ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 12, No. 1, pp. 55-64, January 2022
DOI: https://doi.org/10.17507/tpls.1201.07

A Comparative Study of EFL Instructors' Essay Rating: Holistic Versus Analytic Approaches at a Tertiary Institution in Saudi Arabia

Shifa Alotibi
English Language Institute, King Abdulaziz University, Saudi Arabia

Abdullah Alshakhi
English Language Institute, King Abdulaziz University, Saudi Arabia

***Abstract*—This study seeks to explore the factors that influence EFL instructors' rating decisions while using holistic and analytic rubrics. Few studies have been conducted on the factors that influence the rating practices of EFL instructors, specifically, in the Saudi EFL context. This study addresses this gap and contributes more broadly to the understanding of the interplay between EFL instructors and the use of holistic and analytic rubrics. The data were collected in a Saudi university at a preparatory year program (PYP). This study utilizes semi-structured interviews with eleven EFL instructors from different nationalities. Guided by the critical language testing as a theoretical framework and with qualitative analysis, the study reveals that critical language testing can minimize the negative consequences of writing assessment done by graders; however, students' low English proficiency, time constraints, heavy workload can negatively affect the rating practices. Finally, several pedagogical implications, insights, and recommendations for future research are put forward in the conclusion.**

***Index Terms*—critical language testing, EFL writing, holistic and analytic rubrics, Saudi EFL context**

I. Introduction

Assessing students' writing is considered a complex, however, an essential process of language performance assessment (Barkaoui, 2008; Connor-Linton, 1995; Huang, 2012, as cited in Han & Huang, 2017). Research shows that the variability of ESL/EFL students' writing scores may be natural or stem from various factors such as age, proficiency level, first language (L1), fatigue, and cultural impacts on writing assessment practices. Furthermore, scores variability might be caused by raters, since they are central to students' writing performance assessment. Raters' different backgrounds, experience, or understanding of rating criteria may lead to different assessment of EFL essays (Barkaoui, 2010). Also, raters' reliability plays a crucial role in making the decision while correcting students' or test takers' writings (Kayapinar, 2014). As Bacha (2001) discussed to obtain valid rating scores, it is important to adopt the best evaluation instrument. She also commented that choosing accurate essay evaluation criteria has been considered problematic in many EFL/ESL programs due to several factors. Thus, it is important for teachers to be aware of different writing assessment tools they use or those they should adopt (Bacha, 2001). In addition, raters' rating should reflect students' ability and raters should avoid personal bias (Schaefer, 2008).

II. Theoritical Framework

The Critical Language Testing

The theoretical framework that provides a foundation for this current paper is Shohamy's (1998) CLT. This theory considers testing as a non-neutral action with social, ideological, and political plans that frame the lives of the participants, i.e., teachers and students. CLT refers to tests' uses and consequences in education and society. The basic tenet of this framework is that it emerged from the realization that tests are powerful tools which may lead to unintended consequences. In CLT, test takers are viewed as political subjects and language tests as success tools hidden in educational, political, and cultural fields. It was established based on the tests' relationship to social, political, and educational contexts. Tests are seen as benchmarks for acceptance and rejection that control other educational tools such as textbooks, curricula, and teaching methodologies. In addition, Shohamy (1998) points out the misuse of tests and their impact that goes beyond learning and teaching into political and educational domains.

As CLT calls for fair and equal consequences of tests, they must be valid and have positive impact through the use of appropriate assessment tools and rubrics. Validity is a dominant feature of language testing. It is the meaning and interpretation of test scores that indicate the overall evaluative judgment regarding the appropriateness of a test. In addition, it is an increasingly important area in language testing and assessment as it sheds light on test developers' responsibilities with respect to the uses of tests, the relevance of test scores, usefulness of tests, and their consequences for test takers. Test developers have to validate the inferences derived from the test scores rather than the test itself. This is to make sure that the test is grounded on ethical and evidential bases. To better understand validity, Messick (1995) defines it as “nothing less than an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use” (p. 742).

Test validation has been viewed differently by different scholars over the past few decades. Messick (1996), Bachman and Palmer (1996), and Kunnan (2004) three models of test validation have one aspect in common, which is consequences, that can affect test takers and teachers, society, and educational systems. This aspect of the three test validation models is what CLT focuses on, which is the impact and consequences of language tests in educational and political contexts as they are important and powerful tools for determining an individual’s future.

It can be inferred that CLT is interrelated with the development of language assessments, if assessment tools, i.e., rubrics are used appropriately, they will deconstruct power of tests. Shohamy (1998) supported the view that power of language tests must be shared with local bodies: test takers, students and teachers, so the assessment follows democratic principles and improves learners’ or test takers’ language proficiency. Thus, CLT may provide theoretical grounds for democratic language assessment. As noted above, Shohamy (1998) stressed the pivotal role of the social, cultural, political, educational and ideological factors that affect teachers and students in language testing. In this study, CLT is used as theoretical grounds by understanding EFL raters’ perceptions of rating and different factors that underline their performance through using holistic and analytic rubrics.

III. Literature Review

This section provides an overarching discussion of the literature related to the research topic. It reviews studies on historical overviews of writing assessments in the field of English language teaching. Moreover, it includes studies conducted in the Saudi EFL context on the issue of EFL writing assessment, and various types of rating scales; the definitions, advantages, and limitations of holistic and analytic rubrics. Finally, it reviews literature on factors influencing raters’ performance in rating writing essays using the two rubrics.

A. History of Writing Assessment

Writing assessment has been in existence for thousands of years, while writing composition started in the 1960s. The sons of the nobility in the period 111–771 BC prepared for Imperial Service with writing as a form of art among five others. University exams were orally conducted through the Middle Ages and until the late 19th century. Writing assessment underwent developments in US high schools and universities in the late 19th and early 20th centuries. US universities were influenced by Harvard University’s changes in replacing oral examinations with written exams. In the 1950s and 1960s, many educational assessment researches focused on “objective” testing. At the time of the emergence of this “objective” testing, there was a strong objection in the UK, particularly in Wiseman’s work (1949, 1956). Wiseman’s argument was rigorously for validity and what he called the “backwash effect”, more frequently called “washback”, which is the effect of testing on teaching (Hamp-Lyons, 2002).

Yancy (1999) portrayed the changes in writing assessment from 1950 onward in three waves, in which one wave contributes to another without a complete displacement of the previous wave. In the first wave (1950–1970), writing was assessed in the form of objective tests. During the second wave (1970–1986), holistic scoring took place. Finally, in the third wave (1986–present), writing assessment was, and is, done through portfolio and program assessment. In the first wave, students’ ability to write was evaluated through direct assessment and objective indirect assessment, which dominated the writing assessment practices of that period. This objective assessment of writing was in the form of multiple-choice questions, through which the students’ knowledge of word usage, grammar, and writing mechanics was tested. In the second wave, direct writing assessment came as a reaction to multiple-choice testing as each student produced a single writing sample that was “holistically” read and scored by trained readers. Moreover, these readers were asked to agree on specific scoring guidelines to give reliable scores. The last wave changed this trend, and students had to submit a collection of samples of their writing (portfolio) (Burkhart, 1999; Hamp-Lyons, 2002; Hout, 2003; Graham, 2000).

B. Writing Assessment Studies in the Saudi Context

Even though Saudi Arabia was never under the British empire, the Saudi government introduced the English language to its people 85 years ago, mainly due to the great expansion of the oil industry. Since English served as a lingua franca, the staff of the Arabian American Oil Company (ARAMCO, 1968) needed to acquire proficiency in English to communicate with the outside world. Similarly, people in general required to interact in English language with Muslims who annually visited Saudi Arabia to perform *umra* (a religious rite) (Al-Seghayer, 2005).

English is the only foreign language taught from kindergarten to higher education in the Saudi educational system (Al-Seghayer, 2005; Alshakhi, 2019). In addition, it is used as the medium of instruction in most universities’ technical, science, engineering, and medicine departments; industrial and governmental institutions; private and public organisations; private-sector job advertisements; and the media (Al-Seghayer, 2005; Shukri, 2014).

Several previous studies have tackled various issues related to the teaching and assessment of writing in the Saudi context (e.g., Aldukhayel, 2017; Alsamaani, 2014; Alshakhi, 2019; Ghalib & Al-Hattami, 2015; Hamouda, 2011; Obied, 2017). According to Alserhani (2007), traditional paper-and-pencil tests are still dominant in the teaching and assessment of writing in Saudi Arabia. Alserhani (2007) commented on the traditional grading system, stating, "According to the traditional grading system, writings are usually checked, given grades, and returned while students are passive participants in the assessment process" (p. 17). He highlighted the issues of teachers ignoring the writing process and being error-hunters rather than facilitators, which result in students' low writing performance.

According to Ansari (2012), more than 50% of EFL students do not know how to write in English language, which can be attributed to the learners' lower level of proficiency. There are several contributing reasons for Saudi students' low EFL proficiency levels. For instance, Arab EFL learners in general, write like information reporters rather than knowledge transformers. Moreover, Arab students' learning experiences during primary and secondary school often affect them negatively (Richardson, 2004). Furthermore, negative transfers from L1 to L2 writing result in unsatisfactory written production, short duration of terms, and the complexity of assessing writing (Hamouda, 2011; Hussein & Mohammad, 2011; Javid & Umer, 2014; Mohammad & Hazarika, 2016; Shukri, 2014). Al-Seghayer (2005) and Shukri (2014) attributed Saudi students' deficiencies in writing to the traditional methods used in teaching, which consist of drills and structured writing exercises.

Obied (2017) explored the Saudi EFL teachers' and students' perceptions and beliefs about writing assessment at a university in Saudi Arabia. She distributed two slightly different questionnaires to both teachers and students, which asked the participants about their perceptions of the writing assessment and rubrics and the development of the writing assessment. The data from her study indicate that writing assessment is a sensitive issue, and there is resentment among EFL students as the majority reported unfair grading. While the students affirmed their need to learn about the rubrics prior to the writing exams to achieve high scores, the teachers declared that there was insufficient time to go through the rubrics with their students, especially in big classes. Furthermore, the teachers added that they were not consulted in the design of the rubrics for their courses.

C. Factors Influencing Raters' Performance while Using Rating Scales

As previously discussed in the introduction, various factors may negatively affect the raters' quality of rating. Factors such as teaching experience, raters' fatigue, L1 influence, background, age and other factors. To ensure the validity and fairness of tests and test takers' scores, it is critical to eliminate any factor that may negatively affect the raters' rating performance. This section of the chapter presents some studies on various factors that impact the raters' rating.

Lim (2011) examined how novice raters' rating quality develops over time and the extent to which they maintain their rating quality. Lim's study was a part of a larger longitudinal study in an English language institute at the University of Michigan, UK. He found that the novice raters' severity and consistency were not distinguishable from those of the experienced raters, and their rating quality was a bit lower than their experienced counterparts. However, he demonstrated that, overall, the novice raters' ratings quickly improved with time.

Mahshanian et al. (2017) conducted a study that examined the effect of fatigue on the accuracy of rating writing essays in an EFL context. Their study showed a significant influence of fatigue on the raters' judgment and rating quality. That is, the more papers the raters rate and comment on, the more fatigued they become, and as a result, fatigue can endanger raters' judgement.

Rating accuracy and quality in writing among EFL learners is critical in determining their English language proficiency. Raters have to follow scoring protocols and guidelines to provide unbiased and reliable ratings; however, despite that, little is known about the factors that may negatively affect the accuracy of rating. The importance of understanding these factors lies in their impact on the raters' rating performance and on test takers' scores accordingly. Therefore, the raters must be critical and aware of their own performance in order to avoid any negative impact on the test takers.

Research Question

What are the factors influencing EFL raters' performance in rating writing essays using holistic and analytic rubrics?

IV. Methodology

This section discusses the methodology of this study, research design, context, participants, data collection instruments, data analysis methods, and ethical considerations.

A. Research Design

It is important to outline the research design that I have chosen for this paper and the way in which the instrument and the data yielded enable me to answer the research question. As this research uses the qualitative approach to answer the question *what are the factors influencing EFL raters' performance in rating writing essays using holistic and analytic rubrics?* Moreover, the paper looked into teachers' point of views and ideas; therefore, a qualitative interview-based approach was conducted in this study.

B. Context

This study took place at the women's campus of the ELI at a Saudi University located in Jeddah, Saudi Arabia during the academic year of 2019/2020. The program includes four levels: level one (beginners), level two (elementary), level three (intermediate), and level four (advanced). In each level, students are taught all four language skills: listening, speaking, reading and writing, in an integrated way. Every level has goals and learning outcomes, but in all four levels, the focus is on improving students' sentence structure and enhancing their written English language. In addition, students are taught all the required grammar rules and vocabulary to write short paragraphs, therefore, there is a heavy focus on teaching writing skill. Writing is assessed through three timed essays in each level, in which they choose from two different prompts to write about in forty minutes. The writing assessment rubric has a five-point analytical scale, and each level has its own rubric. Almost each level's rubric focuses on the same areas which are: structure and length, content and cohesion, grammar and mechanics, and lexical range.

C. Participants

The participants of this study were selected through convenience sampling, which is the most common sampling method in L2 research. I selected participants purposefully for the study as they "have to possess certain characteristics that are related to the purpose of the investigation" (Dörnyei, 2007, p. 99). Another reason for choosing participants through convenience sampling is the need for individuals' rich insights, which is best achieved by 'purposeful' sampling (Dӧrnyei, 2007). Participants were female teachers (due to gender segregation in Saudi education system), who hailed from different nationalities, had different qualifications, and varied L1 backgrounds and numerous years of teaching experience. They were five native English speakers, four native Arabic speakers and two instructors who were speakers of other languages. The extent of their teaching experience varied; there were novices (6 years) and others who had up to 26 years of teaching experience. Three of them were PhD holders, three MA holders, and five bachelor holders (see Table 1). To ensure confidentiality, I referred to the raters with ID codes consisting of T+ a number (1 to 11) and pseudonyms.

TABLE 1
DEMOGRAPHIC INFORMATION ABOUT PARTICIPANTS

Name	L1	Years of experience in teaching EFL	Postgraduate qualifications	Education qualifications
T1 Laura	English	18 years	BA in Elementary Education MA in ESL	None
T2 Fatima	English	7 years	BA in Molecular Biology MA and PhD in Business	None
T3 Mila	English	7 years	System Analysis	Celta/TEFL/IELTS trainer
T4 Caroline	English	16 years	Modern European Studies MA in TESOL	Cambridge Certificate in Teaching English as a Foreign Language
T 5 Aliya	English	10 years	Business	None
T6 Natalie	Arabic	23 years	PhD in English Literature & TESOL	Special Diploma in Education in Testing and Evaluation
T7 Maryam	Arabic	26 years	BA in English	None
T8 Nicole	Arabic	8 years	BA in English	TEFL
T9 April	Arabic	6 years	BA in English MA & PhD in Education Technology	None
T10 Suzan	Other languages	20 years	BA & MA in English Literature	TEFL/TESOL
T11 Mary	Other languages	9 years	Mechanical Engineering	CELTA

D. Data Collection Instruments

To meet the primary research objective, data for the study were obtained from one main instrument: immediate interviews. To help participants produce useful data, I provided them with precise instructions and some training about the grading panel and the rating task at the beginning. The data collection process included individual interviews with all the participants. Meanwhile, I recorded the interviews using voice memos on my iPhone. This stage of the data collection process took about two hours.

The interviews were consisted of two phases; first phase was face-to-face interviews while the second phase was online, and took place few months later (due to Covid-19 and summer break at the university). I chose semi-structured interview which is easily controlled and flexible at the same time. In addition, a set of interview questions was written in advance with some follow-up questions. In addition, the participants were allowed to explain their thoughts and views on their experience of rating and the factors that usually affect their rating performance. Interview questions comprised two sections: the first section asked about background information/demographics, while the second one asked about the participants' rating experience related to holistic and analytic rubrics and what affects their rating decisions. The first section of the interview was needed to individual differences between instructors that may affect their rating practices, while the second section was designed to collect additional data on the writing assessment practices and related challenges. The second phase took place in later with some of the participants in order to clarify a few responses from the previous phase that required further exploration.

E. Data Analysis Methods

All interviews recorded and transcribed for the purpose of thematic analysis. First, I started identifying tentative or initial codes to describe the content of the interview transcripts. Then, I searched for identical patterns and themes across all interviews, and later defined and labeled these themes. Finally, I arrived at a final list of themes that described the data in a way that answered the research questions.

F. Ethical Consideration

This research project rigorously followed the institute's ethical guidelines throughout its process. Several ethical considerations were taken into consideration to ensure that the study was conducted in an appropriate manner. The data collection form was submitted to and approved by the Graduate Studies and Academic Research Unit at the ELI. Consent forms including a brief description of the nature of the study were distributed among the participants prior to start of the data collection process. The participants read the forms and signed them once they agreed to participate in the study. I reassured them about confidentiality and their right to withdraw at any point during the interviews if they wished. Moreover, I obtained their permission to record the interviews. It was further explained to the participants that their information would only be discussed with the supervisor. As mentioned earlier, to ensure confidentiality, I referred to the raters with ID codes consisting of T+ a number (1 to 11) and pseudonyms.

V. Findings

This section answers the research question in one coding scheme (Factors influencing EFL raters' rating performance). A list of categories and subcategories for this coding scheme are listed below in Figure 1.

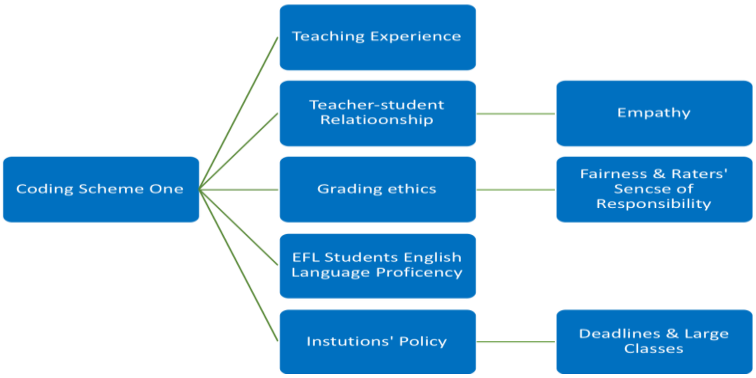
Coding Scheme One
Teaching Experience
Teacher-student Relatioonship
Grading ethics
EFL Students English Language Proficency
Instutions' Policy
Empathy
Fairness & Raters' Sencse of Responsibility
Deadlines & Large Classes

Figure 1: Coding Scheme One List of Categories and Subcategories

In the interview, the participants were asked about the factors that influence their rating performance. Their responses differed and accordingly five distinct themes emerged from the interviews' analysis.

A. Theme 1: Teaching Experience

There were nine responses to the question related to the effects of teaching experience on the participants' rating practices. They were asked if they had changed over their years of teaching experience and in what way? Multiple participants reported that their years of teaching experience has shifted their rating performance and changed them from being strict and firm raters to less strict and lenient sometimes. Aliya stated: "I am definitely not the person I started teaching. I've mellowed with time. I used to be stricter and now I am, I wouldn't say lenient, I would say less strict" (Interview 2). Another participant Mary also said:

> When I first came into teaching English, I think I had different expectations, but now after doing it for so many years, my outlook is different. I became more lenient. (Interview 1).

In contrast, Laura and Mary reported that they became stricter with time. Laura reflected on her rating experience and talked about how she became stricter by saying "I think I became a lot stricter over the years and you have to avoid nepotism or favoritism because you might be called to justify". She also added "I came from an environment where if they thought that the teacher was too lenient, they would question her" (Interview 1).

B. Theme 2: Teacher-Student Relationship

In addition to mentioning the effects of teaching experience on EFL raters' rating practices, interviews also elicited various ways of establishing relationships between the participants and their EFL students. The theme of a teacher-student relationship is described in two parts: a) empathy and b) boosting students' autonomy.

C. Empathy

The participants were asked during their interviews if their relationship with their students affects their rating decisions in any form. Field notes taken during each interview recorded that the participants tended to pause before answering this question. The delay in their responses to this question seemed to be more about the sensitivity of this topic or not wanting to be judged. The participants had different point of views about emphasizing with their students. During one of the interviews, Suzan reported:

> It affects all my other dealings except convictions of my formal exam papers and why, because I believe it's very important to have ethics in place, one of the reasons I was recruited as a teacher is because I value ethics. Anyone coming into this profession I believe, needs to be guided by ethics. And if I let, my relationships interfere with my work, then obviously that wasn't going to be ethical. And then in this case, we're talking about writing, yes, it doesn't get to in a way marking formal writing assessment, if it isn't assessment, I don't mind. (Interview 2).

Another participant, April, who is an Arab native speaker talked about how being a non-native EFL teacher affected her rating; "I think the second language teachers are more, let's say understandable, because we know the situation, we've been boxed by this situation of the second language learners" (Interview 1).

D. Theme 3: Grading Ethics

A third theme that emerged from the participants' responses pertained to grading ethics. Specifically, the participants discussed the responsibility they undertake while rating students' work and the urge to be fair enough when they give grades. These have been collected under the boarder idea of 'grading ethics'.

E. Fairness and Raters' Sense of Responsibility

Evidence of the importance of fairness in grading students' assignments or exam papers was reported by most of the participants, who highlighted instances such as being professional raters, students' future studies and jobs, and credibility issues. For example, Mary said this about the difficulties and challenges a rater might face to achieve fairness:

> In my very honest opinion, I don't think that we will ever be like a hundred percent fair when it comes to the grading of writing. I believe that everyone has his own personal coin in it. I think that this is something personal that teachers are not trying to be fair, but maybe even the more you are trying to be fair, the more it gets like you don't know how to decide. It's difficult. It's just the student, especially if you have in your mind that by giving her this grade, this will affect her GPA. (Interview 2).

Similarly, Mila stated:

> It's a struggle within yourself because you don't want to give a passing grade without any effort on part of the students. And you don't want to give them the grade just because for example, these girls are going into medicine and every form counts. So, you have that inside yourself thinking, okay, I don't want to mess up her chances of getting into medicine, but again, you don't want to send them off. (Interview 1).

Mila went on to say:

> And do you want someone who's working in the hospital who doesn't know how to read English? They're giving medicine, they're giving medical treatment and they can't read English. That can cause a problem. So that is in my head. These people are going to be doing research. These people are going to be saying that I taught them English. (Interview1).

Other participants also talked about giving unfair grades and their negative influence on students' performance. For example, Laura said: "we should be a little lenient and this is nice, but that doesn't benefit the students". (Interview 1). Similarly, Caroline commented on being professional in giving grades and helping students:

> I think they come to me and say, 'teacher, this is my last chance. Please help me'. But in the end, you know, we have to be professional and, you know there are issues of credibility here. So, I wouldn't say I'm generous. I want to be lenient. I do try to see what the students can do rather than what they cannot do. (Interview 1).

F. Theme 4: EFL Students English Language Proficiency

In addition to the previous themes which emerged from the participants' responses, a fourth factor that affects their rating practices was found. Four out of eleven participants believed that students' low English language proficiency and being EFL students is a factor that should be taken into consideration by EFL raters. Aliya described her experience in teaching EFL students and how it is different that ESL teaching, and the way it affected her teaching and rating practices.

> I've been in this country for long enough to understand the difference between ESL and EFL. So, it's very different from somebody who's studying English as L1, completely different as grammar, the vocabulary, and the way of teaching is different. I knew, I'm coming somewhere where English is not spoken. As, you know, not even as a second language. English is mostly spoken as a foreign language, so it's mostly EFL. So, I knew that, so bearing this in mind, it was easier to adapt to the teaching context. My background didn't really affect. The one thing that did affect, I will have to say is that I had to simplify my speaking to suit the needs of the students. (Interview 2).

Moreover, Caroline's statement is indicative of the belief that raters should take into consideration that these students are EFL students, and native English teachers or raters should have realistic expectations from them.

> The fact that they are EFL learners makes you feel anxious. It’s very easy as a native speaker to think, Oh, dear, look at all of these spelling mistakes, look at all these punctuation mistakes and expect them to write something perfect, which we shouldn’t do. You’ve got to be careful that you’re not expecting too much of them simply because you’re a native. (Interview 1).

Mila echoed this belief as well: “I take into account that they’re learners of a foreign language and speak English as a foreign language, and not English as a second language. So, we need to take that into consideration that this is not their language”. (Interview 1).

Mary also mentioned that the level of the students and the track they are in often affect her rating decisions and practices: “I’d say the level of your students. This is what I told you that it really depends on a lot. Whether I have a good group or a bad group, if I have general track students or science students”. (Interview 2).

G. Theme 5: Institution Policy

As all educational institutions have plans developed to guide their actions, the ELI policies where the participants of this study work, influence their rating performance. Some participants expressed that having a big number of scripts to grade or not having an ample time to grade accurately, might negatively affect their rating decisions.

H. Deadlines and Large Classes

Aliya reported that unreasonable deadlines don’t allow her to grade papers the way she wishes:

> Time is a big factor. You are given some time, too many papers to grade. And I remember, I had over 60 or 70 copies to grade over a weekend. It was too much. Then of course, my concentration doesn’t remain the same. (Interview 2).

Mary explained how she gets affected by grading the number of papers. She talked about her different rating experiences when having a few or many papers to grade:

> How to say that kind of feeling that you have, like, I only have 15 papers. So, I take my time and I really relax during that and so on. But if I have 43 scripts, I have to divide them on two days and you know, like I’ve read so many things and even the things that I read, they influenced you somehow. So, I believe that the number of the papers you have to grade influences you as well. (Interview 2)

Suzan also affirmed that time is an important factor. She said:

> One of the things that genuinely affects me, what I shared previously were like circumstantial things that generally affect me. For example, the administrative demands and this, I believe plays a real role. So, for example, if the administration is going to be like, you have to submit that by tomorrow, or like, you know, just immediately after the weekend. I think it’s very unfair and it definitely affects my performance because it is not possible to be human and, you know, like correct, 30 or 40 papers back-to-back and do justice to everyone. So, time is a main factor. (Interview 2).

VI. Discussion

This section aims to discuss the results in the light of the reviewed literature previously. This study set out with the aim of exploring the factors that affect EFL instructors’ rating performance.

With respect to the research question, which aimed to identify the factors influencing EFL raters’ rating performance, five distinctive factors were found. These factors are: teaching experience, teacher-student relationship, grading ethics, EFL students’ English language proficiency, and institution’s policy. This question looks into the raters’ performance from different aspects; their emotions and feelings while rating students’ writing, and any aspect that may have an influence on them while rating.

The findings of this study confirmed the effect of the teachers’ teaching experience on their rating performance, as nine responses out of eleven confirmed that they became less strict with time. Only two participants, Laura and Mary were the opposite as Laura said: “I think I became a lot stricter over the years and you have to avoid nepotism or favoritism because you might be called to justify” (Interview1, 2020). This finding contrasts with Lim’s (2011) conclusion who found no distinct differences between novice and experienced raters’ severity while rating. A possible explanation for these different findings might be due to the different context where the study took place, the participants’ backgrounds, educational level, and native language. Hence, future studies will most likely reveal different results.

In the current study, most of the participants highlighted a significant factor which is the importance of being as fair as possible to their students while rating their writings as this is an important part of being an ethical and professional EFL rater. This result ties well with the theory we foregrounded in this study, the CLT that considers the impact of tests on learners by discussing fairness and bias in language testing.

In addition, the findings showed a high level of responsibility towards giving students what they deserve as this will affect their future learning and professions. Mila commented saying:

> And do you want someone who's working in the hospital who doesn't know how to read English? They're giving medicine, they're giving medical treatments and they can't read English. That can cause a problem. So that is in my head. These people are going to be doing research. (Interview1, 2020).

As an EFL rater, I totally agree with this belief, it is my responsibility as a rater to be fair and responsible as my rating will affect the students' future. In my personal opinion, language testers have a huge responsibility on their shoulders, as Shohamy said "the conflict between professionalism and morality, and the unequal power relations between test makers and test takers" (p. 331). I also believe that rubrics make assessing students' work efficient, consistent and objective, and provide them with a clear understanding of what is expected from them and as a result, they improve their overall learning.

The participants' sense of fairness and justice confirmed in this research study highlights the power of rubrics as guiding principles while rating. According to Shohamy (1998), tests can move beyond concerns over validity and reliability issues in language assessment as they are vehicles affecting political, educational, and social domains. Tests are potentially used by testers in authority to control, manipulate, or change individuals, institutions, and society at large. They are also used as powerful turning points in test takers' lives such as passing a course, attending a university, getting a job etc. Thus, CLT calls for democratic testing practices in order to reduce the negative impact of tests on individuals and society at large. As shown in the excerpt above by Mila, the participant revealed her awareness of the effect of using rubrics strictly on students' future and on the society in general, therefore, this implies that from CLT perspective, having an accurate testing tool will result in fair and just students' evaluation. The participants' awareness of their assessment practices is what CLT came as a movement, for CLT raises testers' awareness of their own rating practices and results in more justice for test takers.

Moreover, several participants, especially non-native English raters admit that they sympathize with their EFL students, as they have been through a challenging language learning experience themselves. A number of native English instructors also indicated that the low English language proficiency and the fact that these learners study English as a foreign language should be taken into consideration. For example, one native-English speaker participant, Caroline, stated that:

> The fact that they are EFL learners, anxious, Um, it's very easy as a native speaker to think, Oh, dear, look at all of these spelling mistakes, look at all these punctuation mistakes and expect them to write something perfect, which we shouldn't do. You've got to be careful that you're not expecting too much of them simply because I am a native. (Interview1, 2020).

However, the effect of this sympathy does not exceed classroom teaching and informal assessment. This finding is in agreement with the study's theoretical framework, CLT. CLT perceives testing as tools that are related to social and educational contexts, however, they are designed to minimize the negative consequences of tests if used appropriately. In this study, even though the participants listed several factors underling their rating performance, the existence of rubrics reduced the effect of these factors on students' scores which resulted in giving fair grading.

The last factor that influences EFL raters' performance, according to several responses, is the institution's policy. More specifically, being asked to rate a large number of papers in a short time as raters may not be able to rate all papers equally with the same concentration. These responses imply that EFL institutions or schools are encouraged to devote an ample time to raters to ensure an equal and fair grading process. Take Aliya for example talking about the negative consequences of having a short deadline:

> Time is a big factor. You are given some time, too much papers to grade. And I remember, I had over 60 or 70 copies to grade over a weekend. It was too much, then of course, my concentration, it's not the same. (Interview2, 2020).

In this respect Suzan added:

> One of the things that genuinely affect me, what I shared previously were like circumstantial things that generally affect me are, for example, the administrative demands and this, I believe plays a real role. So, for example, if the administration is going to be like, you have to submit that by tomorrow, or like, you know, just immediately after the weekend, I think it's very unfair and it definitely affects my performance because it is not possible to be human and, you know, like correct, 30 or 40 papers back to back and do justice to everyone. So, time is a main factor. (Interview2, 2020).

The findings illustrate that heavy workload can negatively impact raters' and students' performance. Raters may have to provide students with inaccurate rating or feedback, harsh or easily unjustified grading, etc., due to the unmanageable workload they have to handle and the pressure they feel.

VII. Conclusion

The findings of the study evidently suggest that EFL educational institutions should give raters adequate time to rate and provide the appropriate rating criteria that suit the learners' level and language proficiency. Moreover, EFL raters should adhere to the given rubric, give their feedback on the effectiveness of these rubrics, share the rubrics with their students, and provide suggestions for rubrics' improvements if needed. Finally, the findings of this study contribute to the body of literature on the effects of rubrics on EFL instructors' rating performance, their overall teaching and assessment practices, and the EFL learners' learning outcomes.

Language assessment is of importance to practitioners in all fields of education. This work may have some implications for all educators. It proposes that all stakeholders in language teaching/assessment contexts need to be aware of the role of raters, their different attitudes towards rating or rubrics, and their understanding of the rating criteria assigned to them.

Course managers need to think carefully about what form of rating to use for exam/test purposes. There needs to be forms of language assessment that is consistent with teachers' and course managers' instructional goals and with learners' learning development.

REFERENCES

[1] Al-Seghayer, K. (2005). Teaching English in the kingdom of Saudi Arabia: Slowly but steadily changing. In G. Braine (Ed), *Teaching English to the world: History, curriculum, and practice* (pp. 125-134). Mahwah, NJ: Lawrence Erlbaum.

[2] Aldukhayel, D. M. (2017). Exploring students' perspectives toward clarity and familiarity of writing scoring rubrics: The case of Saudi EFL students. *English Language Teaching*, *10*(10), 1-9.

[3] Alsamaani, A. (2014). Evaluating classroom assessment techniques of novice Saudi EFL teachers. *Journal of Arabic and Human Sciences*, *270*(1687), 1-40.

[4] Ansari, A. A. (2012). Teaching of English to Arab students: Problems and remedies. *Educational Research, 3*(6), 519-524.

[5] Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System*, *29*(3), 371-383.

[6] Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

[7] Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74.

[8] Burkhart, F. (1999). Samuel Messick, 67, leader in the educational testing field. *European Journal of Psychological Assessment, 15*(1), 87-87.

[9] Dörnyei, Z. (2007). *Research Methods in Applied Linguistics*. Oxford: Oxford University Press.

[10] Ghalib, T. K., & Al-Hattami, A. A. (2015). Holistic versus Analytic Evaluation of EFL Writing: A Case Study. *English Language Teaching*, *8*(7), 225-236.

[11] Graham, B. (2000). *Trends & Issues in Postsecondary English Studies*. National Council of Teachers of English, 1111 W. Kenyon Road, Urbana, IL 61801-1096.

[12] Hamouda, A. (2011). A Study of students and teachers' preferences and attitudes towards correction of classroom written errors in Saudi EFL context. *English Language Teaching*, *4*(3), 128-141.

[13] Hamp-Lyons, L. (2002). The scope of writing assessment. *Assessing writing*, *8*(1), 5-16.

[14] Han, T., & Huang, J. (2017). Examining the impact of scoring methods on the institutional EFL writing assessment: A Turkish perspective. *PASAA: Journal of Language Teaching and Learning in Thailand*, *53*, 112-147.

[15] Hussein, A., & Mohammad, M. (2011). Negative L1 impact on L2 writing. *International Journal of Humanities and Social Science*, *1*(18), 184-195.

[16] Javid, C., & Umer, M. (2014). Saudi EFL learners' writing problems: a move towards solution. *Proceeding of the Global Summit on Education GSE*, 4-5.

[17] Kayapınar, U. (2014). Measuring essay assessment: Intra-rater and inter-rater reliability. *Eurasian Journal of Educational Research, (57)*, 113-136.

[18] Kunnan, A. J. (2004). Test fairness. In Milanovic, M. & Weir, C. (Eds). *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27-48). Cambridge University Press.

[19] Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543-560.

[20] Mahshanian, A., Ketabi, S., & Abbas, R. (2017). Raters' fatigue and their comments during scoring writing essays: A case of Iranian EFL learners. *Indonesian Journal of Applied Linguistics*, *7*(2), 302-314.

[21] Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*: 741–9.

[22] Messick, S. (1996). Validity and washback in language testing. *Language testing*, *13*(3), 241-256.

[23] Mohammad, T., & Hazarika, Z. (2016). Difficulties of learning EFL in KSA: Writing skills in context. *International Journal of English Linguistics*, *6*(3), 105-117.

[24] Obeid, R. (2017). Second Language Writing and Assessment: Voices from within the Saudi EFL Context. *English Language Teaching*, *10*(6), 174-181.

[25] Richardson, P. M. (2004). Possible influences of Arabic-Islamic culture on the reflective practices proposed for an education degree at the Higher Colleges of Technology in the United Arab Emirates. I*nternational Journal of Educational Development, 24*, 429-436.

[26] Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493.

[27] Shohamy, E. (1998). Critical language testing and beyond. *Studies in educational evaluation*, *24*(4), 331-45.

[28] Shukri, N. A. (2014). Second language writing and culture: Issues and challenges from the Saudi learners' perspective. *Arab World English Journal*, *5*(3), 190-207.

[29] Yancey, K. B. (1999). Looking back as we look forward: Historicizing writing assessment. *College composition and communication*, *50*(3), 483-503.

Shifa Soror Alotibi has a Bachelor of English language, King Abdulaziz University, Jeddah, Saudi Arabia 2011. She received her masters' degree in TESOL, King Abdulaziz University, Jeddah, Saudi Arabia, 2021. With particular research interest in language testing and assessment. She worked as an EFL instructor at King Abdulaziz University from 2011 until 2020. She was also a member of Research Committee, Translation Committee, Quality Control of Supplementary Material Committee, and a learning management support teacher at King Abdulaziz University.

Abdullah Mohammed Alshakhi is an Associate Professor of Applied Linguistics and the Head of Curriculum and Testing Unit at the English Language Institute at King Abdulaziz University. With particular research interests in language testing and assessment literacy, construct validity, writing assessment, and language policy. He has published in both local and international journals. He is actively involved in several workshops involving language assessment and testing through the Educational Testing Service (ETS), Cambridge Assessment, and the European Association of Language Testing and Assessment (EALTA)