European Journal of English Language Teaching
ISSN: 2501-7136
ISSN-L: 2501-7136
Available on-line at: www.oapub.org/edu

doi: 10.5281/zenodo.1079871 Volume 3 | Issue 2 | 2017

RATERS' PREJUDICES IN ORAL PERFORMANCE ASSESSMENT

Murat Polat[i],
Emel Akay
Dr., Anadolu University,
Turkey

Abstract:
Research on testing speaking claims that raters' beliefs, perceptions and even their prejudices may be involved in the process of grading although they are given a set of rubrics to stay on the same track and have stable qualities on the assessment of oral production; therefore, many researchers have studied the rationale of such beliefs and the amount it affects the scores. This study aimed to find out whether the perceptions and beliefs of raters play a significant role in testing speaking and question the role of experience in the involvement of such beliefs. To do that, a group of raters were asked to grade the audio recordings of ten voluntary students twice with one-month-interval in between, being misinformed about the students' physical appearances each time with the help of different pictures, and were interviewed later to identify whether their pre-conceptions on students' physical appearances play a role in their grading oral performances. Also, the data obtained were used to draw some conclusions whether the raters intentionally or unintentionally used their beliefs in the grading process. The analysis revealed that student appearance may be significantly effective in teachers' grading and this is true especially for experienced teachers who believe their judgements are true and unbiased more than the less experienced ones.

Keywords: performance assessment, rater prejudice, bias, halo effect, physical appearance

[i] Correspondence: emailmpolat@anadolu.edu.tr

1. Introduction

In the assessment of foreign language performance, the assessors or graders are mostly involved in the direct evaluation of spoken or written responses of the students. Most of the time scores assigned by those raters are not double-checked or verified and considered as the genuine interpretations of learners' language performance; thus, the assigned scores or those graders directly have an impact not only on students' test results but also on assumptions made about testees' language learning and their overall performances. Consequently, what graders do in these performance tests have been considered as a possible source of variance that may affect the correctness of the students' ability (Akay & Toraman, 2015; Bachman, 2004; Crisp, 2012; Kim, 2015). Therefore, no matter what the assessment method is, rater judgements & possible errors in those judgements and the source of those errors in foreign language performance assessment have gained considerable importance and studied in order to sustain reliability and validity evidence of such tests and their scorings (Johnson & Lim, 2009; Kenyon, 1992; Polat, 2017; Reed & Cohen, 2001). That is why most of the studies that have focused on rater effect on performance tests dealt with potential rater bias and halo effect which may directly or indirectly change the students' performance ratings significantly (Banks 1998; Eckes, 2005; Read et al., 2005).

2. Rater Bias and Halo Effect

In foreign language testing a number of rater effects such as being biased, halo effect, considerable leniency or severity of raters' and scale shrinking can be direct sources of method variance which is at the same time a source of systematic variance that must be associated not with the students' performance but with the raters' beliefs and judgements (Cronbach, 1995; Eckes, 2005; Hoyt, 2000). Among those, rater bias, its possible sources and effects have been studied by many researchers in language testing (Myword & Wolfe, 2003, Newstead & Dennis, 1990; Kondo, 2002; Centra & Gaubatz; 2000; Aydin et al., 2016). In the assessment of speaking performance, there might be a number of bias that could operate in a performance marking process. Test scores are considered biased if a test design, or the way results are interpreted and used by the decision makers, systematically disadvantageous to certain groups of students over others, such as students' colour, gender or age, students from lower-income backgrounds, students who are not proficient in the target language, or students who are not typical of what raters used to see (Centra & Gaubatz; 2000, Crocker & Algina 1986). Regardless of their psychological reasons and derivations, there is a considerable variety of bias types that could function in scorings. Of these, some biases may be

almost completely undeliberate that is why raters never consider this affecting their grades whereas others can be highly aware of this fact and there are even some cases that raters defend their subjectivism while grading (Dennis & Plymouth, 1990; Hedge, 2000). Also, in some cases bias in testing may be based on raters' stereotyping, for instance, the anticipation that a certain group of people or students will perform in a certain way or up to a certain level while another group would never do that or over perform (Schaefer, 2008). For instance, in a study done by Wigglesworth (1993), the bias was based on some subjective knowledge of individual students and the rater could depend on if the student is a well-behaved one or not while grading his/her speaking performance.

Another and sometimes very commonly seen form of bias is on students' gender and appearance. Although literature reveals that there is little firm evidence of gender bias in scoring, for some graders, gender of the student may be the cause of possible bias in evaluating the performance (Boyce, 1979; Eagly & Mladinic, 1994). Newstead (1996) noted that among all kinds of bias in performance assessment, gender bias and its reasons has been the most widely researched topic since too many candidates complains about the existence of such variance. Though it is known that rater bias cannot be completely removed, many studies were made in order to find out the psychological reasons of it. Read et al. (2005) stated that since there is human effect in performance evaluation no one can call it completely arbitrary, what stakeholders have to do is to minimize potential bias agreeing on the fact that raters' judgements are affected by their social positioning, the way they use their professional experience and human relations. Accordingly, a number of researchers claimed that performance rating can be gendered or even the appearance of the testees may be taken into account while scoring how well they perform in an exam (Francis et al, 2001; Harding, 1991; Murphy & Elwood, 2002). Thus, the last but not the least form of rater bias is associated with the appearance.

Langlois et al. (2000) stated that while making decisions people quite often make judgements based on simply the appearance and argued that "if it were not true, they wouldn't remind their offsprings not to judge books by their covers (p 392). It is commonly accepted that some personality traits like the shape of your face, the way you look, how you smile, dress, your hair style or even the shape of your glasses can give some ideas to people about your way of thinking, the group of people you belonged to or may ring some bells reminding someone who had a place in that person's experiences. Also, Umberson & Hughes (1987) stated the same fact and summarised that while making evaluations graders may sometimes tend to organize their judgements, perceptions and expectations on learners around their sex, age, clothing and appearance which are observable characteristics of the individual that may

give a rough idea on its status in the society. This notion is known as the basis of Status Characteristics Theory. Berger et al. (1977) stated that expectations frame how the individuals are evaluated by the teachers or raters. For instance, if you are a student at a school, you should look like a student rather than a model or a punk. This state is also associated with another form of score variance, halo effect. This notion is also known as "*implicit bias*" and Talamas et al., (2016) explained "*if a rater knows something favourable about someone, s/he tends to judge other things favourable and often the first thing picked up is biased."* If you are wearing thick glasses it could be referred that you are a frequent reader, or if a 17-year old- college girl never wears make up, it could be assumed that she spends most of her time on her studies. Krawccyzk (2017) assorted that gender and physical attractiveness may be two very different dimensions of the bias problem but also interrelated most of the time, that is why, some justification of treating them jointly in a study could be a good idea to see if beauty could be a curse at times.

Clearly, the question of sex and appearance bias in performance assessment has not been clearly and fully resolved since some studies reflect no significant bias proof. The study reported here, therefore, may address on these issues and could underline how much prejudices of raters could affect their own scorings when consciously or unconsciously they operate their judgements. Unlike the previous studies, the raters all scored students from their own gender and this can be another perspective on sex bias since some raters can be more lenient to the other sex or could be harsh when they see someone more beautiful and attractive then the stereotypes in their surroundings. Finally, the role of the halo effect is also considered important in this study and participants were asked questions about this phenomenon to clearly state their expectations from students apart from performing well in exams.

3. Method

The aim of this study is to investigate whether the participants (speaking raters) assess students more harshly or leniently because of their physical appearance although they use the same scoring rubric. This descriptive study tries to find answers to the following research questions:

1. Is there a relationship between the scores of raters and students' physical appearance?
2. Do the raters score a student's oral performance differently considering his/her physical appearance?
3. Is there a significant difference among the scores of certain groups?
4. Do the raters' scores vary according to the sessions held?

3.1 Participants

20 female teachers working at Anadolu University School of Foreign Languages were randomly selected and invited to contribute to the research. Of these 20 teachers, 18 accepted to take part in the study voluntarily. Total year of experience of these teachers in teaching English ranged between 2-21 years (11 teachers were experienced between 2-5 years and 7 teachers were experienced more than 6 years up to 21 years), and all of them had a previous experience in scoring oral performances as they did in midterm or proficiency exams of the institution. Male teachers were deliberately excluded from the study to avoid gender bias in scoring as the audio recordings belonged to female students. Also, 10 volunteer female students who were learning English in the pre-intermediate level of a language teaching program participated in the study.

3.2 Instruments

Prior to the study, 10 female students were recorded in an exam condition. They were asked a description question to answer, which is "*How can you describe your ideal partner?*" Each student was given the same question and answers were audio recorded. These recordings were used in the scoring procedure by the teachers. In addition, 10 pictures, 5 showing beautiful models and 5 showing normal student girls, were used to show the graders while scoring. These pictures were all false and did not belong to the actual students. Also, an analytic scoring rubric to assess the spoken performances was given to the graders. In fact, teachers had been using this rubric for several years to assess speaking achievement exams, that the graders had already been familiar with the rubric.

3.3 Procedure

Two grading sessions were held during the study to collect data. 10 recordings were divided into two sets including 5 in each. In the first grading session, the raters were given the first set and they were shown model pictures as the owners of the voice and they were asked to score the oral performance according to the speaking rubric. After a 20-minute break, the raters were given the second set of recordings and this time student pictures were shown as the owners of the recordings. The same procedure was followed after a one-month interval in the second grading session. In the second session, the grading order of the sets was switched. The raters initially scored the second set of audio recordings with the student pictures in the background, and after a 20-minute break, they graded the first set seeing the pictures of models. After all the scorings were done, semi structured interviews were held with voluntary participants on the results and the reasons of those results.

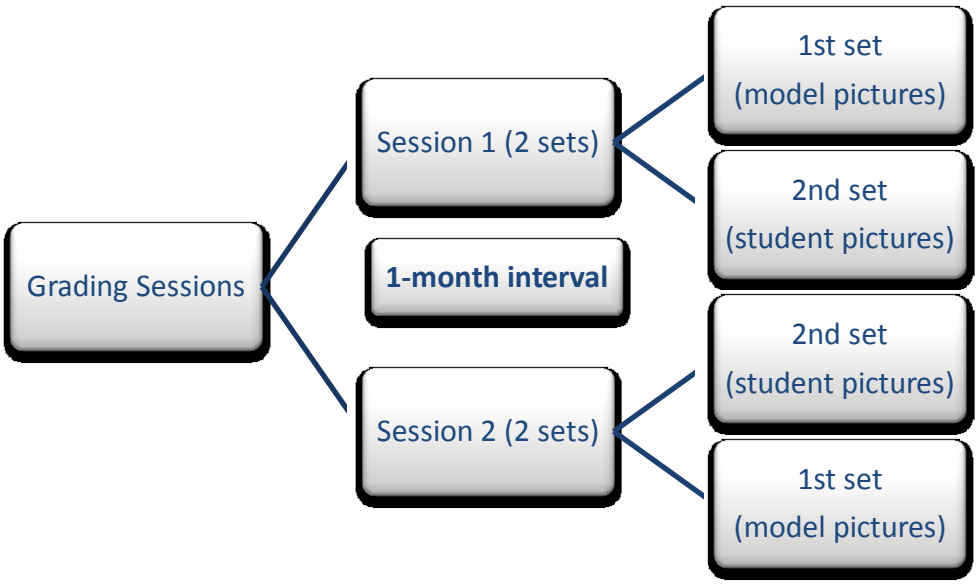


Figure 1: Conceptual framework of the study

3.4 Data Collection and Analysis

The data consisted of the raters scores on each grading and the recordings which include semi-structured interviews with 13 of the participant teachers. The scores were collected during the two grading sessions from the teachers and used in the analyses. In order to identify the difference in scores according to the pictures shown and the grading sessions, Partial Correlation, Independent Samples T-test, and Paired Samples T-test were calculated by using IBM SPSS version 20 software.

4. Findings

In order to identify the analyses to conduct, the normality of the scores was tested. The results are presented in Table 1.

Table 1: Test of normality results

	Picture	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Statistics	df	Sig	Statistics	df	Sig
Session 1	model	.124	180	.200*	.985	180	.945
	student	.097	180	.200*	.965	180	.405
Session 2	model	.074	180	.200*	.988	180	.977
	student	.104	180	.200*	.977	180	.736

As can be seen in Table 1, the scores assigned to the recordings showed normal distribution according to the results of both Kolmogorov-Smirnov and Shapiro-Wilk

tests (p>0.05). Thus, parametric tests were utilized to find answers to the aforementioned research questions.

4.1 Relationship between scores and appearance

To answer the first research question, Partial Correlation Analysis was conducted on each session scores. As the aim of this study was to analyze the effect of physical appearance on raters' scores, the main effects of raters and recordings were controlled.

Table 2: Partial correlation results for session 1

Correlations						
Control variables			**Score**	**Picture**	**Rater**	**Audio**
None	**Score**	Correlation	1.000	.465	-.212	.239
		Sig.(2-tailed)	.	.000	.105	.066
		df	0	178	178	178
	Picture	Correlation	.465	1.000	.000	.870
		Sig.(2-tailed)	.000	.	1.000	.000
		df	178	0	178	178
	Rater	Correlation	-.212	.000	1.000	.000
		Sig.(2-tailed)	.105	1.000	.	1.000
		df	178	178	0	178
	Audio	Correlation	.239	.870	.000	1.000
		Sig.(2-tailed)	0.066	.000	1.000	.
		df	178	178	178	0
Rater & Audio	**Score**	Correlation	1.000	.552		
		Sig.(2-tailed)	.	.000		
		df	0	176		
	Picture	Correlation	.552	1.000		
		Sig.(2-tailed)	.000	.		
		df	176	0		

a. Cells contain zero-order (Pearson) correlations.

Table 2 gives the zero-order Pearson correlations when no variables were controlled. The results indicate that there is a positive correlation between scores and pictures shown ($r = 0.465$, $p < 0.001$). In the data set, model picture was coded as 1 and student picture was coded as 2. Thus, positive correlation explains that raters gave higher scores to the audios accompanied by student pictures. On the other hand, no correlation between raters, audios and scores was found (p>0.05). When the rater and audio variables were controlled, the correlation between scores and pictures increased ($r = 0.552$, $p < 0.000$). $R^2 = 0.304$ value of the correlation was estimated, which means the pictures shown are responsible for the variation of the scores at about 30.4%.

The same Partial Correlation Analysis was conducted on data collected in second session. Table 3 gives the results of this procedure.

Table 3: Partial Correlation results for session 2

Correlations						
Control variables			**Score**	**Picture**	**Rater**	**Audio**
None	**Score**	Correlation	1.000	.286	-.241	.075
		Sig.(2-tailed)	.	.027	.064	.571
		df	0	178	178	178
	Picture	Correlation	.286	1.000	.000	.870
		Sig.(2-tailed)	.027	.	1.000	.000
		df	178	0	178	178
	Rater	Correlation	-.241	.000	1.000	.000
		Sig.(2-tailed)	.064	1.000	.	1.000
		df	178	178	0	178
	Audio	Correlation	.075	.870	.000	1.000
		Sig.(2-tailed)	.571	.000	1.000	.
		df	178	178	178	0
Rater & Audio	**Score**	Correlation	1.000	.465		
		Sig.(2-tailed)	.	.000		
		df	0	176		
	Picture	Correlation	.465	1.000		
		Sig.(2-tailed)	.000	.		
		df	176	0		

a. Cells contain zero-order (Pearson) correlations.

Table 3 gives the zero-order Pearson correlations when no variables were controlled. Similar to the results reached in the previous analysis, no correlations between raters, audios and scores were identified ($p>0.05$), whereas the correlation between scores and pictures was found significant ($r = 0.286$, $p<0.05$). When the rater and audio variables were controlled, the correlation between scores and pictures increased ($r =0.465$, $p<0.000$). $R^2 = 0.216$ value of the correlation was estimated, which means the pictures shown are responsible for the variation of the scores at about 21.6%. The analysis of the scores in the second session confirm the findings by indicating that there is a relationship between scores of raters and appearance of students.

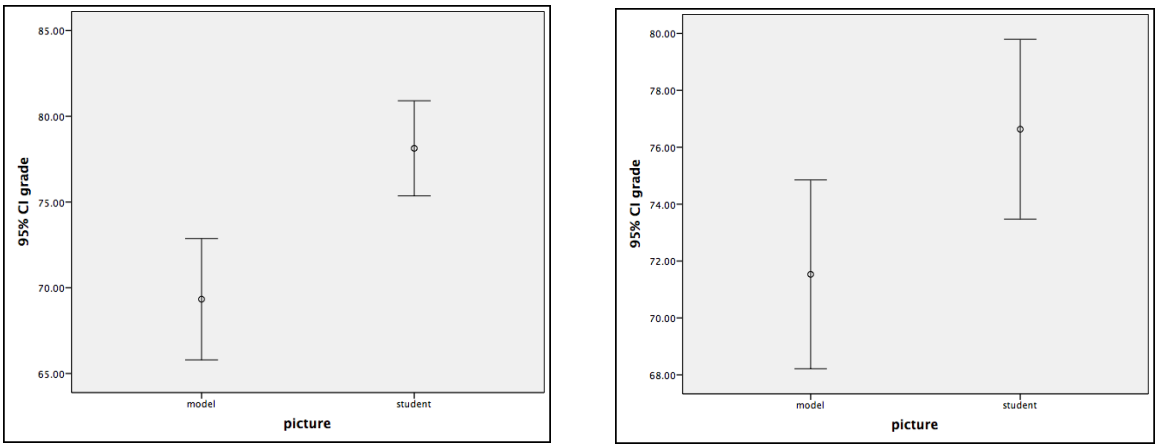
4.2 Difference of scores according to physical appearance

In order to find out whether the raters score a student's oral performance differently considering his/her physical appearance significantly, Independent Samples T-test was conducted. Actually, the descriptive statistics give clues about their different attitude in scoring.

Table 4: Descriptive statistics of scores based on pictures

Picture	Session	N	Mean	SE
Model	**1st session**	90	69.333	1.570
	2nd session	90	71.533	1.570
Student	**1st session**	90	78.133	1.570
	2nd session	90	76.633	1.570

Table 4 reveals that the mean scores of raters differed when different pictures were shown during grading. The participants favored the audios when they saw student pictures as the owner of the voice (In the first session M=78.13, and in the second session M=76.63), but their mean scores decreased when they saw model pictures as the owners (in the first session M=69.33, in the second session M=71.53). Box Plot figures confirm this result visually in Figure 2.



Session 1 *Session 2*

Figure 2: Difference of mean scores in sessions

As seen in Figure 2, means of the scores accompanied with student pictures are higher than the other group in each session. Thus, it can be assumed that female raters, especially, scored more beautiful female students more harshly while giving higher scores to the normal students. Independent Samples T-test verifies the difference found in descriptive statistics.

Table 5: Independent Samples T-test results

			Levene's Test for Equality of Variances		t-test for Equality of Means				
			F	**Sig**	**t**	**df**	**Sig(2-tailed)**	**Mean difference**	**Std. Error Difference**
Score	**Session 1**	Equal	.664	.419	-	178	.000	-8.80000	2.1982

		variances assumed			4.003			
		Equal variances not assumed			-4.003	174.86	.000	-8.80000	2.1982
	Session 2	Equal variances assumed	.003	.957	-2.276	178	.027	-5.10000	2.2408
		Equal variances not assumed			-2.276	177.866	.027	-5.10000	2.2408

Levene's Test proves that the variances of scores in both sessions were equal ($p>0.05$), therefore t-test results when equal variances assumed were taken into consideration. Significance values of t-test were found lower than 0.05 in each session. Thus, it can be inferred that the mean values of oral performance scores given by the raters differ significantly in each group based on students' physical appearance (in session 1 $t= -4.003$, $p<0.001$ and in session 2 $t= -2.276$, $p<0.05$).

4.3 Variation of scores according to the sessions held

In this study, 2 grading sessions were held. The second session was a kind of confirmation of the first session with the same raters and the same audio recordings. In this case, Paired Samples T-test was conducted to identify whether the raters' scores varied according to the sessions held.

Table 6: Paired samples statistics for sessions

Paired Samples Statistics					
		Mean	**N**	**sd**	**Std. Error Mean**
Pair 1	**Session 1**	73.73	180	9.536	1.231
	Session 2	74.08	180	8.981	1.159

Table 6 indicates that the mean values of the scores in each session are almost equal to each other (M= 73.73 and M=74.08). Paired Samples T-test result verifies the situation noticed in descriptive statistics.

Table 7: Paired samples t-test results for grading sessions

Paired Differences							
		Mean	**sd**	**Std. Error Mean**	**t**	**df**	**Sig.(2-tailed)**
Pair 1	**Session1 - Session2**	-.350	6.161	.795	-.440	179	.662

Table 7 reveals that scores given in session 1 and session 2 do not show a significant difference according to t-test result (t= -.440, df=59, p>0.05). Thus, it can be inferred that disregarding the pictures shown, the raters assigned similar scores in each session although they had a 1-month interval between the scoring sessions. Time variable is considered insignificant in this case. However, students' physical appearance created a variation on raters' scores according to the results mentioned before. Female raters, especially, scored more beautiful female students more harshly while giving higher scores to the normal students, which may indicate that beauty can sometimes be a curse.

4.4 Interview results

After the analysis of the findings, voluntary participants were invited to participate in a semi-guided interview whose focus was if this difference on student scores was intentional or not since there are cases that raters have some rough prejudices against students' characteristic features but even they themselves are unaware of this fact. 13 teachers joined the interviews and answered a set of questions which were prepared by the researchers. Table 8 reveals the questions and the responses of the participants.

Table 8: Interview results

Questions	Yes	Undecided	No	Total
Do you think that physical appearance makes a difference on raters' scoring students' oral performance?	8	4	1	13
Do you think that physical appearance makes a difference on your scoring students' oral performance?	2	7	4	13
Do you think that considering students' physical appearances in scoring their oral performances is fair?	0	1	12	13
Have you ever felt that you have been treated differently just because of your physical appearance?	4	5	4	13
Did you intentionally give lower grades to these (model) students?	8	4	1	13
Will this study affect your perceptions on scoring students' oral performances regardless of their physical appearances?	4	6	3	13

When the participants were asked whether the physical appearance makes a difference on raters' scores of students' oral performances, 8 teachers agreed, 4 felt undecided and 1 disagreed. Thus, the majority of the participants confess that, in an exam condition, the graders are influenced by the appearance of the testees. One of the teachers asserts:

"When a student enters from the door to the room in speaking exams, his or her first impression gives a little clue about how he or she can do. Generally, students who are dressed better or better-groomed receive better grades."

The participants were also asked if they are influenced by the students' physical appearance and 2 agreed, 7 felt undecided and 4 disagreed. Contrary to the statistical results explained earlier, only 2 teachers confirmed that they are affected by the physical appearance of students. 7 teachers were most probably unaware of this fact and the rest may have insisted that they are not affected although their scores did not show the same. Another teacher states that:

"I feel more positive when a better-looking student comes in the exam room."

Also, a teacher asserts that

"I've never paid attention to this issue before. Maybe I was affected, I don't know".

However, another teacher rejects that she was affected by physical appearance:

"Of course it is not ethical. I try not be influenced by the outlook of students."

Another question in the interview asked whether considering students' physical appearances in scoring their oral performances is fair or not. Almost all of the participants agreed that it is not fair to assign scores based on the students' appearances. One of the participants mentions:

"So why bother with exams? We could only look at the pictures and classify students easily as passed or failed. But it is not ethical and does not reflect the actual performance of the students and it also raises a question about validity of the exam we use."

When the teachers were asked whether they have been exposed to a similar treatment because of their own physical appearance, the number agreeing and disagreeing responses were the same. Some of the teachers assert:

"No, I don't think I was judged with my appearance",

"In a job interview I felt that I was rejected because of my appearance. There were much prettier girls waiting in front of the interview room."

After showing their grades on particular group, the teachers were asked whether they intentionally gave lower grades to better-looking students. Interestingly, eight of the teachers agreed and one disagreed. One of the participants' states:

"I deliberately gave a lower score to this student because she wears a heavy make-up and looks as if she has nothing to do with learning English."

Another teacher asserts:

"She looks overly daring and I am sure she gives her teacher a hard time during the lessons."

Finally, when teachers were asked whether this study will affect their perceptions on scoring students' oral performances, almost half of the participants were undecided and the number of the ones who agreed and disagreed were similar. In fact, one of the undecided teachers claimed that this situation is the psychological reflection of the phenomenon and it would be impossible to control it:

"This is something psychological. I subconsciously gave these grades. I think it is quite impossible to change or control these feelings. I understood that I was affected by the physical appearance of the students but I cannot guarantee that I won't be influenced next time."

Another teacher stated:

"It was a kind of surprise. I always thought I was an objective teacher. I should be more careful or critical about the scores in the next exam."

5. Discussion & Conclusions

The aim of this study is to investigate whether the participants (speaking raters) assess students more harshly or leniently because of their physical appearance although they use the same scoring rubric. First, it was found that raters gave higher scores to the audios accompanied by student pictures. On the other hand, no correlation between raters, audios and scores was found. When the rater and audio variables were controlled, the correlation between scores and pictures increased. $R^2 = 0.304$ value of the correlation was estimated, which means the pictures shown are responsible for the variation of the scores at about 30.4%. In other words, apart from what students say or

how well they use the target language, they were graded not only because of their language performance but by their appearance as well. Similar to the results reached in the first grading, no correlations between raters, audios and scores were identified, whereas the correlation between scores and pictures was found significant in the second grading. When the rater and audio variables were controlled, the correlation between scores and pictures increased. $R^2 = 0.216$ value of the correlation was estimated, which means the pictures shown are responsible for the variation of the scores at about 21.6%. The analysis of the scores in the second session confirm the findings by indicating that there is a strong relationship between scores of raters and appearance of students. In order to make it clear that the raters score a student's oral performance differently considering his/her physical appearance, an Independent Samples T-test was conducted. It was again seen that the mean scores of raters differed when different pictures were shown during grading. The participants favored the audios when they saw student pictures as the owner of the voice, but their mean scores decreased when they saw model pictures as the owners. Significance values of t-test were found lower than 0.05 in each session. Thus, it can be inferred that the mean values of oral performance scores given by the raters differ significantly in each group based on students' physical appearance.

In this study, 2 grading sessions were held, therefore, a question of rater reliability may born in minds that could be a possible reason of this difference among raters' scores assigned for two groups. The second session indeed was a kind of confirmation of the first session with the same raters and the same audio recordings. That is why, Paired Samples T-test was conducted to identify whether raters' scores vary according to the sessions held. It was found that the mean values of the scores in each session are almost equal to each other. To illustrate, scores given in session 1 and session 2 do not show a significant difference according to t-test result ($t= -.440$, $df=59$, $p>0.05$). Thus, it can be inferred that disregarding the pictures shown, the raters assigned similar scores in each session although they had a 1-month interval between the scoring sessions. Time variable is considered insignificant in this case. However, students' physical appearance created a variation on raters' scores according to the results mentioned before. Female raters, especially, scored more beautiful female students more harshly while giving higher scores to the normal students, which may indicate that beauty can sometimes be a curse. What is more, it was found that more experienced teachers score the model pictures more harshly and score the student pictures more leniently than the less experienced raters and there was a significant difference between the experienced and less experienced teachers' scores in this sense ($p>0.05$) although they used the same rubric and evaluated the same recordings.

Finally, the results of the interview confirmed the statistical findings and drew a more vivid picture of the phenomenon, appearance matters. Contrary to the common belief that "What is beautiful, that is favored.", female graders favor the less attractive and charming girls and they are mostly aware of this fact. Having such a bias that if a girl wears a strong make up and dresses more care-taking than her friends, they assume that she studies less than the others and this idea drives them perform more critically while grading. Is this fair? Most of them know that this is not fair and may sometimes mislead them while judging student performance, however they know that most of the data they use while grading are based on their personal feelings, assumptions and experiences and this is really hard to change in one or two days or after such a study.

To conclude, even if it is generally ignored and not accepted by most of the teachers, students appearance and teacher beliefs play an important role in performance assessment. The reason may vary; it could be the gender bias or the appearance of the testee which was mainly underlined in this study. It is commonly agreed that to take raters' beliefs, personal judgments out of performance evaluation is almost impossible so what is logical and possible to is to raise awareness against this issue and let the raters see how they perform against different variable while scoring and individual performance. In the end no one would tolerate evaluating someone's ability more harshly or leniently just because of his/her sex, origin, religion, ethnicity, sexual preference, color or nationality. This would be not only a problem of validity and reliability of the test but a failure of the whole educational system which would be utmost fair to anyone who seeks for equality and justice.

References

1. Akay, E. & Toraman, C. (2015). Students' attitudes towards learning English grammar: A study of scale development. *Journal of Language and Linguistic Studies*, 11(2), 67-82.
2. Aydin, B., Akay, E., Polat, M. & Geridonmez, S. (2016). Türkiye'deki Hazırlık Okullarının Yeterlik Sınavı Uygulamaları ve Bilgisayarlı Dil Ölçme Fikrine Yaklaşımları. *Anadolu University Journal of Social Sciences.* 16(2), 1-19.
3. Bachman, L.F. (2004). Gender Bias in the Classroom. *Journal of legal education.* 23(4), 137-146.
4. Banks, L. B. (1998). Teacher Cognition in Grammar Teaching: A Literature Review. *Language Awareness, 12*(2), 96-108.

5. Berger, J., Hammit, F., Norman, R. & Zelditch, M. (1977). *Status Characteristics and Social Interaction: An Expectation State Approach.* NY: Elsevier Scientific Pub. Co. Inc.
6. Boyce, M.W. (1979). Physical attractiveness- a source of teacher bias? *Australian Journal of Teacher Education*. 4(1), 34-44.
7. Centra, J. & Gaubatz, N. (2000). Is there gender bias in student evaluations of teaching? *The journal of higher education.* 70(1), 17-33.
8. Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice,* 31(3), 10-20.
9. Crocker, L. & Algina, J. (1986). *Introductıon to Classical and Modern Test Theory*. CBS Collage Publishers Canpany. USA.
10. Cronbach, L.J. (1995). Giving method variance its due. In D.T. Gilbert & S.T. Fiske (Eds) *Personality research methods and theory (145-157).* Hillsdale, NJ. Lawrence Ebaum Ass. Inc.
11. Dennis, I. & Newstead, S. (1990). Blind marking and sex bias in student assessment. *Assessment and evaluation in higher education.* 15(2), 132-139.
12. Eagly, A. & Mladinic, A. (1994). Are people prejudiced against women? Some answers from research on attitudes, gender, stereotypes and judgements of competence. *European Review of Social Psychology.* 5(1), 1-34.
13. Eckes, T. (2005). Examining rater effects in TESTDAF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly.* 2(3), 197-221.
14. Francis, B., Robson, J. & Read, B. (2001). An analysis of undergraduate writing styles in the context of gender and achievement. *Studies in Higher Education.* 26(3), 313-326.
15. Harding, S. (1991). *Whose science? Whose knowledge?* Buckingham, Open University Press.
16. Hedge, T. (2000). *Teaching and learning in the language classroom.* Oxford, England: Oxford University Press.
17. Hoyt, W.T. (2000). Rater bias in psychological research: When is it a problem and what can we do about it? *Psychological Methods.* 4(2), 403-424.
18. Johnson, J. & Lim, G. (2009). The influence of rater language background on writing performance assessment. *Language Testing.* 28(4), 485-505.
19. Kenyon, D.M. (1992). *Introductory remarks at symposium on development and use of rating scales in language testing.* Teachers Forum, Columbia University.
20. Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly.* 12:239-261.

21. Kondo, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing.* 19(1), 3-31.
22. Krawczyk, M. (2017). Do gender and physical attractiveness affect college grades? *Assessment & Evaluation in Higher Education*. *DOI: 10.1080/02602938.2017.1307320*
23. Langlois, J.H., Kalakanis, L., Rubenstein, A.J., Larson, A., Hallam, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychol Bull.* 26(3):390-423.
24. Murphy, P. & Elwood, J. (2002). Constructions of achievement and the positioning of students: a gender perspective. *Pedagogy, Culture and Society.* 10(2), 134-152.
25. Myword, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many facet rasch measurements: Part1. *Journal of Applied Measurement.* 4, 386-422.
26. Newstead, S. (1996). The psychology of student assessment. *The Psychologist: Bulletin of the British Psychological Society*. 9, 543-547.
27. Newstead, S. & Dennis, I. (1990). Blind marking and sex bias in student assessment. *Assessment and Evaluation in Higher Education*. 15, 132-139.
28. Polat, M. (2017). Teachers' attitudes towards teaching English grammar: A scale development study. *International Journal of Instruction.* 10(4):379-398. DOI: 10.12973/iji.2017.10422a
29. Read, B., Francis, B. & Robson, J. (2005). Gender, bias assessment and feedback: analysing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education.* 30(3), 241-260.
30. Reed, D.J. & Cohen, A.D. (2001). Revisiting rater and ratings in oral language assessment. *Studies in Language Testing11: Experimenting with uncertainty (pp82-96).* Cambridge, UK.
31. Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, 25(4), 465-493.
32. Talamas, S., Mavor, K. & Perrett, D. (2016). Blinded by beauty: Attractiveness bias and accurate perceptions of academic performance. *Plos One.* 11(2), 1-18.
33. Umberson, D. & Hughes, M. (1987). The Impact of Physical Attractiveness on Achievement and Psychological Well Being. *Social Psychology Quarterly.* 50(3), 227-236.
34. Wigglesworth, G. (1993). Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction. *Language Testing.* 10(3), 305-319.