelmiştir ve iki okuma arasında gözlenen fark azalmıştır. Bu çalışmanın sonuçları rubric kullanımında uygulamalı eğitim yapılmasının önemini vurgulamaktadır.

Anahtar Kelimeler: sınav okuma pratiği, açık uçlu sınav değerlendirme, standartlaşma

ABSTRACT

The evaluation of written exams is highly subjective. Therefore, instructors must be equipped with the set of guidelines and the criteria in the form of rubrics. In addition, instructors should be trained in the usage of rubrics. Some problems might occur such as instructor's inference of each scale on the rubric, utilization of rubrics effectively and each exam paper's uniqueness. It is obvious that rater trainings enable instructors to use rubrics effectively and in a standard way. Especially, writing standardizing workshops is crucial to be fair to all students in prep schools which have a lot of students.

This study was conducted to analyze the effectiveness of workshops on rating process of instructors at a private university in İstanbul. Twenty-eight prep school instructors and 842 writing exam papers of 421 students were involved in this study. The data was collected via quantitative and qualitative data collection instruments. Interviews with workshop conductors after the standardization sessions provided the qualitative data; exam results of the students, ratings of instructors and benchmark analysis supplied the quantitative data. Moreover, workshop conductors were interviewed after the sessions.

The results indicate that workshops are highly effective on the feedback process. Instructors started to use rubrics in a more efficient way and discrepancy between two papers graded by instructors lowered. This study highlights the importance of standardizing workshops on feedback process of university instructors.

KeyWords: rater training, writing workshops, standardization

DEFINITIONS AND ABBREVIATIONS

Benchmark: A benchmark is something whose quality or quantity is known and which can therefore be used as a standard with which other things can be compared. https://dictionary.reverso.net/english-cobuild/benchmark+paper.

Inter-rater Reliability / Rater Agreement: Inter-rater reliability is the level of agreement between raters or judges.
https://www.statisticshowto.com/inter-rater-reliability/

Rater Training: Eliminating rater differences McNamara (1996) (p. 232).

Standardization Sessions: In which teachers rate sample scripts of oral or written performances of students and compare their marks.
https://doi.org/107.1007/978-3-319-77177-9 (p.139).

Z-score: Z-score or standard score, is used for standardizing scores on the same scale by dividing a score's deviation by the standard deviation in a data set. The result is a standard score. It measures the number of standard deviations that a given data point is from the mean.
https://corporatefinanceinstitute.com/resources/excel/functions/z-score-standardize-function/

CEFR: Common European Framework of References for Languages

ELT: English Language Teaching

ICC: Intraclass Correlation Coefficient

MA: Master of Arts

SPSS: Statistical Package for Social Sciences

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

1. INTRODUCTION

1.1. PRESENTATION

In this chapter, brief information about the background of the study is given briefly and the purpose and the importance of the study are explained. The problem is presented, and research questions are listed.

1.2. BACKGROUND OF THE STUDY

Rubrics have been debated for a long time on the pages of many journals. They were all defined differently as sequencing machines or necessary guiding tools. Moreover, it has been mentioned that standardization occurs when rubrics are used, rubrics enable one to be fairer, they make the feedback process of teachers more meaningful as a result, and this helps student development. There is no doubt that students' skills improve by writing and getting proper feedback from teachers. The work that students produce can be compared with others and its rank can be classified so that the quality of the work can be defined. It is possible to name students work numerically and define it with numbers. Rubrics help teachers to increase the level of fairness in the assessment process Turley and Gallagher (2008) (p.88-89). It is obvious that rubrics have been used as guiding tools and educators and learners have benefited from them. Andrade (2000) also indicates the importance of rubrics by mentioning that rubrics represent scales to classify the product from excellent to poor. This classification assists the language learning process, but this is only possible if this assessment process is reliable. Because improving the weak parts of the students or school systems is only possible by defining problematic parts well. Moreover, people count on large scale exams and formative or summative assessments give information about their language proficiency or progress. Fairness becomes a matter when all these types of exams are considered. People, educators, and students all look for more reliable methods to evaluate the work that students produce. Standardization becomes more important since it ensures fairness with the help of guidelines and a set of rules. Most of the schools and institutions use rubrics to set their standards in writing exams and criteria help raters be more objective. (p.13)

Each band of the rubric is defined to inform the rater how to assess the student's performance. But at that point, the rater's interpretations create differences in the scoring process. Every rater has some priorities in their mind, for that reason even if they refer to the rubric, scoring results might differ. The issues arise at institutions, where a lot of raters take a role in the assessment process. This becomes an important problem and it should be dealt with to make the assessment process more reliable. Standardization sessions or training is done to make the raters familiar with the rubrics. So, this study was conducted at a school where more than 400 students study specifically when rubric utilization became inevitable. Because of the accreditation process, instructors were asked to do the scorings with the rubrics. Instructors were newly introduced with rubrics to standardize rating procedures. Analytic rubrics were used in this study since they have more detailed descriptors on each scale and raters can analyze writing better. Those rubrics provide more input about the quality of the writing papers Knoch (2009) (p.276). Although rubrics supply guidelines to instructors during the assessment process, there are still some concerns about the utilization of rubrics. Some teachers have different beliefs; they give priority to different language aspects while grading the writing exams if they do not use rubrics. Some problems such as quality of writing, language usage, content knowledge, grammar and mechanics might be graded in different ways and one might be regarded as superior to another. As a result, different raters come up with different results and reasons when they are evaluating papers. As Yeloğlu, (2013); Rezaei and Lovorn, (2010) mention in their studies that instructors need longer training to be able to utilize rubrics properly before the assessment process. During training, instructors will be able to comprehend each band of the rubrics more efficiently by having discussions about concerns and they will practice the utilization of rubrics until they get standardized. The decrease in the discrepancy point differences between two raters before and after the workshops highlights the importance of standardization workshops**.** (p.378;18).

1.3. PURPOSE OF THE STUDY

This study has been designed to find out how effective writing standardization workshops are on the utilization of rubrics**.** The aim of this study is to provide data that would help us to realize how effective the workshops are and the data about the challenges faced throughout the standardization sessions. Therefore, the study takes place in a school where a writing evaluation rubric is currently being introduced, and exam papers are expected to be checked twice by different teachers with the rubrics. To make teachers familiar with the new phenomena in the school, writing papers evaluation workshops were done. To evaluate the effectiveness pre and post-tests were done. The interrater reliability rate of the instructors was noted before and after the sessions. This data was used to present the difference before and after workshops (standardization sessions). Next, benchmark analysis was done to check the current situation of the instructors in the last workshop and how close the instructors were in scores to each other and the benchmarks. Whether instructors had the same performance in the last workshop and the final exam was checked by comparing the inter-rater reliability of the raters. Interviews with committee members were done to present the difficulties faced throughout this new process and their beliefs about the effectiveness of the workshops.

This study will investigate the following research questions:

1. What were the ICC and individual scores (z-scores) of the raters in the last workshop? Is there a difference in ICC scores of the raters between the last workshop and post-test (final exam)?

2. What is the difference in ICC scores of the raters before and after writing workshops?

3. Do workshop conductors believe that workshops were effective? What were the difficulties workshop conductors faced during the process?

1.4. IMPORTANCE OF THE STUDY

Many other studies have focused on the effectiveness of the rubrics and it has been highlighted in those studies that rater training is necessary since raters have different interpretations while they are grading scripts with the same rubric. So, training is an inevitable requirement for raters. Many studies were done to check the effectiveness of rater training. They either focused on inter-rater reliability, individual leniency and harshness of raters. Inter-rater reliability analysis was done mainly by the Intraclass Correlation Coefficient (ICC). In some other studies, interviews are done to present the effectiveness of the training process. In this study, ICC rates in pre and post-tests were compared. 28 instructors scored 421 scripts before and after the writing workshops and they worked on the scripts for about five hours during the workshops. Moreover, in the same study raters' performance in the last workshop was evaluated and ICC scores from them were presented. Interviews were done to gain more information about the training sessions such as questions of the raters, parts they struggled with and solutions found by the committee. An example of a school that underwent new regulations was presented in this study. This study shows why an institution needs writing standardization workshops, how they conduct them, the problems which occur during the process and how they deal with them. As a result, the effectiveness of the workshops was checked by comparing the inter-rater reliability results.

1.5. LIMITATIONS OF THE STUDY

There are some limitations in this study. First, one school was chosen as a research area. The results cannot be generalized to other institutions or instructors but can definitely provide some insights on this topic. Next, the study was done at the beginning of the newly introduced process. After a few weeks, the results were gathered and compared. If a longitudinal study could be done to observe this process, better results might be gained. Because of the workload of the instructors, workshops were limited in numbers. Raters’ opinions are highly important in such kind of studies but due to the institutional constraints, interviews could not be carried out with the instructors. Workshops could not be recorded since the school authority underwent the accreditation process at the time of the data collection.

CHAPTER 2

2. LITERATURE REVIEW

2.1. PRESENTATION

In this chapter, the literature about assessment and evaluation, writing assessment and difficulties related to scoring constructed response items, scales types, reliability of scoring (inter-rater, intra-rater reliability), rubrics and standardization training is explained briefly. The studies conducted regarding rubrics and rater training are presented.

2.2. ASSESSMENT AND EVALUATION

Assessment and evaluation processes might be regarded as the same process, they are combined to enable stakeholders to make some assumptions related to the learning process, but they refer to different terms in a consecutive process.

Assessment is a process to analyze a student's performance within the frame of a previously set of goals or objectives. Teachers, learners and school administrators obtain information about learners' performance through assessment. It gives them an awareness of what students can do and whether the main objectives of a curriculum are achieved or not Stoynoff and Chapelle (2005). It means that each assessment is designed with some pre-determined objectives so, the assessments are done to check whether the objectives actualize or not.

According to McAlpine (2002) assessment is a mutual communication for learners, teachers, administrators, and curriculum developers which means they have the chance of evaluating teaching, learning, resources and the curriculum. In this mutual communication, there are two sides: the educational process and learning outcomes. The educational process is the procedure carried by educators to achieve their goals and learning outcomes are the yields of that process or can be defined as the results of the assessment. (p.4).

There are two different types of assessment: formative and summative assessment. The former aims to give feedback during the learning process, and it has the aim of improving learner's current knowledge. The latter aims to give an overall idea about the progression of a learner at the end of the course. McMillan (2007) describes summative assessment as a search for information after completing a term or a course and investigating its effectiveness by assessing learners. On the other hand, formative assessment is described as a guideline to help the educational process of learners. In other words, summative assessment gives information about the final point or learning outputs while formative assessment supplies information about the ongoing process, moreover, its goal is improving students' current level during the educational process.

Although evaluation and assessment are in correlation, defining evaluation might be regarded as complex since assessment is sometimes used alterably for evaluation. Evaluation is having a decision about learners' performance, having some data about learners' strengths and weaknesses. According to Weiss (1972) evaluation is gathering data with the aim of making a judgment or decision. After learning outcomes are assessed, the results are evaluated to reach a final decision.

Whenever it is crucial to acquire a language, assessment of skills becomes important. If the matter is a foreign language and assessment of skills, writing is not an exception Weigle (2007). (p.194).

2.3. ASSESSING WRITING

Since writing assessment is an important part of learning, it has been a concern for English Language Studies. Students learn how to write in different contexts so, being able to give them feedback for their improvement is very crucial. Therefore, providing them feedback in various contexts should be a matter to find new methods to assess writing Huot (2002). (p.61-79).

When writing assessment becomes a matter, writing ability comes to mind as a second term since the former assesses the later. So, there is no doubt that the best way of assessing writing is by directly making learners write Hughes (1989). In other words, everything produced by learners is commented on or scored, the written data is collected and evaluated. (p.75).

Gallagher (2009) describes writing assessment as gaining information about learners' writing proficiency level for summative and formative assessment requirements. Several questions should be dealt with when assessing writing such as the purpose of assessing writing, learner profile, criteria or standards, who needs the results, reliability and validity and the limits of the assessment. (p.34). According to Bachman and Palmer (1996) making assumptions about the language ability and consideration of the results of these assumptions are essential purposes of language assessment. The deductions gathered as data might be used for some different purposes such as identifying the program level or effectiveness, individual success, or a classroom level. (p.61).

Reliability is a concern for every assessment. Concerns about writing assessment even date back to 1912 Starch and Elliott (1912). These concerns are related to the reliability of the test results and the assessment process. There are two different kinds of answer options for writing assessment: selected response items and constructed response items Popham (2003). The former is choosing the correct answer from the given options while the latter is providing data without depending on any cues which refer to a productive skill, directly writing the answer. Constructed response items might be the answer to a short question or an essay type question. Learners get the best chance of showing their writing skills ability by responding to essay types of questions. However, it has two disadvantages; essay types of questions require time consuming scoring and this scoring might have some inaccuracies. (p.72-86).

All in all, the evaluation of constructed response items requires interpretation and recognition, it is highly subjective. Raters in the writing assessment, especially in large scales, could be different as a result, it causes rater reliability problems.

2.4. WHY IS IT DIFFICULT TO ASSESS WRITING IN LARGE SCALES?

Assessment, even independently can be regarded as difficult because determining the goals to assess, choosing the methods to conduct tests and giving feedback already creates some challenges for the teachers and administrators. Apart from that, assessing writing comprises more challenges. Since writing is a productive skill, it is mostly assessed through essay type questions. Especially, essay type questions can be used as a formative and summative assessment at schools or institutions. But the problem occurs at that point and more questions arise. Is it possible for the rater to be fair? Is it always possible to be consistent? What type of guidelines or criteria do the raters use? Is it possible to get correlative results in a similar kind of exam? All of these questions create a matter of reliability. When this assessment is considered in large scales, the problem increases because, at that point rater training, rubrics, scales and their usage are questioned.

Reliability and validity are inseparable components of writing assessment. Both should be conducted in every test, but the occurrence of this requirement is not enough for test results to be conducted and it does not provide us trustable results. It is necessary to have a reliable rating process to complete an assessment and evaluation process. Organizing a reliable rating process is possible by using certain sets of rules, guidelines, scales, rubrics and rater training.

Although writing assessment is necessary for teaching, it might be problematic and complex. Since writing assessment is subjective, teachers might have difficulty in evaluating exam papers in terms of assessment criteria or sets of rules. In other words, there are some obstacles throughout the evaluation of exam papers such as what to take into consideration, what to ignore, how to deal with minor mistakes and teachers' beliefs, personal differences. Especially at schools where more than 20 teachers work, it is necessary to have some set of rules to evaluate exam papers Yurekli & Ustunluoglu (2007).

Huot (2002) emphasizes on interrater reliability, no matter what the writing assessment type is, the focus should be on rater training, guidelines and techniques to certain interrater reliability. In other words, whatever the assessment type is, no matter how reliable and valid it is, if it is not rated in a proper way the results will not be valid or trustworthy. As a result, the usage of rubrics, scales and rater training is incredibly important. (p.61-79).

2.5. HOW TO SCORE WRITING ASSESSMENTS?

As Stratman & Hamp. Lyons (2003) emphasizes that since direct writing assessment is a performance-based test it includes different elements such as topic, examinee, examiner… As a result, in such assessments, the score depends on the assessment criteria. Direct writing assessments are subjective which means test developers are not able to prepare one answer key that has an answer for all the other exam papers which are all unique in terms of context, style and language. Popham (2003) defines the important drawback of constructed test items, essays, as the probability of inaccuracies in the rating process but also, he highlights that it is certain to overcome that problem if a proper rubric is used. So, the usage of some previous set of rules is inevitable to be fair in testing.

First rubrics were only designed to help scoring. These are the first examples of the rubrics which helped graders to give points (grammar, mechanics, content, language usage) but they were not really expressing information about each criterion and how they should be considered and evaluated. Great rubrics can reflect how much the learners know, and it also reflects their insights. In time, rubrics are improved, and teachers start to evaluate the learning process according to rubrics. They start to give detailed information about each criterion and help teacher's cognition of writing assessment Popham (1997).

Rubrics are regarded as guidelines or tools which assist raters to define the quality of the written product. Rubrics have some bands on them, each of which helps raters while scoring papers. Related to the quality of the performance, each band has numbers on the rubric. Higher points on the rubric refer to a better performance Schafer (2004).

Rubrics are defined by Tierney & Simon (2004) as a helpful device to lower the subjectivity level in the assessment process. Gaining higher standards of objectivity is the aim of the utilization of rubrics. Markbook (2011) defines rubrics as a guide to putting the student's performance into a grade. It evaluates the paper in terms of grammar, content, structure and mechanics. While evaluating an exam paper rubric make the written work comprehensible and it helps to turn it into a grade. Brualdi also (1998) defines rubrics as frameworks that help to show the level capability of learners. Rubrics are prepared according to the knowledge and experience of the teachers. Teachers could have assumptions about learners' knowledge within that framework. In other words, rubrics are guidelines for teachers during the assessment process. (p.2). They ascertain the criteria which will assist teachers. Rubrics help teachers to evaluate their learners within a framework which will also help their objectivity. With the help of rubrics, teachers will be fairer since they have some set of rules and criteria to guide them. A. Qasim & Z. Qasim (2015) also suggest that rubrics reduce the amount of subjectivity while reading writing exams and it neutralizes this process. Moreover, they mention that it could also provide feedback on the improvement of the students.

Learners should be able to gain explicit information about their writings. They should know what they can do well or not. Each language aspect should be clearly explained to them. They should be informed about their assessment process. It is known that assessing writing is highly subjective and it gives all the hard work to the teachers. Rubrics are available to clear away these concerns and give reasons behind the assessment Broad (2000).

Turley & Gallagher (2008) argue that there is no certain type of criteria that can be regarded as perfect or suitable for every different type of writing assessment. The type of criteria used should be decided according to the purpose of the program, place, content and focus. Their effectiveness might change according to the aim they serve.

Moreover, Rubrics are also useful to improve self- assessment skills of the learners. As Broad (2000) suggests rubrics also help self-evaluation. Rubrics should give an idea to learners about how they are assessed and what their work should include. Students can know what they are expected to do and with the feedback, they receive from teachers and students monitor their learning process as they improve their writing skills. Andrea (2007) also put forwards the idea that if the rubrics are used with the right input, they assist the self-evaluation process. It helps to arouse self- awareness. Moreover, rubrics play a key role in the completion of the tasks. Students will have a chance to criticize their work so that they will be able to complete the task. Kutlu (2010) analyzed instructors using rubrics for their assessments. He realized that instructors with a positive attitude towards assessment rubrics take advantage of the rubrics and in time they redesign their rubrics according to learners' needs and they also involve learners in this process. Before they set the assignment, they first introduce rubrics to their learners so that they will be able to know what they are supposed to do as a result it improves learners' writing skills.

Weigle (2002) mentions that some steps should be taken into consideration while designing scoring criteria. The first question should be related to the type when designing criteria, the designer should choose a primary trait, multi-trait rating, holistic or analytic scale. The second concern of the designer will be the user of the criteria. Who is going to use it? Next, there will be some different aspects of criteria to measure the writing and according to program objectives these aspects might differ, but they should be in balance. For some programs, one aspect of the written product might be superior compared to others so they should be decided in advance by the criteria developer. Third, parallel to different performance expectations from the students, there should be some lines to score learners' performance and the number of these lines should be between six and nine. It will help assessors to make inferences about performance such as poor, good, needs improvement… These lines or the bands should be clearly distinguished from the other as the assessors are supposed to rate the performance properly. (p.108).

Finally, the developer needs to decide how the grades will be reported. For example, if it is necessary to give information about each writing aspect on the scale, then analytic scales will be appropriate to use otherwise, holistic rubrics might be used to report overall scores.

2.6. WRITING ASSESSMENT SCALE TYPES

According to Weigle (2002), There are four different types of criteria described to assess writing ability. Each of them has a different purpose of usage and place to be conducted. First, the appropriate type of criteria should be chosen to assess learners' skills properly. The selection of proper criteria is also crucial to get valid results, moreover, it will be a concern of the reliability of the test. (p.108).

2.6.1. Primary Trait and Multi Trait Rating Scales

Primary trait and multi trait rating scales are specifically designed for each task. What is expected from each task and assessment of which language aspect is prioritized in every assessment might differ. One single criterion will not be suitable for every other task. It means that for every other test criteria developer needs to prepare specific criteria to assess the product of students which is neither time efficient nor user friendly.

2.6.2. Holistic Scoring

Holistic scales do not focus on each writing aspect of the product, but they mostly give information about the learner's proficiency level in language. So, the raters are not supposed to deal with content, grammar, language issues separately which means they are supposed to choose the proper score band for each written product. There are some expectations in each score band for learners' products but still, the raters do not score each aspect of writing Klimavo (2011) For that reason holistic rubrics are mostly used in large scales to identify proficiency levels of learners. They are mostly time and cost efficient which is significant when the assessment is done for many learners for written international exams Bacha (2001). (p.375).

Finson (1998) states that if the target is focusing on product, holistic rubrics are preferable since they give information about the general proficiency level and it is the final assessment to get an overall view about the student, but if the aim is process, then analytic rubrics should be used because they supply information about language proficiency in detail so in the improvement process of the language they will be more beneficial.

Martin and Kniep (2000) highlight the fact that although the preparation of holistic scales needs more effort since bands are designed as general criteria to ascertain the proficiency level of the learner, they are time efficient in practice. In other words, when it is attempted to be analyzed there is not much to say about the features of the writing, but it is still not less reliable since it can determine the proficiency level and it is user friendly and time saving in usage. (p.35).

The study conducted by Becker (2010/2011) focuses on the selection of rubrics. The study tries to define the better rubric among holistic, analytic and primary trait rubrics. 82 instructors were involved from the US and a questionnaire that had 10 questions was designed to collect the data from instructors. Interviews were also done with the instructors. After collecting the data, the researcher realized that none of the instructors were using primary traits rubrics, two of the instructors were not using any rubrics during the scoring process and the rest who were using holistic rubrics doubled the numbers of instructors who were using analytic rubrics. This study also investigated the reason why instructors prefer to use holistic rubrics. The answer was scoring papers by using certain numbers was easier than scoring each feature with numbers. So, it can be understood that instructors find it difficult to score each feature on the rubric differently. Overall impression and defining one number to score papers were recorded as preferable. (p.114).

2.6.3. Analytic Scoring

Analytic scoring enables raters to make decisions about each aspect of writing samples. Analytic rubrics have several futures representing the quality of the written product. It means that the rater can give information about the formation of the written product. So, analytic scoring supplies more information than holistic scoring. Moreover, the reliability of the analytic scale is higher than the holistic ones. Since the language aspects of learners improve differently, analytic scoring is more appropriate to determine each aspect of the written product. Even combined usage of analytic and holistic scoring is possible, the rater can define the level of learners' performance by using holistic scoring and identify items analytically Weigle (2002). (p.109).

On the other hand, Knoch (2009) put forward the idea that interrater reliability is higher in analytic scoring because raters can define each band with an explanation or a reason. This helps raters to give reliable decisions. (p.276). Barkaoui (2010) supports this idea by mentioning that holistic scales do not assist raters while scoring written products in terms of various aspects of language such as language usage, vocabulary depth or organization. Kuisma (1999) supports the usage of analytic scales in writing assessment by expressing that raters can come to an agreement more easily when they use analytic scales since they are able to discuss each feature on the scale and they justify their reasons in a more concrete way. Being able to discuss each aspect and analyzation can make these scales preferable. Barkaoui (2011) supports the idea of Kuisma (1999) by mentioning the same issue, analytic scales help raters to get closer scores than the holistic ones.

Moreover, Crehan (1997) claims that since analytic scales supply more detailed information, they provide more input to students what to study more or which subject to focus on more and it provides the same guidance to teachers to analyze students' strengths and weaknesses. This process turns into a clear feedback benefit for teachers and learners.

Even though some measures are taken to prevent reliability problems, during the writing assessment process some concerns still stay there to be dealt with. Some problems can be eliminated throughout the rating process by the usage of some set of guidelines, tools and scales. They help to determine or diagnose the proficiency level of a student or reveal the student needs but Lumley (2002) states that raters follow up criteria step by step, they know what each band refers to but still, there are some problems that cannot be covered during the writing assessment process which means raters need training and to spend more time on the usage of criteria. So, this process will be more efficient. (p.246).

2.6.4. Image-Based Scoring

As it happened in many different areas, technology has affected scoring processes, especially on large scales. In the late 1990s, electronic scoring was developed. Constructed response items can be scored by using technology. In this case, technology grabs students' responses electronically and in seconds it helps related responses to meet with the scorer proficient in his/her field. With the aim of being more reliable and valid, related responses are transferred to the scorers directly. The machine matches proficiency levels in the field with the responses. Moreover, equally and automatically responses are sent to the second and third checks then the averages are taken by the machine to be announced Way&Vickers&Nichols (2008). (p.3-4).

2.7. RELIABILITY OF SCORING

When writing assessment is considered, another issue comes to mind, fairness. Besides being helpful and assisting learners, writing assessment should be fair to enrich its aim. White (1996) mentions that writing assessment should be fair. Students should get the result that they deserve. A set of rules or guidelines make the writing assessment process fairer. It helps instructors to evaluate writing papers with guidance. Spandel (2006) argues that grading writings might be a matter of objectivity. All the instructors evaluate the writings following the criteria so that teachers could have guidelines to lead them in this process with safety. It would be useful to evaluate writings objectively which is also a matter of fairness.

On reliable examination, learners are expected to get the same results even if there is a change in the scorer, the day which is the exam scored or the person who is scored. These are the differences that should not make differences in the score to make the exam reliable. This is the purpose of the assessment. When the assessment is the matter, two different types of reliability become important: interrater reliability or intrarater reliability Moskal and Leydens (2000).

2.7.1. Inter-rater Reliability

It is most common to hear students' critiques after an exam whose results based on subjective scoring. It means that the exam results might differ according to the rater. If the exam paper is scored by another rater, the results might change. Because each rater has a specific way of scoring and they might prioritize different features of the product. To eliminate this problem some set of rules, guides or methods are used even though they are not able to resolve the problem entirely Moskal and Leydens (2000).

Stemler (2004) defines inter-rater reliability as a feature of assessment but not as a tool. It is a commitment between raters within the frame of some set of rules, rubrics. The final decision on a specific performance at a certain time. So, it is the requirement of the assessment process. For the determination of accuracy and consistency of the rating process, there are three approaches mentioned: consensus estimates, consistency estimates and measurement estimates. (p.1).

2.7.1.1. Consensus Estimates

Consensus agreements are mostly used in studies to show the percentage of the agreement between raters**.** Consensus estimates stand for the interrater reliability that raters give similar scores to the papers. They are closer to each other in scores. This is especially expected when the raters are trained to use rubrics. By using the rubrics sensible raters should reach a compromise. Agreement levels should be not less than %70 Stemler (2004). (p.3). If the agreement rate reaches %70 or above, it is considered as reliable. It should be also kept in mind that the consensus agreement depends heavily on a rubric. If the rubric has fewer parts on it like in holistic rubrics, higher consensus agreement results become more possible to get. If there are more bands, reaching a consensus agreement becomes more difficult. In some studies, Cohen's Kappa is used to figure out an agreement rate that occurs between raters by chance. Kappa's value between .40 and .75 represents fairness more than a chance Stoddart, Abrams, Gasper, & Canaday (2000).

To analyze nominal data Kappa can be used but continuous data can be analyzed with Pearson Correlation, paired sample T-test. These tests analyze reliability or correlation. Intraclass Correlation Coefficient (ICC) both analyzes rater agreement and correlation, so it is the best option to analyze inter-rater reliability. ICC values between 0.75 and 0.9 are regarded as good and values more than 0.9 are excellent. There are different types of ICC forms. The correct ICC form should be chosen for analysis. One-way random, two-way random and two-way mixed models are used. One-way random model is used for large scales. In such cases, one part of the assessment can be rated in one institute while the other part is assessed in another place. This is a very rarely used model. If the raters are chosen randomly from a large population two-way random effects are used. If the raters are set and if the same set is used for ratings, two way- mixed model is used Koo and Li (2016). (p.155-157).

2.7.1.2. Consistency Estimates

The raters may not come to an agreement; they might have different decisions about the performance. While one rater gives 1 point to the performance the other may give 3. Consistency Estimates become clear when the data is continuous. When there is always the same discrepancy between two raters it is called consistency estimates in interrater reliability. For example, two different raters score papers and one of them always gives two points more than the other. This means it gives consistency in scores but then it means that the scores should be checked to see the correctness Barrett (2001). In most of the studies, correlation coefficient is used to estimate consistency but, in many cases, it is not specifically explained which coefficient is used. In some articles, they specify it and they mostly use Pearson's, Spearman's or Kendall's W coefficients. Cronbah's Alpha is also used to report consistency. Jonsson & Svingby (2007). Consistency estimates are acceptable if the value is .70 or more Brown (2004). (p.134)

2.7.1.3. Measurement Estimates

This approach is based on the idea that all the data gathered from the raters should be collected and there must be some reasons and justifications behind the decisions. So, the data is valuable for the assessment process. Raters might prioritize the different parts of the performance but overall, the data is very important to get the average. The collected information is crucial not the scores Linacre (2002).

2.7.2. Intra-Rater Reliability

Intra-rater reliability refers to the consistency in the scores of exam papers of the same rater at different times which means a scorer might have different points of view on different days or the scorer can prioritize different aspects of language more at another time. In other words, a scorer can rate the same performance differently at different times. For a different case, the rater might score the same performance of a strong student better than a weak student because of the background knowledge about them the rater might be misled Moskal and Leydens (2000). Studies that are working on intra-rater reliability focus on Cronbach's Alpha and this helps them to get information about the consistency of an individual marker. If the value is more than .70 Brown, Glasswell, & Harland (2004) accept it as enough. So, intra- rater reliability might be a concern in reliability but when the raters use rubrics, this is not a major concern anymore. (p.105).

2.8. IS IT NECESSARY TO HAVE STANDARDIZATION SESSIONS (WORKSHOPS)?

The data collected from a writing assessment is highly subjective because every paper is unique on its own. Every single paper has its exceptions, differences that cannot be put into any band of a scale. So, scoring those exam papers is regarded as less reliable since they are subjective, and raters cannot come to an agreement about which features of writing to score Turgut (1990). Raters face with some unknown, not anticipated problems on each paper, they might not know how to evaluate that specialty on the paper. Making decisions or evaluating the paper properly becomes difficult. Since the paper on evaluation is different from the other papers and some obscureness comes out as a problem, the rater must deal with and spend time to find the correct band on the scale. Moreover, the rater must be still fair while evaluating that obscureness. (p.65).

The raters develop some strategies to keep up with the scale and they tried to deal with the problematic parts that they cannot name. Unfortunately, even if the raters try to stay closer to the scale, the first impression or complexity they got from a paper affects them a lot. As a result, they are influenced, and they cannot be that much transparent to evaluate the paper. At that point, Lumley (2002) focuses on the point that raters play a more important role than the scales do since they are at the center of the writing assessment process. In other words, raters decide at critical points which features of scales to focus on more. Inevitably, there are some confusions related to meanings of each word on the scale and the rater must justify his/her judgement depending on the scale. So, these issues make raters the most important factors of the scoring process. 5(p.267)

Huot (1990) also supports this issue by mentioning that there is lots of information, explanations and inferences related to scales and debates which are going on about the influence of criteria in the writing assessment process but there is not much information about the rating process which means it is not for sure that the scoring process totally depends on criteria. Whether the raters depend on criteria at critical points or for exceptional papers or situations is not known. This obscureness raises a question. Do raters have some judgements about the papers before they analyze the quality of the papers? In other words, raters might have some decisions in their minds when they first see the paper and they try to transmit or justify that judgement by adapting it into criteria. So, the criteria function as a justification tool for in the assessment process. (p.258).

Lumley (2002) studied this issue and the question was whether the experienced raters understand criteria in the same way as the other raters and by using think aloud technique the researcher focused on raters' explanations and justifications about the exam paper. What was revealed is that the raters can be trained to be able to talk on the same features of a scale. Moreover, they can focus on and discuss expectations from the learners. Discussion on each band of a scale leads to a better understanding for a rater. (p.253)

Rater training is supported by another researcher Yancey (1999) as it is mentioned in his study criteria, scales and set of rules help in scoring and they assist raters in the assessment process but there are point differences in two different ratings which means the discrepancies might be lowered in time. Every difference in papers should be negotiated and well explained until it is fully justified. This negotiation requires more time to make interrater reliability higher. But this time which is separated for discussions enable raters to come to an agreement and it will be beneficial. This is exactly what is done in standardization sessions to decrease discrepancy points between two raters.

2.9. STANDARDIZATION SESSIONS (WORKSHOPS)

As it is quite common to see human scorers rating constructed tests of items, standardization is crucial, especially in large scales. Standardization refers to interrater reliability; different raters should reach closer results. With the aim of gaining more reliable results, papers should be rated twice by different scorers to see discrepancy points between the papers. If the aim is getting lower discrepancies, the scorers should be trained.

Alderson, Clapham &Wall (1995) states that the first step should be familiarization with rubrics second, consistency in using them and then getting ready to cope with unexpected problems and papers. Constant training and updating should always take its place even if the raters are experienced. They should be reminded and updated. The Chief Examiner defined as the person or a group of people who are responsible for the rating process, guidelines and preparing or choosing rubrics according to the expectations. It is advised that after getting exam papers before the rating process begins the Chief Examiner either alone or with a group of collogues which is suggested, should choose immediately 15 or 20 problematic papers. It is known that some paper might be kind of off-topic, short in word limit and differently answered. Those papers should not be checked properly with rubrics but should be selected as an example of common mistakes that occurred in the exam. After choosing problematic papers, the Chief Examiner and the committee members should rate the papers individually by using rubrics before scoring the paper all together with the other raters. Members should compare and discuss the results while taking down notes. The Chief Examiner will be the leader in the standardization sessions and should have excellent knowledge about the rubrics. One whole day training is suggested and before raters come to the meeting room, they should get samples of rating practice exam papers and the rubrics. Sessions should be done before the rating process. When all raters come to the meeting room, problematic parts of the exam papers are together, but the committee should not give any explanations since this part is essential to not to influence the raters. As a next step, problematic papers which the raters have not seen yet should be presented. Papers should be again rated and discussed at the meeting. The CE should observe raters while they are scoring and if anybody needs help or an explanation, the CE should provide help. It should be noted that after the session if there is any need for more explanations on the rubrics or changes, the CE is responsible for doing this, but it should be noted that too many changes will cause confusion and obscureness among the raters. These standardization meetings should be constant. (p.105).

According to Way, Vickers &Nichols (2008) another standardization process is described as; raters should be monitored while they are rating. If it is possible previously rated papers should be given to the raters and raters should be asked to score those papers individually, later the results should be discussed, and questions should be answered. Some specifically chosen papers should be distributed. For instance, the ones which are borderline will help raters to discuss critical papers. Raters will understand how to define those papers. (p.4). Another problematic point can be focused on during the standardization sessions and the topics might change according to commonly made errors or how to focus on confusing papers. To eliminate the rater errors, teacher training should be done but at the same time, the individuality of the rater and the scorer should be protected Weigle (1998). So, it means that rater and scorer differences should be respected while the errors are being reduced.

Guillot and Delgado (2017) also support the idea of decreasing the number of errors made by the scorers through training. The individuality of the raters should be preserved. (p.1334). Raters should be familiarized with Common European Framework of Reference (CEFR) to be able to understand the level of a learner and the expectations from that learner. CEFR familiarization helps teachers to know more about learner language proficiency levels Little (2007).

Furthermore, raters should be informed about the rubrics in advance and they should be familiarized with the rubrics. After teachers are informed about CEFR and the rubrics, some sample exam papers and rubrics should be given to raters in groups and they should be asked to put the papers into score order by using the rubrics. The orderings gathered from the participants should be shared and anonymized to discuss the obscure parts of the rubric and the justifications should be clarified. Next, the answers should be shown, and the leading team of the raters should be asked for the justifications.

In that way, raters have the chance of rating separately but discussing the results justifications and clarifications together. Since it is anonymized, raters are encouraged to ask more questions and as a result, questioning will lead to a better analyzation of the papers. Brainstorming is crucial at this point and talking to team leaders will enable raters to clarify obscure points in the rating and rubric. If there are still problems in the rating process which means the discrepancies are higher, raters should be encouraged to ask questions individually.

Another point that is important for standardization sessions is rater tendencies. Observations should be done accordingly to identify what type of raters join trainings. If the aim is using analytic rubrics, the behavior-driven training (bottom-up) will be suitable. In this type of scoring, raters do the rating step by step when they finish one phase, they can move to another one, in this way they analyze each feature of the rubric and they have answers and justifications for each band. On the other hand, if the aim is using holistic rubrics, schema- driven training (top-down) is suggested. Raters do not focus on each band of the rubric to make clarifications or adjustments, but they are thought to make decisions by putting all judgements and decisions related to the whole performance into correct category Lievens (2011).

2.10. STUDIES ABOUT RUBRICS

Broad (2000) mentions that rubrics should prevent speculations about the assessment, it should provide learners a brief and clear explanation about their assessment process. In other words, it should decrease the subjectivity level in the assessment. However, it is known that teachers also have contradictory ideas about utilizing rubrics. Although Andrea (2000) thinks that rubrics are useful tools for teachers during the assessment process, they might not really reflect what students express. Tarkan-Yeloğlu, Y. Seferoğlu, G., & Yeloğlu, H. O. (2013) worked with 55 instructors of a private university and they put forward in their study that rubrics are helpful in the assessment process and it helped to establish a standard grading but still there are problems such as; perceptions and approaches of instructors while they are using the rubrics. Moreover, the study put forward the idea that instructors need training on how to use rubrics. Training is necessary to utilize rubrics effectively. So, the results of this study highlight the importance of writing workshops to make instructors familiar with the rubric. (p.373).

Rezaei, A. R. & Lovorn, M (2010) also suggest that raters should be well trained before utilizing the rubrics. They conducted a study on raters to see the importance of types of mistakes for scorers and a group of writers prepared two different scripts, the first one was contextually well developed but it was poor in grammar, the other one was grammatically perfect but poor in context and covering the questions asked. They found that mechanical errors affect the raters more than the contextual ones even if they use analytic rubrics. Instructors prioritize different types of mistakes according to their beliefs and getting standard results is impossible in this way. So, it is suggested at the recommendations part of the study that the workshops to teach raters how to utilize rubrics are crucial. (p.28).

A study was done at a University in Pakistan. Perceptions of 3 literature teachers and writing teachers were taken on the effectiveness of using rubrics throughout the semester. The data was collected through questionnaires and the results show that teachers found using rubrics effective in the assessment process but at the limitation part it was expressed that teachers have some serious problems for the utilization of the rubrics. So, the problem that the teachers face here is related to the training process of using rubrics Qasim and Qasim (2015). (p.53-54).

A similar study was done in New Zealand. 17 writing teachers were involved in the study. Those teachers attended 5 sessions to score papers with rubrics and in each session, they spent 4 hours to check the papers. In the end, the feedback was taken from teachers through a questionnaire. Results revealed that teachers have positive feelings about rubrics, and they believe that it helped them a lot during the assessment process. They also mentioned that when they got used to using rubrics, it became easier to utilize them. In other words, the results show that familiarity with rubrics makes the rating process more effective and easier. Teachers also believed that writing exam results became more consistent Glasswell (2001). (p.15).

Another study was done by Beyreli and Arı (2009) on 200 papers and 6 raters aimed to show the concordance between two papers checked by two raters by using analytic rubrics. They found out that concordance among raters is enough but during the assessment, raters sometimes scored differently from the rubric because of time, topic, the difference between two texts, lack of concentration, and expression of the writer. Those problems can be solved through rater training. (p.109).

Eckes (2008) puts forward another point of view for the necessity of standardization training. It is claimed in this study that discrepancy points between two ratings might be caused not only from rater experience, rubrics' variety and field experience differences but also individual differences in scoring. It means that raters have some tendencies while they are rating. They might give importance to some bands more than the others. As it is displayed in this study six different rater types were found. The study was conducted on 64 experienced raters and a 4-point scale was given them to number 9 scales that they commonly use. Criteria have some different bands of features such as grammar, completion of task and fluency. The raters chose what was important for them. For sure, some parts were superior for raters and during the scoring process something in their minds gave more importance to those sections than the other ones. Results revealed that fluency, structure, correctness, non-fluency type, syntax and non-argumentation types of raters exist. For example, while a syntax type of rater prioritizes language proficiency in linguistics, a fluency type of rater cares more about the task and task achievement. This information will be useful for trainers and chief examiners in standardization sessions. In the discussion part of the study, it is suggested that rater training can overcome these problems but the influence of the training on rater types is not clear. (p.181).

Lumley (2002) emphasizes the importance of rater training by highlighting that the scoring process mostly depends on the rater rather than the rubric. He conducted a study to question raters' role in the assessment process. 4 raters and 24 scripts took place in his study. 4 raters checked the first set of 12 papers individually by using a rubric and the rater agreement was acceptable. Later, the same 4 raters checked the other set of 12 papers, but this time the think aloud process was involved. Raters were asked to explain how they make decisions and they were asked to justify their answers and this process was recorded. Results showed that even though the discrepancy points were acceptable, raters' explanations and justifications did not fit the rubric all the time. As it was observed, teachers gave their reasons to score in that way and they looked for justifications from the rubrics. At the recommendations part of the study it is highlighted that this problem can be overcome by training or meetings where some explanation, agreement discussions, group and individual work happens. (p.253-268).

Debates about rubrics and rubric usage have been a concern for raters for a long time. The studies above mention some of these reliability concerns of rating procedure. Discrepancy differences between raters are one side of the assessment of writing while the other side is related to the students. As Reddy and Andrade (2010) suggest rubrics can serve as a tool to support the learning process. With the help of rubrics, students have a better idea about what is expected from them and they improve their writing. This made the explanations necessary for the students and this time, students tried to understand the rubric that they were assessed with. As we can infer this led to many perception differences among students. Some other studies investigated the discrepancies between students and teachers. With the aim of defining these criteria, interpretation differences between students and teachers Li & Lindsey (2015) carried out a study. 5 teachers and 119 students took part in the study and qualitative and quantitative data were collected through interviews and questionnaires. Teacher and student interpretations about holistic rubrics tried to be identified. As it was expected the discrepancies between teachers were less than the discrepancies between students and teachers. While teachers focused on the whole sentences on the rubrics, students' focus was at word level and they made some other, unrelated interpretations. This study shows that discrepancy discussions will last and even some new concerns will arise more in the future. Rater- rater, student-rater conceptions and interpretations will vary, more rubric application studies will be necessary to make use of the assessment and learning producers. (p.68).

2.11. STUDIES ABOUT STANDARDIZATION WORKSHOPS

Some educators believe that using rubrics do not give more objective results. Untrained raters even can use rubrics to reflect what is in their mind and just to find some justifications for their scores they depend on rubrics. Kohn (2006). As it is stated in the studies above; rubric application is a concern. They mostly faced problems with rubric usage in the rating process and interpretations of raters. As a result, at the recommendation part of their studies, they suggested rater training on the application of rubrics. (p.14).

According to the study done by Brown (2004), even lower levels of training might be helpful to the assessment process. Interrater reliability increases because of these trainings. In their study, they designed a half-day training program for state school teachers who have no experience in writing scripts scoring in large scales. (p.117) Language Test Construction and Evaluation Alderson (1995) was used as a guideline for workshops. Experienced raters were the Chief Examiners. They observed raters during scoring, and they helped them. The half day training took 4 hours in total and they found out that even 4-hour training could help high school teachers to score papers. The only difference was that teachers were slower than usual, but this was regarded as normal since it was their first time. On average 7.2 papers were checked per hour by the raters. The areas that teachers needed more instruction were grammar, punctuation and complex sentences. Consensus, consistency and measurement approaches were considered while reliability was calculated. Overall, the close agreement rate of the teachers was 75 %. And the results of this study suggest that a 2-day training program will be enough to reach better results in writing exams.

Wolfe (2009) conducted a study to show the results of context difference of rater training and scoring on rating quality, time spend for scoring, training time and rater perceptions.120 raters took place in the study. 3 sets of 40 raters joined one of these training sessions. Raters had almost the same backgrounds and experience.

These training sessions were online distance training, face to face training, and online training in a center. 400 papers were checked by the raters. In all this training, raters scored papers individually and they have a chance to ask questions when they need to through telephones, e-mails, or face to face. 20 papers were rated together with the experienced raters. An online distribution system was used to give the papers to the raters. Raters' performances in three settings were recorded and a questionnaire was done to investigate the perception of raters. A holistic rubric was used as a guideline. Training time and scoring time in each set were recorded. Correlation between the scores and back reading agreement was noted, frequencies for assistance were recorded. Results of the study show that the highest training time was spent in face to face training and online training in a center since face to face communication causes more questions, discussion and raters introduce themselves and speech goes on. Face to face training was three times longer than the online versions. Scoring time difference was not significant but online rating took less time compared to others. Agreement rates were better in online training and face to face training but all the groups had %70 agreement or higher. For back reading, the difference between the groups was not significant but there was a slight difference in online training groups. Back reading (the agreement between the raters and experienced raters) was better in online groups. There was no significant perception difference of raters about the three different trainings. While raters asked for more help about the scoring process in online training, raters in a center asked more questions related to computer intervention. It can be inferred from the study that training context does not cause significant differences but there is a difference in training time. In that case, online as opposed to face to face training there is also no significant difference in raters' scoring time in three different settings. In online versions, raters do not have to wait for the slowest rater but in every case, standardization effort makes the rating process better. Neither the quality of the performance nor the perceptions of raters change in three contexts. (p.5-17).

As we can infer from the studies standardization training can be done in various ways such as online training, face to face training, rating papers together with experts, getting help from experienced colleagues or Chief Examiner. All types of training improve raters' skills in scoring. There is an example of a study done by

Greer (2013) to prove the idea that even experienced colleagues can help inexperienced teachers in rating training. The study was conducted on inexperienced teachers and those teachers were supposed to rate composition papers of ESL learners, teachers followed a guideline that would be helpful to them in their rating process. Inexperienced teachers were given the commented papers of experienced teachers so, the comments on the papers helped them to understand the rating process. After that new teachers commented on their feelings after the training. It is reported that inexperienced teachers gained confidence in their paper rating performance. (p.38).

Another study based on a two-month training program; 1700 writing samples were given to 13 novice teachers as a weekly assignment. Group discussions and individual working times were allocated to the raters. Teachers spent a lot of time working on those papers and the result was not the number of papers, but the discussions and questions made the training process effective. Teachers had various tasks to complete related to revising and analyzing the features of bands on the scale. So, in this research making an agreement on each band and discussion on that were prioritized Harsch and Martin (2012). (p.239).

Fahim and Bijani (2011) explain the decrease in discrepancy points after training. The study was done on 60 student papers and 12 EFL raters. The study shows a difference in the ratings of teachers before and after the ratings. First, 60 writing exam papers were collected from advanced learner students and 45 papers were chosen randomly out of 60 papers. IELTS trainers read 45 papers and they were kept checking the differences with the ones 12 raters scored. All the papers were typed not to influence the rater because of their handwriting. The first 15 papers were given to the raters to be scored before the training and the results were recorded and compared with the results of IELTS trainers. Papers were checked by using a rubric which has four aspects of writing. During the summer, raters were informed and familiarized with the rubrics by making scoring practices on several papers. They were given some CDs to revise the training process. Immediately after training 15 papers were scored by the raters and discrepancies between IELTS trainers and raters were taken for before and after training. The discrepancies were converted into z –

scores and z values were obtained. The last 15 papers were given to the raters as a post-test and the same procedure was followed to get z values. The results of their study point out that raters become consistent in scoring and they mostly overcame the problems of giving lower or higher points, but they could not eliminate them totally. (p.11).

Moon and Hughes (2002) conducted a similar study and they worked on a large scale. They wanted to see the change of discrepancy points after training as Fahim and Bijani did in their study, but the difference was two different training methods were compared, sequential and spiral models, and the study investigated which method was more effective. As it is described in the study the sequential model refers to training where raters take exam papers, guidelines and explanations and rate some exam papers that were previously checked by the committee. After rating the papers, discussions are held to give explanations and clarifications. The spiral model is explained again with the same procedure but with a difference in the cognitive process. During the discussions, raters are supposed to turn back to the rubric for each prompt and they should be sure that prompts that are taken from students, justification and scoring mesh correspond to each other. This process makes raters rethink the process and refer to the rubrics again and again. It requires more time to rate than the sequential model needs. So, the study was conducted on 3,660 essays of 342 students. 94 raters took place in the assessment process. These were all experienced raters. Each rater scored about 60 papers. Rater agreements were recorded after each rating with the training of two different models. The result presents the idea that the spiral model is more effective than the sequential model since the rater agreement is higher. (p.16-17).

Another concern arises when raters see different prompts which means different kinds of exam questions. During the training before scoring the exam papers, raters are trained to check those papers. After scoring many papers, raters have some criteria in their minds, and they lessen their reference to the rubrics. When they have new prompts on the exam papers to score, they may tend to use the same criteria in their minds, or they still look for the same kind of features on the exam paper. This result was gained from a study that Weigle (1994) conducted. Her study involved 8 novice and 8 experienced raters. And 30 papers were checked by them, there were also two different prompts. The results reveal that there might be interference of other elements in the rating process even though rater reliability increases. Training also helps to have a better understanding of the criteria and the bands on the criteria are clarified. Moreover, raters felt that they could use rubrics more properly thanks to training and during sessions they mentioned it.

Smith (2000) supports the findings of the previous study with the results of his study. He worked with 6 experienced raters and he used the think aloud technique. The papers were all at borderline. Raters were all consistent and they could analyze the papers truly and during the assessment process raters referred to rubrics several times. But it was revealed that raters also mentioned some other features of the papers which were not expressed on rubrics. Furthermore, it was found out that teachers or the raters did not understand the same thing from the criteria. Their inferences were also different. The raters got closer in terms of holistic scoring but when the case was a specific item or features on the paper, they could not agree on that much. So, Smith also put forward the same idea as Weigle (1994), there might be some factors related to other things that are affecting the process. These different findings of the training process show us that we do not have enough information about what really happens during the rating process.

What is really happening in the rating process cannot be defined but it is mostly related to inexperience in scoring and studies mostly focused on training issues and how to design and conduct them. They focused on the frequency of the meetings and discussion on the rating process, referring to the criteria, selecting appropriate rubrics and adapting and reforming them according to feedback. So, these are all helpful to make the assessment process more objective but on the other hand, the existence of other factors has been discussed a lot. For example, Lindsey and Crusan (2011) stated that raters might make different inferences for each learner even if they use the same rubric for all the learners because of the predictions about writers' backgrounds. The identity or the cultural differences of the writer might be the cause. Another factor in the rating process can be the topic selection, Yang, Lu & Weigle (2015) emphasizes that each topic requires different grammar structures, use of language and content knowledge to be fully written. (p.64) Using the same rubric to assess all these different topics might cause a scoring reliability problem. There might be many other problems that cannot be listed. Even in this complex procedure, some educators do not believe in the effectiveness of rater training. There are still some arguments that inter-rater reliability does not change and get better after the training process Pufpaff, Clarke & Jones (2015). It is believed that whatever is done will not be helpful because the trainers are humans and there will be always other factors that cannot be even defined. While this created questions, Weigle (1994) explains the situation as training does not make raters better in scoring but it helps them to be self-consistent. Raters reach a standard after training. They know how to judge language elements and they refer to rubrics more frequently and this leads to consistency which makes raters more reliable markers, and this happens because of the positive effect of the training process. Another explanation came from Huot (1990), even if there are some other factors or some questions related to training processes and their quality, we should continue to train raters since this is the best alternative. Cumming (1990) also supports rater training by mentioning that experienced raters improve some techniques to score the papers. Moreover, they evaluate themselves in the rating process. In time, raters feel more comfortable in the assessment process.

The studies investigating rater agreement mostly benefited from Kappa calculation for categorical data or Intra-class Correlation Coefficients (ICC), if the data is numerical for quantitative researches Fleiss (1973) p (613). ICC measures were used to analyze rater agreement in some researches while some others benefited from it to support their main research question and they eliminated obscurity in their study. ICC measures were also used in this study as the main source of quantitative data analysis and some other studies' ICC measures used are listed below.

Engemann and Gallagher (2006) focused on inter-rater reliability rates of 10 teachers. Two discussions sessions, each of which, lasted 3 hours were held. Teachers were involved in conversations related to the assessment process. They talked about the rubrics and interpretations of them. Teachers rated scripts by using rubrics and overall scores for each paper were noted. To be able to know how close the teachers were in ratings, the intra-class correlation coefficient (ICC) was used to determine the rater agreement rate. For three different rubrics, the results were presented as 0.95, 0.94, 0.05. These were named as high rater agreement rates and these results were gained after a long discussion session. It was mentioned in this study that discussions on the assessment process helped raters to give consistent scores. It was also mentioned that either during workshops or after workshops teachers' conversations helped rater agreement. Moreover, rubrics were used more appropriately after workshops p (40).

(ICC) Intraclass correlation coefficient was used in another study to investigate the inter-rater reliability Cho (2003). The study mainly focused on the methods for the writing assessment. They used different types of writing assessment for placement tests in their school. More reliable results were needed for the evaluation of direct tests. For that reason, two different raters graded 57 scripts. Raters underwent 4 training workshops, after that 3 raters rated those scripts. Every week each rater graded 20 scripts. ICC was used to analyze rater agreement. It was found that the values changed between 0.77 and 0.95. Value 1 represents the exact agreement in ICC p (175).

Dunsmuir (2015) conducted a study that investigates the validity and reliability of writing assessment measures. The study involved 97 primary school children in England and Wales. The assessment item was found valid and reliable. Inter-rater reliability evaluations were done to check whether the items give similar results when they are rated by different scorers. Rater agreements were checked with an ICC evaluations two-way random model. 4 different raters graded 20 scripts. The average ICC was 0.97 which was regarded as high p (9). It was found that raters agreed more on structural and mechanical items rather than ideas. p (12).

In some other studies, Spearman correlation coefficient measures were used to determine rater agreement Park and Stapleton (2003). The main aim of the study was not to see rater agreement but still, correlation measures were needed to define correlations between L1 internal voice and success in L2 academic writings. To investigate this correlation, they must be sure that writing assessment was done fairly. As a result of this, scripts of 63 students were rated by 6 different scorers and they also used two different types of rubrics to be sure that every single item was controlled. Before the scripts were scored, a two-hour rater training was done. Raters graded some sample papers during this training. The inter-rater correlation was found to be 0.73. As a result of the study, they could not find any relationship, but this might be the consequence of the scale chosen. The correlation remained in doubt. p (251).

Neumann (2014) also benefited from the ICC calculations two-way mixed model which was used to check the reliability in error coding. The main aim of the study was to investigate the effectiveness of the feedback given by writing teachers to students. First, teachers coded writing errors on student papers. To check the correctness photocopied papers were coded by the second rater after a training session. Rater agreement was defined with an ICC value range from 0.88 to 0.96. p (89).

Song and Caruso (1996) also worked on inter-rater reliability with another method. The measures were done by using two-way ANOVA. 30 English and 32 ESL professors took part in the study. 4 scripts were chosen from a real exam. Professors were grouped in fours. They also evaluated the papers with holistic and analytic scales. Rater agreement of two different groups was compared in the end. There was no meaningful difference in analytic scoring but in holistic scoring English, faculty gave higher grades. p (166).

ICC calculations were used by Sweedler-Brown (1993) to show the correlations of analytic and holistic scoring. Correlated scores were found as a result of the study. The main aim was to conduct a study to show the focus of English writing teachers while they were rating exam papers. Whether they gave more importance to structures or rhetorical features of wiring was investigated. They used two different types of scoring criteria to strengthen their analysis and to show the correlations or the consistency between two scoring criteria. Instructors gave more importance to structures and analytic and holistic scorings had correlated results. p (3).

CHAPTER 3

3. METHODOLOGY

3.1. PRESENTATION

In this chapter, research design, methodology, setting and participants, data collection methods; benchmark results, rater correlations and interviews are presented. The importance of the study was mentioned.

3.2. RESEARCH DESIGN

This research displays the effectiveness of new regulations of a private university throughout the accreditation process. Regulations were redesigned to increase the rater reliability in the writing assessment process. As descriptive studies present the situation, or issues without having control of the researcher, this study also analyzes the results of a process. The phenomenon is displayed in its natural setting in descriptive studies. There is also a chance of using qualitative and quantitative methods at the same time Ethridge (2004). Using mixed methods in research provides numerical and in-depth information that strengthens the study. In this research, while correlational and z-value analysis provides numerical information related to the process, interviews provide in-depth information to the phenomenon. (p.24).

Mixed methods ensure the analysis of quantitative and qualitative data but at the same time, the researcher might choose one of them as the main research item while the second item supports the research. This issue is named explanatory design, which is one of the three types of mixed- methods. In explanatory design, the researcher chooses quantitative data as the main domain of the research and provides some extra information to support the research by qualitative data. So, in this research numerical data is regarded as the main domain of the research while interviews provide some supportive data to the research Frankel (2012), (p.557,558,560).

3.3. SETTING AND PARTICIPANTS

3.3.1. Background of the English Preparatory Program

English Preparatory Program provides English support courses to students in their faculties or schools. The program has four modules and each module lasts 8 weeks. Students take skills and main course lessons separately. Apart from the quizzes, students have skills (writing, reading, speaking, listening) and main course progress exams in each module. Most of the students start the course from the elementary level and they finish at the upper-intermediate level. At the end of the 4th module, according to average scores of four modules they pass or fail. The data of this study was collected from the progress exam results of the students.

3.3.2. The Participants

842 scripts of 421 students were collected in total. The first half of the scripts were taken while students were at the intermediate level and the second half was collected when they were at the upper-intermediate level. Opinion essays were asked for in both exams. 621 students were at an intermediate level when they took the first writing exam in the study. After two months passed there were 421 students left at the upper intermediate level due to dropouts and absenteeism so, 200 students' papers could not participate in the study. The first half of the intermediate level scripts were checked by the same 28 raters before the standardization sessions and the second half of the upper intermediate level 421 scripts were rated after the training (2- month period). It was the students' first year at a university, they were all prep students. So, their ages mostly ranged between 18 and 22.

28 instructors participated in the workshops (sessions). 28 instructors were assigned randomly to rate the exam papers before and after the workshops since. The same set of raters also graded benchmarks in the last workshop. 5 of the instructors were male and 23 were female. Instructors were non-native speakers. Their experience in the program differs From 1 to 4 years and their ages ranged between 24

and 32. Instructors graduated from ELT or English Language and Literature departments. None of them had gone through a rater training or standardization session before. All the instructors attended 4 standardizations sessions.

Standardization sessions were conducted by 3 testing unit members, a department chair and a vice principal. They were guided by Alderson (1995). I also personally took part as a testing unit member in this study. Testing office members' experiences in teaching vary from 2 to 4 years and they all graduated from ELT departments and two of them are MA students of ELT departments. The department chair has 7-year experience in teaching and acts as Chief Examiner within the unit. The vice-principal has 25 years of experience in teaching. Standardization scripts were selected and rated by this committee before standardization sessions. They also rated 3 exam papers randomly from each exam pack to see if the raters are in the right band. Papers with high discrepancy problems were checked again by this committee.

3.3.3. Data Collection Methods

The data is collected both qualitatively and quantitatively. Qualitative data is collected first through interviews with 4 committee members of standardization sessions. Interviews were done after all sessions were over, committee members answered the questions about problems they faced, how the process went, and the improvement related to the scoring process. The quantitative data is first collected through 28 rater's discrepancy points on 842 papers. The correlation difference in the rater agreement was analyzed before and after the workshops. Next, rater scores on 3 scripts were collected from 28 raters in the last workshop and discrepancies were compared with benchmarks.

3.3.3.1. Benchmarks

Benchmark analyses were made to see how close raters get to the benchmarks in the last workshop before the post-test of the research. So, this analysis provided the data that would be helpful to make interpretations in the post test. It might be regarded as a while test in the research since it investigates the correlation results of raters before post-test and analyses individual discrepancy points from the benchmarks. It provides information about raters individually and defines their current situation before the post-test was done. First, the script samples were chosen from a previous exam. Especially, scripts that had high discrepancy differences between two raters were selected and they were rated by committee members. Each committee member got a sample and rated the script individually. Next, they discussed the results within the committee and explained the reasons behind them. Before the last workshop, benchmark samples were collected and rated by committee members and an acceptable discrepancy gap was determined Alderson (1995). During the last session, raters graded papers without any discussions with partners. They were observed by committee members while they were rating the scripts. After the scored rubrics were collected, group discussions were done. Raters had discussions on papers with committee members. Discrepancy differences were explained, and an appropriate range was presented. After that committee members noted the grades given by the raters. 3 different scores were noted from 28 raters. To analyze how close the raters were to the benchmarks, z-value analysis in SPSS was used. Z-value gets the mean scores of raters and presents how far the raters are from the mean. First, benchmarks were determined by getting the means for each script. Means and benchmarks prepared by committee members were at the same range. So, the scores for each paper were mean scores. Z- value defined how close the raters were to the benchmarks. ± 1 defined the leniency or severity of the raters Fahim and Bijani (2011). (p.8).

ICC two-way mixed effects was used to analyze the interrater-reliability rate in the last workshop. Raters were a set of raters and the papers were not given to them randomly. In other words, raters were not distributed papers randomly. 28 raters rated same set of 3 scripts. The ICC results presented the closeness of raters in scores to each other. This analysis provided data about the current situation of the raters before the post-test. The Inter-rater reliability rate in the last workshop to see the effectiveness of the workshops and to check whether there is a difference in real exam and training environment situations Koo and Li (2016). (p.157).

3.3.3.2. Rater Correlations (ICC)

Correlational research might be sometimes called associational research since it analyzes the relation between two or more variables. It might be sometimes identified as descriptive research because it describes the situation between variables. The correlational study clarifies the relation by providing a degree in quantitative studies and it does this by using a correlation coefficient. When there is a correlation between two variables a correlation coefficient is used, and this is mostly used in educational studies. It can be also used to check the reliability of two different scorers Frankel (2012). (p.331-340). Correlation does not only describe reliability but also the rater agreement between scorers. ICC defines the agreement rate between two scorers or more. So, discrepancies between raters can be defined with ICC scores. Two-way random effects model was used since the raters were selected randomly and assigned random classes to score Koo and Li (2016). In this research, to analyze rater correlations some steps are taken, first scripts were collected, an appropriate rubric was chosen and developed by committee members then standardization sessions were held, and scripts were scored and finally, discrepancies between two raters were noted in pre and post-tests. (p.156).

4.3.3.2.1. Scripts

Since Weigle (1994) mentions that there might be some changes in results if you change the type of the essay, an opinion essay type of question was asked to students in both progress exams. 621 students attended the first progress exam at the intermediate level but some of them did not attend the second progress exam at the upper-intermediate level so two different scripts of the same students were collected from the participants of both exams. So, the number of students lessened to 421. After two exams were done, 842 papers were collected to be rated by 28 teachers. Because of the number of the students' papers to be scored, it was impossible to type them. So, all scripts were handwritten.

4.3.3.2.2. Rubric Selection

Weigle (2002) emphasizes the importance of rubrics in the assessment process. The design of the rubric can affect the rating procedure. In large scale assessments, the most common types of rubrics are holistic or analytic ones. (p.72). Weigle (2002) mentions that analytic rubrics give more reliable results compared to holistic ones. (p.73) So, in this study analytic scales were used to score scripts and a Cambridge assessing writing performance (2016) rubric was used as a base in the design of the current rubric (Appendix A). Committee members rated some of the scripts before redesigning it and they realized that some changes were necessary to adapt the rubric to the goal of the course. One more band was added between fair and poor so that raters can identify in between papers in a better way. After the first familiarization, the rubrics were never changed again since it would cause confusion among the raters. The reliability and validity check of the modified rubric could not be done statistically because of the institutional constraints but instead, expert opinion check was carried out by the experienced committee members in the university and the instructors found the modified rubric purposeful and coherent and did not report, either formally or informally, any issues with the new rubric. So, for the trustworthiness of the new rubric expert “expert opinion” and “member check” were done. These two measures are commonly used in the literature, as Nevo (1985) mentions, “expert opinion” is quick, cost-effective and it enhances the credibility of a product. Moreover, the experts' opinions can be used to modify a product. Lincoln and Guba (1985) propose a four-point criterion list to establish trustworthiness in naturalistic inquirers. "Member check" is regarded as the most important technique for establishing credibility. Members can help to develop a more sophisticated understanding of the phenomenon p. (314).

The rubric has 4 categories, for each of them there are 6 bands to describe the quality of the work from ‘excellent’ to ‘poor’. While ‘excellent’ meets the requirements of the task fully, ‘0’ refers to irrelevant content and performance below 1. The categories are content, organization, language and mechanics. Content defines topic comprehension and informing the reader fully by providing some examples and supporting the idea. The organization defines the layout of the text, linkers and cohesive devices while presenting the topic. Next, the language category looks for vocabulary selection and complex grammar items for the presentation of the content. Selection of appropriate vocabulary and its correct usage. Finally, mechanics gives information on whether the structures are used correctly or not. The usage of the grammar is evaluated in this section. The category was not named ‘grammar’ since it was presumed by the committee that instructors would tend to prioritize grammatical errors. Students get points from 4 different categories and each category has 6 bands from excellent ‘5’ to ‘0’ in the end total points gathered from each band is multiplied by 5 to reach the total score.

4.3.3.2.3. Standardization Sessions

The School of Foreign Languages underwent an accreditation program. As a result, new applications were necessary, but the school authority had extremely limited time to adapt to new regulations. After the selection of rubrics, instructors were handed new rubrics in a meeting with a brief explanation and they were supposed to use those rubrics for the exam which would take place in a couple of weeks. Exams were checked by two different raters with new rubrics, but the results showed that standardization sessions were the inevitable need of the instructors since there were more than 200 papers waiting to be checked for the third time. So, before the training, results were taken from that exam as a pre-test to analyze the differences in discrepancies before and after the training sessions.

After the exam results were received**,** raters went through 4 different phases of standardization sessions in two months first, as Way, Wickers & Nichols (2008) suggest raters should be familiarized with CEFR levels and expected performances of those level of learners. A professional trainer from Pearson came to explain CEFR levels. After a long presentation, the practice part started, and the trainer showed a lot of prompts like” who can achieve… “or “ ……. level of students can introduce themselves’ on his slides. Instructors first completed those prompts by themselves and later they discussed with their partners and finally prompts are discussed all together. The total training lasted approximately 2 hours. (p.3).

Table 1. 4 different phases of standardization sessions

content		duration	week
Phase 1	familiarization with CEFR levels	2 hours	1
Phase 2	familiarization with the rubric	1,5 hours	3
Phase 3	paper rating	2,5 hours	5
Phase 4	paper rating	2,5 hours	7

The other 3 standardizations were formed by testing unit members, two vice principals and the department chair. The department chair acted as a Chief Examiner and designed sessions with his colleagues and they together selected papers, rated them and discussed them together before the sessions. As it is suggested in Alderson, Clapham & Wall (1995) familiarization with the rubrics is crucial. So, in the second session instructors are put in groups and handed rubrics to discuss what they understand from each band. Some previously rated sample papers (high, low, average scored papers) were distributed by the committee without final scores on them and the instructors were asked to put them in order after that results were discussed with explanations and justifications from the rubric. The rubric was discussed in detail to be clear. This session lasted about one and a half hours.

In the last two sessions, 7 different scripts that had some different problems like, short but full purpose achieved, or borderline were chosen from a previous exam as Alderson (1995) suggested. Papers were rated first by each committee member and then discussed. Justification and explanations were noted on papers. When the sessions started, instructors first asked to discuss the expectations from the students then the expectations were listed all together. Next, papers were distributed without any grades on them and raters first scored them individually then discussed the scores within groups and later with everyone. Concerns and questions were tried to be eliminated at that point. The chief examiner was observing everyone in case they needed some assistance. Each session lasted about two and a half hours. Sharing concerns, ideas, questions, explanations, justifications required a lot of time.

These sessions were held every other week on Wednesdays for the first half of the group, on Thursdays for the second half of the group since it would be very crowded to form one big group and it would decrease the effectiveness of the workshops. Throughout the sessions, two groups were mixed and formed new groups to give instructors the chance of having discussions with some other raters. Referring to rubrics for each item was highly suggested during the workshops and this was also observed. The total time spent for rater training was almost 8 hours in total. This can be regarded as a good amount of time spent for sessions as Weigle (1994) defines rater training sessions as two hours norming process. Since it was crucial for instructors to feel comfortable discussing and ask many questions, the sessions were not recorded, instead, the scripts they worked on for scoring were collected.

4.3.3.2.4. Scoring Scripts

The first writing progress exam was done, and 28 instructors were asked to do the first checks of the exam papers. Instructors were informed not to put any grades or markings on the papers but to fill rubrics for each writing exam. After instructors had rated the exam papers, they separately submitted those papers and the rubrics to the testing office with final grade sheets. Instructors also sent scores to the testing office via e-mail. Testing unit members took out the grade sheets and rated rubrics from the envelopes and gave them to new raters to be checked. After that, exam papers were distributed to other instructors for the second checks by the testing office. Instructors were asked again to fill a rubric for each exam paper, bring back the exam papers and send the results via e-mail. So, during the first and second checks, instructors put neither the grades nor markings on the papers. Instructors were also not informed about the second checker and papers were not distributed for the second check before all the instructors submitted the first checks. Two different grades collected from two different instructors were noted by the testing office and according to discrepancy points between the papers, some papers were again sent to a third check. Moreover, committee members rated 3 different papers randomly in each class after they received the envelopes from the raters. This helped committee members see two raters' decisions, in other words, if the raters were both in the range or not.

4.3.3.2.5. Discrepancy Policy

After getting two different scores from two different raters, final scores should be adjusted. The rater mean method was used to ascertain the last scores of the students. It is simply getting the average of two scores Johnson (2000). The average can be taken if there are 2 discrepancy points on the rubric. If the difference is higher than that, the papers are sent to a third check. In that case, three different scores appear to adjust. But the problem is which scores should be taken into consideration, all of them or the closest ones. So, the parity method Wolcott (1998) is used to ascertain the score after the paper return from the third check. It means that the paper is checked for the third time by a colleague then the average of the three grades can be taken to adjust the final score. If one of the scores is hugely different from the other ones, then the committee does the third check.

Table 2. The order of the events

	Student	Raters	Papers	Time Period	Duration time
new rubric was introduced		32		week 1	
first exam was done	421	28	421	week 3	50’
1st session was done		32		week 5	120’
2nd session was done		32		week 7	90’
3rd session was done		32		week 9	120’-150’
4th session was done		32		week 11	120’-150’
Second exam was done	421	28	421	a day after	50’

3.3.3.3. Interviews

Interviews were carried out to provide more detailed data. A semi-structured interview model was applied. This model is highly effective in qualitative researches. Especially when the interviews are done through the end of the study. The researcher has a chance to investigate and observe the questions in his/her mind. The realness of the hypothesis can be checked. The information gathered from the interviews supports the analysis of the data Frankel (2012). The aim of the present study is to analyze the effectiveness of writing standardization sessions. So, 4 committee members who were responsible for the process answered the researchers’ questions. The researcher focused on some issues about the process and asked the opinions of the committee members about the effectiveness of the process, whether they came across any problems, how they found solutions, some specific examples from the process if there were any repeating problems, how it ended, and whether they used any methods or referred to consultants. Some further questions were asked depending on the responses of the instructors. Permission was asked to record the interviews, but since it was an ongoing accreditation process, committee members did not want recording done. They only let the researcher take notes during the conversation. Interviews provide the answer to the second question of the research. “Do workshop conductors believe that workshops were effective? “ and “What were the difficulties workshop conductors faced during the process?’ (p.451).

3.3.4. Data Analysis

The data gathered through interviews was analyzed qualitatively. Interview data questions were designed within the frame of writing standardization workshops. Questions were prepared according to needs, observations and problems. First, quantitative data was gathered through the discrepancy points of two different raters before and after the workshops. To analyze the data SPSS 15.0 statistical program was used. The data was analyzed with the help of the Intraclass Correlation Coefficient (ICC). ICC results of pre and post-tests were compared. The second quantitative data was collected from 28 raters’ ratings on 3 different papers in the last workshop. The data was analyzed with the help of z-value analysis. So, the closeness of individual raters to the benchmarks was presented and the inter-rater reliability rate was analyzed in the last workshop and could be compared with the post-test results.

CHAPTER 4

4. RESULTS

4.1. PRESENTATION

The results of the study were presented in this chapter and the analyses were made by using different types of methods. First, a benchmark analysis was made to show how close instructors get to the benchmarks in the last writing workshop. Z-score analysis for individual scores and intraclass correlation coefficient (ICC) for interrater reliability was used to define benchmarks. Next, rater correlations were analyzed with ICC for pre and post-tests and the results were evaluated. Finally, interviews made with workshop conductors were presented in the last section.

4.2. ANALYSIS

4.2.1. Benchmark

Benchmark analysis contributes to the research by providing extra data related to the performances of the raters in the last workshop. So, this data is different than the ones provided in the pre and post-test of the research. So that it aims to support the findings of post-test analysis after the final exam. Benchmark analyses were done by using two different methods. The proximity of each rater to the benchmarks is defined with z-scores of SPSS and overall absolute agreement is reported by ICC in SPSS.

4.2.1.1. Z Score Analysis

In this part of the research, the analysis was done to show the performance of the raters in the last writing workshops. It aims to show how close raters can get the benchmarks before the final exam. 2 points discrepancy in the rubric which equals 10 points out of 100 is accepted by the school authority. So, this difference is regarded as normal. The distance between the benchmark and the rater is defined by the help of z-scores. So, the raw scores were turned into z-scores by using SPSS 15.0 These new scores helped us to understand how close the raters get to the benchmarks. When the discrepancy points increase, raters get far from 0. Raters who have 0, or -0, scores are regarded as they are in the range. In other words, they are in the range of 2 discrepancy points in the rubric. Fahim and Bijani (2011) highlighted that biasedness increases when the z-scores get closer to ± 1 in their research. (p.10).

As it is shown in Table 1 raters who have 0, or -0, scores are in the range. Here we can see that 17 teachers are in the range in the first marking which means that their discrepancy points are normal. R22 seems far from the benchmark in rating Paper 1. When we look at the second paper, we can observe that 16 teachers are in the range, their discrepancies are not problematic. 10 of these raters are the same as with the ones in the range in the first paper which means that 10 raters are in range both in the first and second rating, but the other raters could not be in the range in both ratings. For instance, R3, R6, R10, R13, R22 and R24 are not in the range in the first rating but their discrepancies are acceptable in the second rating. R15 is far from the benchmark in Paper 2 and not in the range in Paper 1 but is in the range in Paper 3. There are 19 raters whose scores are acceptable in Paper 3. R12 and R16 are not in the range only in this rating which means their performances are not as good as they did in rating Paper 1 and 2. R4 is far from the benchmark in rating Paper 3 but has a better result in Paper 2 and is in the range in Paper 1. In total, 7 raters have no discrepancy problems in any of the ratings and there are only 3 raters who are not in the range in any of the ratings. Negative scores show that raters become more severe while positive scorers become more lenient.

Table 3. Z score analysis of the raters

Raters	Paper 1	Paper 2	Paper 3
R1	-0,6	1,0	-1
R2	-0,6	1,0	0,5
R3	-1,7	0,2	-0,1
R4	0,5	1,9	2,9
R5	0,5	0,2	-0,1
R6	1	-0,6	0,5
R7	0,5	-0,2	0,5
R8	-0,6	-0,6	0,5
R9	0,5	0,2	0,5
R10	1	0,2	-0,7
R11	0,5	1,5	-0,7
R12	0,5	-0,6	1,7
R13	-1,2	-0,6	-1
R14	1,6	1,1	1,1
R15	1	2,3	0,5
R16	-0,04	-0,2	-1
R17	-0,04	0,6	0,5
R18	1,6	1	0,5
R19	-0,6	-1	-0,7
R20	-0,6	-0,6	-0,7
R21	-0,04	0,6	-0,7
R22	-2,3	0,6	-1
R23	-0,6	0,25	-0,1
R24	-1,2	0,6	-0,1
R25	0,5	1,1	-0,1
R26	1	-2	1,1
R27	-1,2	1	-1
R28	-0,04	-0,2	-0,1

All in all, we observed that there are only 3 raters who cannot be in the range in any of the ratings while 7 of them are in the range. We also observe that there are different rater decisions for each of the rating procedures. When raters are in the range for two or one of the ratings, they might not be in the range for the rest. There might be different factors affecting the raters while they are grading the papers. We know that these three papers can be placed in the rubric differently, in other words, they are the samples of good, average and poor papers. So, we can only assume that each paper makes raters give decisions in a different way.

4.2.1.2. Intraclass Correlation Coefficient Analysis

In the previous section, how close 28 raters can get to the benchmarks of 3 different papers were demonstrated. Table 2. provided information about individual discrepancy points but still, it is difficult to define the general rate of inter-rater reliability for all rating processes. At that point, ICC two-way mixed absolute agreement model was used to demonstrate the raters' interrater reliability. The values between 0.75 and 0.9 indicate good reliability Koo and Li (1995). As it is shown in Table 2 Intraclass Correlation is 0.99 which indicates excellent reliability. It means that raters gave remarkably close scores to each item, as a result, average measures are very high. (p.158).

Table 4. Intraclass Correlation results of raters in the last workshop

		95% Confidence Interval		F Test with True Value 0			
	Intraclass Correlation	Lower Bound	Upper Bound	Value	df1	df2	sig
Average Measures	0.996	0.984	0.1000	333.359	2.0	54	0.000

Two-way mixed effects model

These results show us that raters have excellent inter-rater reliability scores before they grade the papers of the final exam. ICC and Z-value calculations present the current situation of the raters in the last workshop. Inter rater reliability is depicted as individual and means scores of the raters.

4.2.2. Rater Correlations

Rater Correlations are depicted by ICC, a two-way random, absolute agreement model. Rater correlations of 28 raters were separately analyzed for pre and post-test results. A comparison of two different correlation results gained from the pre and post-test presents us the effectiveness of writing workshops.

4.2.2.1. Pre-Test Results

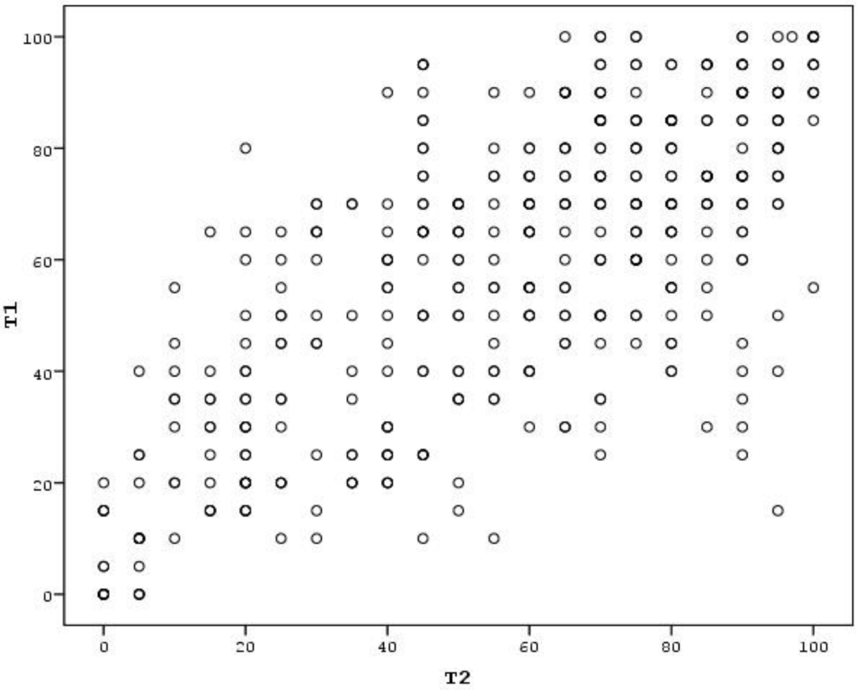
First, final exam results were collected from 421 pre-int students and the same set of 28 raters graded the papers in the pre and post-tests of the research. The pre-test was done after raters were introduced with the new rubrics. Papers were randomly distributed to the raters for the first and second checks. Rater performances were analyzed by using ICC calculations. As Table 3. shows ICC average measures of the raters is 0.83. This is the interrater reliability of the raters before the writing workshops started. This shows a good inter reliability rate Koo and Li (1995). Although this was a good rater reliability rate, the discrepancy point on the rubric was 2 for two different raters so, the school authority decided on having writing workshops since there were more than 200 papers to be checked for the third time and there were a lot of questions asked by the raters. The aim of the school authority was to decrease the number of the exam papers sent to the third check and make the discrepancies lower between two raters. The scatter plot presents the inter-rater reliability rate. (p.158).

Table 5. Intraclass Correlation results of raters in the pre-test

		95% Confidence Interval		**F Test with True Value 0**			
	Intraclass Correlation	**Lower Bound**	**Upper Bound**	**Value**	**df1**	**df2**	**sig**
Average Measures	0.830	0.795	0.860	5.889	420	420	0.000

Two-way random effects model

Figure 1. Scatter plot for pre-test results



4.2.2.2. Post- Test Results

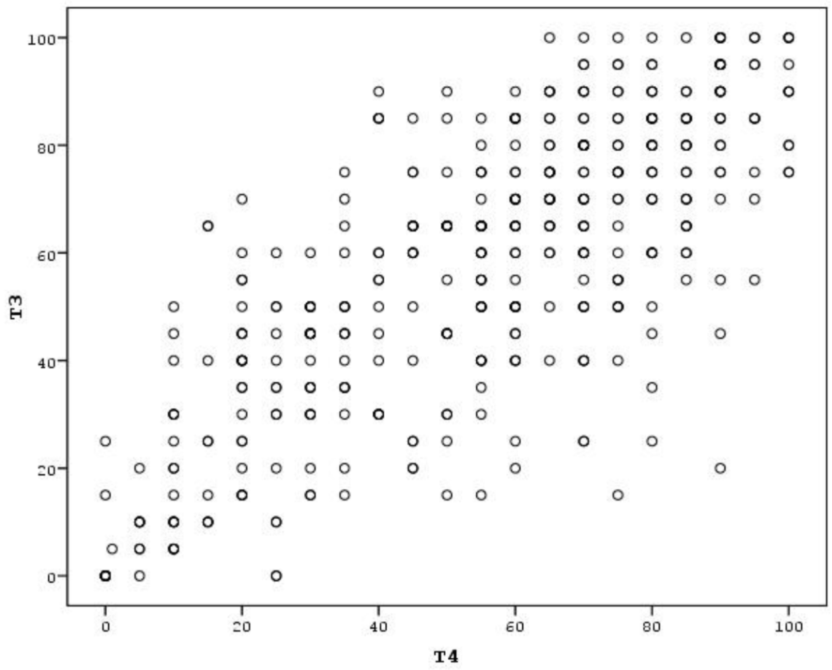
Second, final exam results were collected from the same 421 int students and the same set of but randomly distributed raters graded exam papers for the first and second checks. The collected data was analyzed again with the same calculation method of ICC and the average measures rate changed from 0.83 to 0.86. The new correlation result proves that the interrater reliability of the raters increased which means they improved at giving similar scores to the exam papers. But still, they stayed in the range of good reliability which is defined as values between 0.75 and 0.9 Koo and Li (1995). A scatter plot presents the inter-rater reliability rate. (p.158).

Table 6. Intraclass Correlation results of raters in post-test

	Intraclass Correlation	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound	F Test with True Value 0 Value	df1	df2	Sig
Average Measures	0.865	0.836	0.889	7.513	420	420	0.000

Two-way random effects model

Figure 2. Scatter plot for post-test results



4.2.3. Interviews

4 workshop conductors basically gave answers to 7 questions and some extended questions. Interviews were made to get an answer to the last research questions. Producer related questions were asked to the conductors to see and exemplify what really happened in the sessions. Conductors were asked to generally speak on the most difficult parts of the session and with the help of extended questions, the effectiveness of the writing workshops tried to be understood from the conductor's perspective. All in all, interview results also support the idea that writing workshops were effective maybe not totally but partially for the raters.

1. What are the problems you encounter during the workshops?

All workshop conductors believe that teachers were unwilling to attend to the session since it was an extra workload and a time-consuming rating process due to the rubric. Conductor 3 also mentioned there was resistance from the teachers to cooperate in groups and discuss the problems since everybody has different perceptions of the rating process. Conductor 4 expressed that some teachers tended to severity, and they insisted on giving fewer scores with the help of extended questions they talked more on severe raters. They mentioned that severe raters talked more in groups to express their ideas and they got the chance to share their ideas in different groups, as a result in the last workshop they became less severe for some of the papers but not for all. Conductor 4 also focused on the leniency of the raters. Some raters were more lenient compared to others. By giving them a chance of having different group discussions on the rating process, workshops conductors tried to eliminate the problems.

Conductor 2 remarked that putting raters into large groups of people during the workshops was not a good idea. Letting them have discussions in small but different groups was more effective. Conductor 1 focused on the point that most of the raters were silent in the first session but this attitude changed in the other sessions. They became more talkative and asked more questions related to the rating process. These questions sometimes caused longer discussions and getting lost in conversations.

2. *Have you realized any repetitive problems in each session (writing workshops)?*

If yes, please specify them.

Conductor 4 mentioned that some raters found the criteria confusing and they had a hard time finding where to take points off. Being persistent was mentioned by each workshop conductor. This was the main thing that they observed as a repetitive problem but with extended questions, it came out that those raters started to change their attitudes so, workshop conductors believe that if they had a longer time for training, most of the things would change more. Conductor 1 remarked that raters had a difficult time finding common ground. First, everybody was mentioning their ideas, but they did not tend to have an overall group score after they discussed it in their groups, but later raters got the point and they understood what they needed. Finding common ground was the purpose of the sessions. When they discussed more in groups, the problems became clearer and raters learned to look from a different point of view.

3. *How did you struggle with these problems? Could you please give some*

examples? Do you believe that the solutions you came up with worked?

All workshop conductors believe that rater discrepancy points will be less in the next rating processes. They believe that the best thing that can be done for better results is more workshops since they got less 3rd check papers after the sessions and conductor 4 explains the situation:

"In each training session, as the trainers, we explained the items in detail and brought some sample writing papers and scored the papers together by explaining each descriptor. It seemed that it helped in the training session, however, after training it was obvious that some of the raters still went on the severity in grades and there were constantly gaps between the scores they gave ".

Conductor 3 highlights the importance of workshops by mentioning:

"Repeating sessions of standardization workshops until desired results achieved. I believe that the objectives were achieved for the most part ".

Conductor 1 believes that they found solutions for most of the situations and they found the workshops effective. Raters started to participate more actively in group discussions.

4. What were the questions mostly asked by the teachers during the workshops?
Can you give any examples?

Conductors 1 and 3 stated that raters were mostly unfamiliar with the idea of analytical rubrics and had more of a holistic approach to scoring. It was initially hard for the teachers to see and evaluate a separate criterion on its own. Conductor 4 indicated that students who wrote less than the word limit created problems for raters, and they tended to give too low points. However, this issue had to be dealt with in the content section and they needed to take the points off only from that section. They overly penalize students for that. This was mostly asked by the raters and another question was that if the topic was totally irrelevant from the task. Conductor 2 gives examples to common questions by stating that:

“How to use and evaluate the students using rubrics was the most asked question. Followed by questions to clarify the objectives in rubrics. For example: “It is said in that box that student used a variety of conjunction, so how many are we expecting?”, or “It is said in that box that student gave examples to the topics, do we expect examples for every scenario?”

On the other hand, all conductors received more questions related to structure and grammar. Instructors tended to give more importance to correctness rather than fluency and or content.

5. Do you believe that writing workshops have eliminated the problems teachers face during the rating process?

Conductors 3 and 1 clarify the question by mentioning that the standardization workshops provided a new perspective for teachers who marked the papers and to the greatest extent provided a shared understanding of assessment criteria. Conductor 4 stated that Most of the problems were eliminated. Most of the teachers were new to the concept and they did not know how to grade the students’ product with a rubric. At least, they got familiar with the idea of the use of an analytical rubric. It made the grades more reliable and valid and conductor 2 indicated that it was not eliminated but helped them tremendously in terms of understanding where the middle ground is and how to adjust themselves to the standards set by the rubric. Also, seeing the personal differences and reasons helped them relax about the grading process. It could be understood that workshops did not eliminate all the problems but still it made a point in raters’ minds to think about and improve their rating procedure by knowing more about the criteria.

6. Have you seen any resistance to this new process (writing workshops) from the teachers?

All workshop conductors stated that at first there was resistance from the raters to the new process. Some raters found this new process too complicated. Because they had to refer to the rubric for each paper. Moreover, for each paper, they needed to check the rubric sections continuously. Conductor 2 clarifies this issue by giving the example of the instructor's suggestions. At first, they received many suggestions from instructors not to use rubrics but as they continued doing the workshops, instructors adapted and realized how it makes the grading process as equal as it can be for the students. However, there were still some instructors who did not want to change their severity or leniency but at least they believed that students need an average result of two different raters. Conductor 3 described resistance to this new process as instructors were unwilling and reluctant, and initially saw this as a threat or challenge to their teaching competencies. Conductor 1 also remarked that instructors' approach was not positive at first to the new process, but by the end, they started to hear some different points of view. Some instructors started to talk about the effectiveness of the rubrics and workshops where they could ask a lot of questions but there were also some instructors who wanted to continue with their own style and did not believe the effectiveness of the workshops. Conductor 4 remarked that some raters tended to be severe in grading, they wanted to give fewer points than the official scores. They mostly focused on accuracy rather than fluency. On the other hand, they wanted to see perfect papers to give good scores.

7. *Do you believe that writing workshops were useful for the rating process?*

All conductors mentioned that they had some hesitations about the effectiveness because they obviously observed the unwillingness of the instructors. They saw that instructors tended to continue with their style which was hard to change at once. But later, workshop conductors stated that they had many questions related to the process thanks to the discussion in groups. Instructors started to take part in the discussions more willingly, they became more open and they started to think aloud with their group members. These discussions went on for hours. Moreover, all workshop conductors mentioned that they received questions related to papers and rubrics during the lesson and lunch breaks, which was not expected. Instructors were involved in the new style as a natural process even if some of them were initially unwilling. They found themselves in natural discussions and small talks about the new assessment process. So, this was a smooth pass from the old style to the new one. Workshop conductor 2 stated:

“*there were no set criteria to evaluate students with, so it was very subjective. People were unsure about how many points to add or subtract for general categories such as grammar, content, or spelling. Having an order provided a more stable grading process all around”.*

Conductor 3 emphasized on the decreasing number of 3rd check papers since they were responsible for dealing with the papers which have high discrepancy points. This issue was explained as;

“There is a tendency towards a more coherent and synchronous marking throughout the institution amongst all graders which can be observed by the declining number of discrepancies”.

CHAPTER 5

5. DISCUSSIONS

5.1. PRESENTATION

Discussions were made to evaluate the results of the study. Similar studies were discussed, results were compared, and limitations were presented in this section. Finally, suggestions were made for future studies.

5.2. DISCUSSIONS

This study investigated the effectiveness of writing workshops for raters in the assessment process. As many researchers have suggested in their studies it was expected that rater training would improve interrater reliability and it would decrease discrepancy points between two raters. Yeloğlu (2013) highlighted the importance of rater training in their study after they presented the perceptions of raters on new rubric usage and borderline papers. As it is suggested in many other studies listed above, they also mentioned the necessity of interrater reliability research to investigate the issue more. Writing assessment is a process that should be dealt with in detail since there are some human factors that cannot be eliminated. In many studies, as listed in the literature review part, these factors have been discussed. Rater experience, content, rater beliefs, rubric and student background have some effects on the rating process. Various interpretations might be made by the raters even if they use the same rubric. As it is observed in this study, workshop conductors stated that instructors asked many questions related to rubrics, each scale on the rubric and the papers since they all have unexampled specialties. Evaluation of student papers is a subjective process so, making this process objective is not as easy as it is considered. In many studies, it has been observed that usage of rubrics is not enough to eliminate the problems raters face as a result of this rater trainings are highly suggested to get closer points from two different raters. (p.377). According to Engemann and Gallagher (2006), the best way of lowering discrepancies is discussions. It helps raters to come to an agreement and understand the obscure parts on the rubric. By getting more examples and asking questions to other raters, it is possible to gain a more objective point of view. Raters can get standardized with the help of discussions during workshops or later at any time. That study also proved that two training sessions would be enough to get better results in rater agreement. p (41).

The aim of this study was to observe the difference in rater discrepancies in a short time in training. As Weigle (2004) indicated that the evaluation of writing papers improves with the help of training. It is possible to get better results from raters after they are trained. This training process equals to two hours training process in many workshops. It is also noted that four hours of training would be enough to make a change in the raters' grading process Brown (2004). (p.117). In this study, approximately 8 hours were spent on all the writing workshops. Five hours of it was for benchmark rating practice and discussions and the other three hours were spent discussing and understanding rubrics, and what students can achieve in each level (elementary, pre-int…). The time spent on the training was enough to observe a difference between pre and post-tests. Raters had enough time to discuss and learn more about new rubrics and they had also enough time for practice. Benchmarks were used during the workshops. They assist the rater training process Greer (2013). Every instructor had the chance of rating the same paper with the others so that they could negotiate the evaluation process first within their groups, then with the others. In this study, it was possible to investigate how close raters got to the benchmarks in the last workshop. We could have an idea about the current situation of the raters. In other words, the closeness of 28 raters to the benchmarks was presented. (p.4).

Fahim and Bijani (2011) used the same method in their study to define the closeness of the raters to the benchmarks. They also defined pre and post-test differences by using z-scores. In this study, pre and post-tests were analyzed by using ICC analyzations, but the z-scores were just used to define how close raters got to the benchmarks. Fahim and Bijani (2011) presented in their results that out of 12 raters they had still 3 lenient and 2 severe raters after the training. In this study, similar results were gained out of 28 instructors, 3 of them were severe or lenient in some ratings. So, even after rater training, there are still some raters who grade the papers very differently which may mean that there is a still need of having more workshops. This may lead to rating improvement in the future. In another study, these instructors might be interviewed to understand what they think and believe, or they may be asked to speak about their methods. But in this study, it was not possible to interview the raters since writing assessment with rubrics was a new process and the school authority did not want to disturb any instructors or make them feel offended, they did not even inform those instructors that they were severe or lenient raters. Their aim was to solve most of the problems with discussions and group works. (p.5).

The same data was also analyzed by using ICC which helps us to understand the interrater reliability of all the instructors. The ICC result of raters in the last workshop was 0.99. This is a remarkably high level of agreement for the raters. Wolfe (2009) remarked in their study that 0.70 is a good rater agreement. So, what is achieved in the last workshop was a perfect agreement rate for the instructors. (p.13). Brown (2004) found 0.80 in their ICC results. They worked on 114 instructors across 3 holistically designed tasks. They worked on a six-point rubric which means there are six different bands to evaluate the product. It is also mentioned that a rubric can affect the ICC results. If a rubric has fewer bands on it, it means that there are not many options to think about and it might be easy to decide on so that ICC results can be higher. (p.105-106). In this research, the rubric has five bands on it, and it made instructors think more on it compared to holistic ones. They also mentioned that if the scores are higher than 0.70 they are regarded as enough for rater reliability. For interrater reliability research, Cohen Kappa is another mostly used method for analyses of rater reliability. But Kappa value can be used for nominal data. It is defined as “1” or “0” in SPSS. In other words, if the pass or fail papers are analyzed; fail papers can be defined by “0” while pass papers are defined with “1”. Similar researches can be done for borderline papers Stempler (2004). Kappa value should be at least 70% for good interrater reliability.

However, Greer (2013) used Kappa value for their research’s statistics and they found out that Kappa value changes from 0.330 to 0.235 in post-test results. (p.4) They are both considered as fair agreement but after rater training, interrater reliability decreases a little bit. This decrease is not significant, but it can be inferred that rater training does not change the results.

In this study, because of the continuous data, Cohen Kappa statistics could not be used. The actual grades of the students were analyzed so Kappa value was not appropriate for this type of statistic. Instead, ICC was used which gives better results compared to the Pearson correlation coefficient, paired t-test, or Bland Altman plot. These tests measure either correlation or reliability but ICC measures both Koo and Li (2016). (p.156).

ICC was also used to define the differences in pre and post-tests of this study. Besides comparing results of pre and post-tests, there was a chance to compare pre-post and while (last workshop) results in this study. When ICC results of pre and post-tests are compared after the treatment, the rater agreement value changed from 0.83 to 0.86. Both are regarded as fair agreement, but it is obvious that rater reliability increased. A similar study done by Cho (2003) indicates that ICC results changed from 0.77 to 0.95 after raters underwent four workshops and as a result of this, ICC value increased. In this study, the rater agreement-rate of instructors was 0.99 which can be also regarded as high. So, the ICC results of this study support the findings of the studies mentioned here. Rater- agreements increased after training. These studies showed similar results.

Some studies benefited from ICC measures and they supported or justified their results. ICC was again used to investigate rater- agreements in those studies, although their main research questions were different. For example, Dunsmuir (2015) investigated the reliability of test measuring items. 4 different raters used the same items to grade different exam papers at different times. They checked if test measuring items give similar results in different conditions. This reliability was again checked by using ICC measures. Rater agreement was found high 0,97. p (15). Park and Stapleton (2003) conducted another study to see the correlation between L1 internal voice and success of L2 academic writings. Exam papers were rated by different instructors to be sure about the reliability of exam results. In other words, also in this study, they looked for higher ICC results which equate to a better rater agreement. The correlation could only be investigated if the exam results were reliable. p (260). Neumann (2014) had to check error codes on papers marked by instructors to investigate the effectiveness of written feedback given to students. Rater agreement was also necessary for this research. ICC was used to check agreement value. ICC ranged from 0.88 to 0.96 and it showed a high agreement rate. p (89). Sweedler-Brown (1993) also used ICC to present the correlations between analytic and holistic scorings and ICC results showed that they were correlated. As it is presented, ICC was used to define the agreement rate of different raters. More objective results were gained by getting different raters' scores and correlations of raters were checked to gain more reliable results.

Continuous rater training might increase the rater agreement level so future studies should focus on continuous rater training and they should evaluate the results after repetitive training sessions. Studies might be done to observe rater behaviors. In both pre and post-tests, each instructor checked approximately two sets of different 14 papers for 1st and 2nd checks. So, in total, one instructor checked about 28 papers for 1st and 2nd checks. But in the last workshop, each rater checked 3 papers and their agreement rate was 0.99. If we compare post-test results with while (last workshop) it is obvious that the agreement rate is lower in the post-test 0.86. Normally, we might expect to see while and post-test results closer than actual values but here it is obvious that instructors cannot rate the papers as they do in the workshop. There might be many grounds for it. First, the number of the paper's raters check might be a matter and raters may not focus on the rating process as they do in the workshops. The atmosphere might also be affecting raters to focus on rubrics more and refer to them repetitively. Another point is that instructors are alone while they are rating the papers so what they really think and how they decide is an overly complicated process. Lumley (2002) defines the key element of this complicated process as the raters. The results of this study also support the idea that raters are the most complex links of the evaluation cycle. There should be more researches made to investigate the rater's rating process so that more information will be gained. (p.267).

Although the results for rater agreement are not remarkably similar for the post-test and last workshop, there is still an improvement in rater agreement results when we compare the pre and post-tests. There is a positive change in the results and workload of the school that was caused by the 3rd checks decrease. Less number of papers were sent to a third check. Benchmark samples and discussions in groups helped the raters a lot. Raters became more active participants and they asked many questions related to this new process. Rater agreement came to a better point. So, this research supports the results of Harsch and Martin (2012). (p.239). Discussions and questions of raters were found really useful. The interview results of this study support that instructors found themselves in a natural process, as a result of this they started to ask many questions. They tried to find solutions, and this led them to discuss more the unclear parts of the evaluation. They started to understand each other and by talking more on the topic, they reached a point of understanding. However, there is a still need for continuous rater training. Moreover, rater training might be done just before the exam papers are read since this study proves that raters are more standard during the workshops. So, just a reminder of the rubric and rating a few papers together before grading exam papers individually might be useful. All workshop conductors stated that instructors focused more on structure rather than fluency or context. This supports the findings of Brown (2004). They also mentioned the instructor's priority was punctuation and sentence structure. (p.117). A previous study conducted by Sweedler-Brown (1993) also expresses that writing instructors focus more on structures rather than the other features of writing. They prioritize grammatical errors more. p (3). In this study, workshop conductors mentioned that they received more questions on grammatical bands on the rubric.

All in all, it is known that desired results cannot be gained in many interrater reliability researches Jonsson & Svingby (2007). Rater agreement shows differences depending on many other factors that should be investigated more but, in this study, workshops were effective, and it led to a change in the results. There was an improvement in rater agreement values so, we should continue rater training no matter what the result is as Huot (1990) suggests.

5.3. SUGGESTIONS FOR FURTHER STUDIES

It is certain that a more detailed study should be done to analyse rater behaviours since there are many different factors affecting the rating process.

Workshops can be recorded and interviews with the raters can be done to examine this process in more detail. It was found in this study that rater performances in the workshops and real exams differ so research can be designed to focus more on performances of the raters in real exams. In the workshops, raters can be given more scripts to rate and their performances can be observed in the first and other papers. And the accuracy of the scores can be checked through ratings. Longitudinal studies might be done to observe rater behaviours. Raters can be asked to rate the same set of papers at different times to observe intra-rater reliability.

REFERENCES

1. Books

Alderson, J. C., Clapham, C., & Wall, D. (1995). Language test construction and evaluation. Cambridge,UK: Cambridge University Press.

Bachman, L. F. and Palmer, A. S. (1996). Language testing in practice. Oxford: Oxford University Press. Cambridge (2016). Assessment Series. Cambridge: Ucles.

Ethridge, D.E. (2004) "Research Methodology in Applied Economics" John Wiley & Sons,

Fraenkel, Jack R, Norman E. Wallen, Helen H. Hyun (2012). How to Design and Evaluate Research in Education. New York: The McGraw-Hill.

Guba, E.G. and Lincoln, Y.S. (1985) Effective Evaluation. Jossey-Bass Publishers, San Francisco

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

Huot, B. (2002). (Re)Articulating Writing Assessment for Teaching and Learning. All USU Press Publications**.**

Johnson, R., Penny, J., & Johnson, C. (2000, April). A conceptual framework for score resolution in the rating of performance assessments: The union of validity and reliability. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA**.**

Martin-Kniep, G. O. (2000). Becoming a better teacher: Eight innovations that work. Alexandria: 35. Association for Supervision & Curriculum Development

McAlpine, M. (2002). Principles of assessment. Glasgow: University of Glasgow, Robert Clark Centre for Technological Education.

McMillan, J. H. (Ed.). (2007). Formative classroom assessment: Theory into practice. New York: Teachers College Press

Stoynoff, S., & Chapelle, C. A. (Eds.). (2005). ESOL tests and testing: A resource for teachers and program administrators. Alexandria, VA: Teachers of English to Speakers of Other Languages

Turgut, M. F. (1990). Eğitimde ölçme ve değerlendirme metotları. Ankara: Saydam Matbaacılık.

Weigle, S. C. (2002). Assessing Writing. Cambridge, UK: Cambridge University Press.

Weiss, C. H. (1972). Evaluation Research: Methods for Assessing Program Effectiveness. Englewood Cliffs, NJ: Prentice- Hall

Wolcott, W. (with Legg, S.) (1998). An overview of writing assessment: Theory, research, and practice. Urbana, IL: National Council of Teachers of English.

2. Articles

Andrade, H. G. (2000). Using rubrics to promote thinking and learning. Educational Leadership, 57(5), 13–18.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? System, 29, 371-383.

Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. Language Testing, 27(4), 515-535.

Barkaoui, K. (2011). Do ESL essay raters' evaluation criteria change with experience? A mixed methods, crosssectional study. TESOL Quarterly, 44 , 31057

Becker, A. (2010,2011). Examining Rubrics Used to Measure Writing Performance in U.S. Intensive English Programs. The CATESOL Journal 22.1 113-130

Beyreli, L., & Arı, G. (2009). The Use of Analytic Rubric in the Assessment of Writing Performance-Inter-Rater Concordance Study. Kuram ve Uygulamada Eğitim Bilimleri / Educational Sciences: Theory & Practice 9 (1) • Winter 2009 • 105-125

Broad, B. (2000). Pulling Your Hair Out: Crises of Standardization in Communal Writing Assessment. Research in the Teaching of English, 35(2), 213-260.

Brown, G.T.L, Glasswell, K. & Harland D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. Assessing Writing 9 105–121

Brualdi, A (1998). Implementing performance assessment in the classroom. Practical Assessment and Research evaluation. Vol.6, number 2, 1-3

Cho, Y. (2003). Assessing writing: Are we bound by only one method?. Assessing Writing. 8.165- 191.

Crehan, K. D. (1997). A discussion of analytic scoring for writing performance assessments. Annual Meeting of the Arizona Educational Research Association. (ERIC Document Reproduction Service No. ED 414336).

Cumming, A. (1990). Expertise in evaluating second language compositions. Language Testing, 7 (1), 31–51

Dunsmuir, S., Kyriacou, M., Batuwitage, S., Hinson, E., Ingram V. & O'Sullivan, S. (2015). An evaluation of the Writing Assessment Measure (WAM) for children's narrative writing. Assessing Writing 8. 1-18.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. Language Testing 25 (2) 155–185.

Engemann, J. F., & Gallagher, T. (2006). The Conundrum of Classroom Writing Assessment. Brock Education. 15(2), 33-44.

Fahim, M. & Bijani, H. (2011). The effects of rater training on raters' severity and bias in second language writing assessment. Iranian Journal of Language Testing, 1(1), 1-16.

Finson, K. D. (1998). Rubrics and their use in inclusive science. Intervention in School and Clinic, 34 (2), 79–88.

Fleiss, J. L., & Cohen, J. (1973). The Equıvalence Of Weıghted Kappa And The Intraclass Correlatıon Coeffıcıent As Measures Of Reliability. Educational and Psychological Measurement 33, 613-619.

Gallagher, C. W. (2009). What Do WPAs Need to Know about Writing Assessment? An Immodest Proposal. WPA: Writing Program Administration, Volume 33, Numbers 1-2, Fall/Winter

Guillot, C. P. & Delgado, J. Z. (2017). Scoring validity: Designing and implementing an online training module for rater training. SITE, Austin, TX, United States, March 5-9. 1334-1338.

Harsch, C. & Martin, G. (2012). Adapting CEF descriptors for rating purposes:Validation by a combined rater training and scale revision approach. Assessing Writing, 17, 228-250

Huot, B.A (1990): Reliability, validity and holistic scoring: what we know and what we need to know. College Composition and Communication 41, 201-13.

Jonsson, A. & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review 2 130–144

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. Language Testing, 26(2), 275-304

Kohn, A. (2006). The trouble with rubrics. English Journal, 95 (4), 12–15.

Koo, T. K & Li, M. Y. (2015). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. Journal of Chiropractic Medicine (2016) 15, 155–163.

Kutlu, O., Bilican, Safiye & Yildirim, Ozen. (2010). A Study on the Primary School Teachers'Attitudes Towards Rubrics with Reference to Different Variables. Procedia Social andBehaviorial Sciences, 2(2010), 5398-5402.

Kuisma, R. (1999) Criterion referenced marking of written assignments, Assessment and Evaluation in Higher Education, 24 (1), pp. 27–39.

Li, J. & Lindsey, P. (2015). Understanding variations between student and teacherapplication of rubrics. Assessing Writing 26 (2015) 67–79.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. Journal of Applied Psychology, 86, 255–264.

Linacre, J. M. (2002). Judge ratings with forced agreement. Rasch Measurement Transactions, 16(1), 857-858.

Little, D. (2007). The Common European Framework of Reference for Languages: Perspectives on the Making of Supranational Language Education Policy. The Modern Language Journal, Vol. 91, No. 4 (Winter, 2007), pp. 645-655.

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.

Moon, T.R & Hughes, R. H (2002). Training and Scoring Issues Involved in Large-ScaleWriting Assessments. Educational Measurement: Issues and Practice. Summer 15-19.

Moskal,B. M. & Leydens, J. A. (2000). Scoring Rubric Development: Validity and Reliability. Practical Assessment, Research & Evaluation. Volume 7, Number 10, November 1-6.

Neumann, H. (2014). Teacher assessment of grammatical ability in second language academic writing: A case study. Journal of Second Language Writing 24. 83-107.

Nevo, D. (1985) Experts' Opinion: A Powerful Evaluation Tool. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago

Park, R. H., & Stapleton, P. (2003). Questioning the importance of individualized voice in undergraduate L2 argumentative writing: An empirical study with pedagogical implications. Journal of Second Language Writing. 12. 245-265.

Popham, W. J. (1997). What's Wrong--and What's Right--with Rubrics. Educational Leadership, 55(2), 72-75.

Popham, W. J. (2003). Test Better Teach Better. The instructional role of assessment. 84-88.

Pufpaff, L. A., Clarke, L., & Jones, R. E. (2015). The effects of rater training on interrater agreement. MidWestern Educational Researcher, 27, 117141.

Qasim, A & Qasim,Z. (2015). Using Rubrics to Assess Writing: Pros and Cons in PakistaniTeachers' Opinions. Journal of Literature, Languages and Linguistics www.iiste.orgISSN 2422-8435 An International Peer-reviewed Journal Vol.16, 51-58.

Reddy, Y. M., & Andrade, H. (2010). A review of rubric use in higher education. Assessment & Evaluation in Higher Education, 35(4), 435–448.

Rezaei, A. R. & Lovorn, M. (2010) Reliability and validity of rubrics for assessment through writing. Assessing Writing 15 (2010) 18–39.

Smith, D (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In Brindley, G. (ed.) Studies in Immigrant English Language Assessment. Vol. 1. Sydney: Macquarie University.159-189.

Song, B. & Caruso, I. (1996). Do English and ESL Faculty Differ in Evaluating the Essays of Native English-Speaking and ESL Students? Journal Of Second Language Writing, 5 (2), 163-182.

Spandel, V. (2006). In Defense of Rubrics. English Journal, 96(1), 19-22.

Starch, Daniel, and Edward C. Elliott. (1912). Reliability of the Grading of High-School Work in English. School Review 20: 442–57.

Stemler, S. E. (2004). A Comparison of Consensus, Consistency, and Measurement Approaches to Estimating Interrater Reliability. Practical Assessment, Research & Evaluation. Volume 9, Number 4, March 1-11.

Stoddart, T., Abrams, R., Gasper, E., & Canaday, D. (2000). Concept maps as assessment in science inquiry learning—A report of methodology._International Journal of Science Education, 22, 1221–1246.

Stratman, J., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: Issues for research. In: P. Smagorinsky (Ed.), Speaking about writing: Reflections on research methodology. Thousand Oaks, CA: Sage.89-111.

Sweedler-Brown, C. O. (1993). ESL Essay Evaluation: The Influence of Sentence-Level and Rhetorical Features. Journal Of Second Language Writing, 2 (1), 3-17.

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.

Turley, E. D., & Gallagher, C. W. (2008). On the "Uses" of Rubrics: Reframing the Great Rubric Debate.English Journal, 97(4), 87-92.

Way, W. D, Vickers, D, Nichols, P. (2008). Effects of Different Training and Scoring Approaches on Human Constructed Response Scoring. Pearson. annual meeting of theNational Council on Measurement in Education,New York City, April 1-19.

Weigle, S.C. (1994). Effects of training on raters of ESL compositions. Language Testing, 11, 2: 197-223.

Weigle, S. C. (1998). Using FACETS to model rater training effects. Language Testing, 15(2), 263-287.

Weigle, S. C. (2007). Teaching writing teachers about assessment. Journal of Second Language Writing 16 (2007) 194–209.

Wolfe, E. W, Mathews, S. & Vickers, D. (2009). A Comparison of Training & Scoring in Distributed& Regional Contexts—Writing. Pearson July.1-24.

Yancey, K.B. (1999). "Looking back as we look forward: Historicizing writing assessment." College Composition and Communication, 50, 483-503.

Yang, W., Lu, X. & Weigle, S.C. (2015). Different topics, different discourse: Relationships among writing topic, measures of syntactic complexity, and judgments of writing quality. Journal ofSecondLanguageWriting28(2015)53–67.

Yurekli, A., & Ustunluoglu, E. (2007). Towards student involvement in essay assessment. Essays in Education, 22, 55-64.

3.*Electronic Sources*

Barrett, P. (2001, March). Assessing the reliability of rating data. Retrieved June 16, 2003, from http://www.liv.ac.uk/~pbarrett/rater.pdf

Lindsey, P., & Crusan, D. (2011, December 21). How faculty attitudes and expectations toward student nationality affect writing assessment. Across the Disciplines, 8(4). Retrieved January 6, 2015, from http://wac.colostate.edu/atd/ell/lindsey-crusan.cfm

MarkBook, retrieved March 17, 2011, from http://www.asyluminc.com/mb_manual/MK int3.htm

Schafer, L. (2004). Rubric. Retrieved February 9, 2015, from http://www.etc.edu.cn/eet/articles/rubrics/index.htm

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on consistency of performance criteria across scale levels. Practical Assessment, Research & Evaluation, 9(2), Retrieved February 16, 2004 from http://PAREonline.net/getvn.asp?v=9&n=2.

4.Report

Glasswell, K., Parr, J. & Aikman, M. (2001). Development of the asTTle Writing Assessment Rubrics for Scoring Extended Writing Tasks. Technical Report 6, Project asTTle, University of Auckland 1-27.

5.Thesis

Greer, Brittney, "Assisting Novice Raters in Addressing the In-Between Scores When Rating Writing" (2013). All Theses and Dissertations. 4066.

Appendix A

School of Foreign Languages

WRITING EVALUATION RUBRIC

YAZMA SINAVI DEĞERLENDIRME ÖLÇEĞI

Score by ticking the box in the level's descriptor for each criterion.

Student's Name: ______________________ **Track:** 1□ 2□ 3□ 4□ 5□

Student's No.: ______________________ **Group:** ______________________

	EXCELLENT (5)	GOOD (4)	FAIR (3)	NEEDS IMPROVEMEMENT(2)	POOR (1)	0
CONTENT	□ All content is relevant to the task. Target reader is fully informed	□ Performance shares features of Bands 3 and 5.	□ Minor irrelevances and/or omissions may be present. Target reader is on the whole informed.	□ Performance shares features of Bands 1 and 3.	□ Irrelevances and misinterpretation of task may be present. Target reader is minimally informed.	□ Content is totally irrelevant. Target reader is not informed.
ORGANIZATION	□ Text is well organized and coherent, using a variety of cohesive devices and organizational patterns to generally good effect.	□ Performance shares features of Bands 3 and 5	□ Text is generally well organized and coherent, using a variety of linking words and cohesive devices	□ Performance shares features of Bands 1 and 3.	□ Text is connected and coherent, using basic linking words and a limited number of cohesive devices.	□ Performance below Band 1
LANGUAGE	□Uses a range of vocabulary, including less common lexis, appropriately. Uses a range of simple and complex grammatical forms with control and flexilibity	□Performance shares features of Bands 3 and 5	□ Uses a range of everyday vocabulary appropriately, with occasional inappropriate use of less common lexis. Uses a range of simple and some complex grammatical forms with a good decree of control	□ Performance shares features of Bands 1 and 3.	□ Uses everyday vocabulary generally appropriately, while occasionally overusing certain lexis. Uses simple grammatical forms with a good degree of control.	□ Performance below Band 1
MECHANICS	□ Occasional errors may be present but do not impede communication	□Performance shares features of Bands 3 and 5	□ Errors do not impede communication.	□ Performance shares features of Bands 1 and 3.	□ While errors are noticeable, meaning can still be determined.	□ Performance below Band 1
FINAL SCORE *Adjust the score to 100-point scale by multiplying it by 5, after adding the score for each criterion up.*						