Hacettepe Üniversitesi Eğitim Fakültesi Dergisi (H. U. Journal of Education) 32(3): 581-599 [2017]
doi: 10.16986/HUJE.2017027583

Raters’ Knowledge of Students’ Proficiency Levels as a Source of Measurement Error in Oral Assessments[*]

Konuşma Sınavlarında Bir Ölçme Hatası Kaynağı Olarak Notlandıranların Öğrencilerin Dil Seviyelerini Bilmesi

Fatma TANRIVERDİ-KÖKSAL[**], Deniz ORTAÇTEPE[***]

• ***Received:*** *02.08.2016* • ***Accepted:*** *01.05.2017* • ***Published:*** *31.07.2017*

ABSTRACT: There has been an ongoing debate on the reliability of oral exam scores due to the existence of human raters and the factors that might account for differences in their scorings. This quasi-experimental study investigated the possible effect(s) of the raters' prior knowledge of students' proficiency levels on rater scorings in oral interview assessments. The study was carried out in a pre- and post-test design with 15 EFL instructors who performed as raters in oral assessments at a Turkish state university. In both pre- and post-tests, the raters assigned scores to the same video-recorded oral interview performances of 12 students from three different proficiency levels. While rating the performances, the raters also provided verbal reports about their thought processes. The raters were not informed about the students' proficiency levels in the pre-test, while this information was provided in the post-test. According to the findings, majority of the Total Scores ranked lower or higher in the post-test. The thematic analysis of the raters' video recorded verbal reports revealed that most of the raters referred to the proficiency levels of the students while assigning scores in the post-test. The findings of the study suggest that besides factors such as accent, nationality, and gender of the test-takers and the assessors, raters’ prior knowledge of students' proficiency levels could be a variable that needs to be controlled for more reliable test results.

Keywords: rater effects, intra-rater reliability, paired oral exams, think-aloud protocols

ÖZ: Yaygın olarak kullanılmakta olsa da notlandıran olarak insan faktörünün varlığı ve notlardaki farklılığa neden olan etmenler sebebiyle konuşma sınav notlarının güvenirliği konusunda süregelen bir tartışma vardır. Bu yarı deneysel çalışma, konuşma sınavlarının değerlendirilmesinde, not verenlerin öğrencilerin dil yeterlilik seviyelerini önceden biliyor olmasının verdikleri notları üzerindeki etkilerini araştırmayı amaçlamaktadır. Bu çalışma, Türkiye’deki bir devlet üniversitesinde yabancı dil olarak İngilizce öğreten ve aynı üniversitede konuşma sınavlarında notlandıran olarak görev alan 15 okutman ile ön ve son test olarak iki oturumda yürütülmüştür. Hem ön hem de son testte, notlandıranlar üç farklı seviyeden 12 öğrencinin video kaydına alınmış aynı konuşma sınavı performansları için not vermiştir. Aynı zamanda, performanslar için not verirken, notlandıranlar eş zamanlı olarak ne düşündükleri ile ilgili sözlü bildirimde bulunmuştur. Öğrencilerin dil yeterlilik seviyeleri ile ilgili ön testte herhangi bir bilgi verilmezken, notlandıranlar öğrencilerin seviyeleri konusunda son testte sözlü ve yazılı olarak bilgilendirilmiştir. Sonuçlara göre, önteste kıyasla son testte Toplam Notların büyük çoğunluğunun son testte düştüğü veya yükseldiği saptanmıştır. Tüm notlandıranların video kayıtlı sözlü bildirimleri tematik olarak incelendiğinde, notlandıranların çoğunun son testte not verirken öğrencilerin dil yeterlilik seviyelerine değindikleri gözlemlenmiştir. Çalışmanın sonuçları, not verenlerin ve sınava giren adayların aksan, uyruk ve cinsiyet özellikleri gibi etkenlerin yanısıra, daha güvenilir test sonuçları için, not verenlerin adayların dil yeterlilik seviyelerini önceden biliyor olmasının kontrol edilmesi gereken bir değişken olduğunu ileri sürmektedir.

Anahtar sözcükler: notlandıran etkisi, tek notlandıran güvenirliği, eşleştirilmiş konuşma sınavları, sesli düşünme protokolleri

[*] The article is based on Master’s Thesis “The effect of raters' prior knowledge of students' proficiency levels on their assessment during oral interviews” (Tanrıverdi-Köksal, 2013).

[**] Instructor, Bülent Ecevit University, School of Foreign Languages, Zonguldak-Turkey, fatmatanriverdi@gmail.com

[***] Assoc. Prof. Dr., Bilkent University, Graduate School of Education, MA TEFL, Ankara-Turkey, deniz.ortactepe@bilkent.edu.tr

1. INTRODUCTION

With the growing popularity of the communicative theories of language teaching in the 1970s and 1980s (Brown, 2004), oral interviews have taken its place in academic contexts as an alternative but also a very controversial assessment instrument to evaluate students' spoken proficiency. Oral interviews are widely used in proficiency tests which are conducted to determine whether learners can be considered proficient in the language or whether they are proficient enough to follow a course at a university (Hughes, 2003). Since the testing of spoken language to assess communicative competence is subject to raters' interpretations (e.g., Bachman, 1990) and rating differences (Ellis, Johnson, & Papajohn, 2002), concerns about reliability and fairness have been at the center of the discussion on oral interviews (Caban, 2003; Elder, 1998; Hughes, 2003; Joughin, 1998).

While four main types of rater effects (i.e., halo effects, central tendency, restriction of range, and leniency/severity) are discussed in much detail in the literature (e.g., Lumley & McNamara, 1995; Myford & Wolfe, 2000), the construct-irrelevant factors, the factors other than the actual performances of test-takers that affect raters' behaviors, scoring process, and final scorings, have not been completely explored (Kang, 2012; Stoynoff, 2012). The differences in rater behaviors in terms of leniency/severity toward a particular performance have led researchers to look at another aspect of fair scoring: bias which is an important concept in language testing since test results should be "free from bias" (Weir, 2005, p. 23). McNamara and Roever (2006) define bias as "a general description of a situation in which construct-irrelevant group characteristics influence scores" (p. 83). In other words, bias in assessment refers to an unfair attitude toward test takers by either favoring or disadvantaging them. As a result, low reliability and rater bias in oral interviews can highly affect the decisions made about the test-takers' performances and lead to raters' misjudgments about the test-takers' performances, and thus, prevent raters from assigning fair and objective test results.

As Fulcher and Davidson (2007) suggest, in oral assessments, for which subjective scoring of human raters is at the center of the debate, the attempts to control the construct-irrelevant factors are crucial in order to provide and guarantee fairness in large-scale testing. One way to provide more consistent scoring is the use of a validated appropriate rubric (Hughes, 2003) which provides explicit and thorough instruction for the raters on how to assess the students' performances in terms of what to expect and what to focus on. Yet, sometimes even though the rubrics used are appropriate for the goals of the tests, raters may behave differently both in their own scoring processes and from each other while conducting the interviews, interacting with the test-takers and assessing the test-takers' performances. As a result, if raters are affected by some construct-irrelevant factors during the rating process, it is highly possible that they can misjudge the performance of test-takers which can lead to the misinterpretation of scores (Winke, Gass & Myford, 2011). In other words, rater measurement error, that is, "the variance in scores on a test that is not directly related to the purpose of the test" (Brown, 1996, p.188), can result in a lower score than a test-taker really deserves, which in some cases even lead to failing a test. Considering the fact that human raters may sometimes yield to subjectivity in their ratings (Caban, 2003), investigating rater effects in oral interviews is of great importance for accurate assessments as the results of inaccurate judgments may have harmful effects for test-takers, raters, and the institutions.

1.1. Research on Rater Effects in Speaking Assessment

Rater effect, rater error, rater variation, and rater bias usually refer to the same issue: the change in rater behaviors depending on various factors other than the actual performance of test-takers. Several studies have been conducted to find out how personal and contextual factors affect interlocutors’ and raters’ behaviors and decisions in assessments, and how these factors can be controlled to eliminate or limit the human rater factor in scores (Fulcher & Davidson, 2007).

Previous studies have investigated rater effects on oral test scores from different perspectives such as the raters’ educational and professional experience (e.g., Chalhoub-Deville, 1995), raters’ nationality and native language (e.g., Chalhoub-Deville & Wigglesworth, 2005; Winke & Gass, 2012; Winke et al., 2011), rater training (e.g., Lumley & McNamara, 1995; Myford & Wolfe, 2000), and the gender of candidates and/or interviewers (e.g., O’Loughlin, 2002; O’Sullivan, 2000). For instance, Lumley and McNamara (1995) examined the effect of rater training on the stability of rater characteristics and rater bias whereas MacIntyre, Noels, and Clément (1997) examined bias in self-ratings in terms of participants’ perceived competence in an L2 in relation to their actual competence and language anxiety. O’Loughlin (2002) and O’ Sullivan (2000) looked into the impact of gender in oral proficiency testing, while Caban (2003) examined whether raters’ linguistic background and educational training affect their assessments. Chalhoub-Deville and Wigglesworth (2005) investigated if raters from different English-speaking countries had a shared perception of speaking proficiency while Carey, Mannell, and Dunn (2011) studied the effect of rater’s familiarity with a candidate’s pronunciation. Although there are several studies agreeing on the role of the raters’ beliefs, perceptions and bias in affecting test results, defining those factors that might influence rater judgment is still in its exploratory stage; and to the knowledge of the researchers, no study has focused on whether raters’ prior knowledge of students’ proficiency levels may have an impact on their assessment behaviors in oral interviews.

Joe (2008), emphasizing the complex procedural and cognitive processes the raters go through while assigning scores in performance assessments, suggests that human scoring involves two important principles: “what raters perceive and how raters think” (p. 4). For this reason, due to the fact that statistical approaches fail in providing a full understanding of the decision making process, recent studies have started to show interest in applying cognitive processing models (e.g., think-alouds) in order to gain better insights into how raters assign scores, and why there are differences among raters’ scorings (Brown, 2000). As a qualitative data collection method, verbal report analysis has two types: (a) concurrent verbal reports, also referred to as think-alouds, are conducted simultaneously with the task to be performed, and (b) retrospective verbal reports are gathered right after the performance task (Ericsson & Simon, 1980). Think-aloud protocols are considered as more effective in understanding raters’ cognitive processing during oral assessment scoring because it is sometimes difficult to remember what someone did and why he/she did it (Van Someren, Barnard, & Sandberg, 1994). While employing think-aloud protocols for understanding how raters assign a score can shed light on the rater effects in oral assessment, there is still a limited body of research focusing on think-aloud in oral interview assessments (e.g., Joe, Harmes, & Hickerson, 2011; Orr, 2002).

As discussed above, during the rating process, if raters are affected by some factors other than the actual performances of test-takers, it is highly possible that they can misjudge the performance of test-takers which can lead to the misinterpretation of scores (Winke et al., 2011). Moreover, given that paired oral interviews are also widely used in educational settings to assess learners’ spoken proficiency, due to the performance-irrelevant factors, a student can get a lower score than he/she deserves, or even worse, fail in the test. Since assessment scores should be free from bias and should reflect the actual performance of test takers, exploring the construct-irrelevant factors has recently received much attention in the literature. Therefore, the present study aims to investigate whether raters’ prior knowledge of students’ proficiency levels clouds their judgments about the actual performances of test takers and influences their scores.

The overarching research question addressed in this study is:

- To what extent does raters' prior knowledge of students' proficiency levels influence their assessment behaviors during oral interviews?

2. METHOD

2.1. Setting and Participants

This study focusing on intra-rater reliability in oral interview assessments was conducted at a Turkish university which provides intensive English courses to undergraduate students for one year. The students are required to pass the proficiency exam administered at the end of the academic year in order to pursue their studies in their departments. The rationale for choosing this school is twofold. First, as this was one of the researchers' home institution, it provided convenience sampling to the researchers; second, being one of the few public universities that administer oral interviews as part of their proficiency exam, this institution records and saves these oral interviews in their archives for research purposes.

The participants of this study were 15 (Female=10, Male=5) Turkish instructors who teach English as a foreign language (EFL). These EFL teachers also perform as raters in the oral interviews conducted as part of the proficiency exam in the same institution. The participants vary in the length of their teaching and scoring experience. They were chosen on a voluntary basis, and they were regarded as a representative sample as the total number of instructors working at this university is about 50. (See Table 1 for demographic information of the participants).

Table 1: Demographic information of the participants

Background Information	N (15)	%
Gender		
Female	10	66.66
Male	5	33.33
Teaching Experience		
1-5	8	53.33
6-10	6	40
11+	1	6.66
Scoring Experience		
1-5	13	86. 66
6-10	2	13.33

The variable that is under scrutiny in this study is raters' knowledge of students' proficiency levels. In the institution where the study is conducted, the proficiency level of students is determined according to the results of the proficiency and placement tests which the students are required to take at the beginning of the academic year. While the proficiency exam is administered to decide whether the students are proficient enough to take classes in their departments or should study at the one-year intensive English preparatory program, the placement test is administered to those students who fail in the proficiency exam in order to place them in the appropriate level where students with the same language competency will study. In the institution where the study is conducted, English instruction is offered at three different levels, namely, D, C, and B levels (that is, A1 level, A1+ level, A2 level, respectively) from the lowest to the highest based on the framework proposed according to the Common European Framework of Reference (CEFR). The main course book taught in the institution, a commercial product used worldwide, was based on CEFR, and the three series of the book were developed for A1, A2, and B1 levels. In order to continue their majors, the students are required to take the proficiency exam at the end of the academic year and receive a grade that corresponds to A2 level, which is the minimum exit level in the institution. At the beginning of the academic year when the study was conducted, the students enrolled in this institution started with A1, A1+, and A2 levels, and after a one-year intensive English instruction, their exit levels were supposed to correspond to A2, A2+, and B1 levels.

2.2. Research Design

The data were collected in three sessions: (a) the norming session held to inform the participants about the study, receive their consent, collect demographic information, and achieve standardization for scoring, (b) the pre-test in which the raters were asked to assign scores without the knowledge of the students' proficiency levels, and (c) the post-test in which the information about students' proficiency levels was provided for the raters without making them aware of the actual purpose of the study. The raters were informed that the students' levels were written in the post-test grading sheet because some raters asked for that information in the pre-test. Both in the pre- and post-test, think-aloud sessions were held during which the raters' verbal reports were gathered (See Figure 1 for the study's procedure).

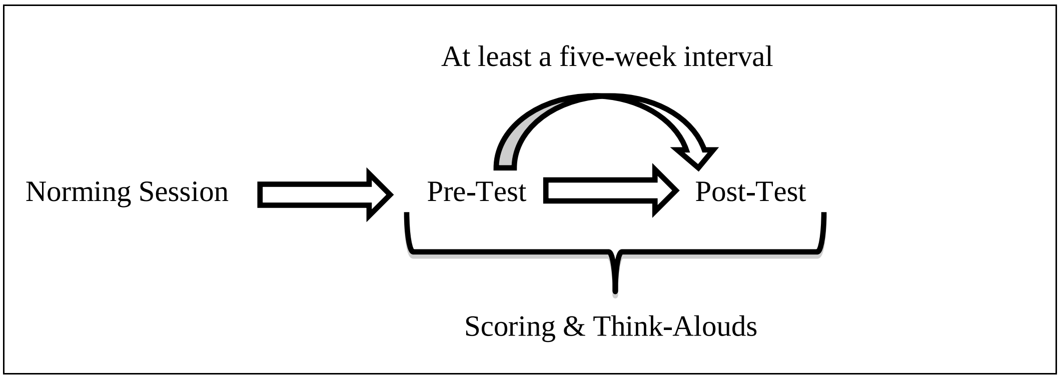

Figure 1. The procedure of the study.

This study adopted a mixed method quasi-experimental research design which combines both quantitative and qualitative data collection techniques. While the scores assigned by the raters for each student's oral interview performance served as quantitative data, the raters' concurrent verbal reports provided during think aloud protocols contributed as qualitative data to gain better insights into the raters' thought processes.

2.3. Data Collection Instruments

Data collection instruments consisted of two sets of data sources; (a) scores from the pre- and post-test, and (b) concurrent verbal reports (think-aloud protocols). The first set of data source, students' oral interview scores were gathered during the pre- and post-test conducted with at least five weeks interval. The scores were collected from raters under two conditions; first, raters' having no information about students' proficiency levels, and then, raters' being informed about students' proficiency levels both in written format and orally.

The second set of data, concurrent verbal reports (Think-aloud protocols), included approximately a five-minute-verbal report of raters during which they commented on each student's performance while assigning scores. Because both the researchers and the participants are native speakers of Turkish, the raters were asked to provide their verbal reports in Turkish so that they would feel more comfortable and provide more data. The raters' verbal reports were video-recorded, and in total, for pre-tests and post-tests nearly one-hour of data was gathered from each rater, which added up to nearly 15 hours of recordings. Both sets of data indicate raters' evaluations of and judgments about students' spoken task performances in relation to the categories of the rating scale.

2.3.1. Rating materials

Rating materials consisted of (a) video recordings of oral interview performances of 12 students conducted as a part of 2011-2012 academic year proficiency exam, (b) the rating scale used by the raters while scoring students' performance, and (c) the grading sheets for raters to fill while assigning scores. The same video recordings were used for the pre- and post-tests. Six video recordings which were recorded during oral interview sessions conducted in 2011-2012 academic year proficiency exam were chosen as the rating material. The students in the video recordings were placed at three different levels according to their placement test scores at the beginning of the academic year, but they were paired randomly at the final oral proficiency exam since the aim of the exam was to assess whether they were proficient enough to continue their departmental studies Thus, they can be paired with a same level, low level, or higher level student. The length of each video was approximately seven minutes excluding the pauses/silences. Each video included an oral interview session of two preparatory school students performing two tasks, one individually with the guidance of the interlocutor, and one interacting with another student. In total, oral interview videos of 12 students with different proficiency levels (D/C/B levels, i.e., A1, A1+, A2, from the lowest to the highest proficiency levels) were used. There were four B level students, two C level students, and six D level students, and the students were randomly paired, either with a same-proficiency-level candidate or with a higher or lower proficiency level student.

In this study, as the videos included the video-recorded oral performances of the students at the speaking component of the final proficiency exam at the institution where the study was conducted, the raters used the same analytic rubric developed by the Speaking Office of the institution. The rubric included five components which are *Fluency and Pronunciation, Vocabulary, Grammatical Range and Accuracy, Task Completion* and *Comprehension.* For each component, the lowest score that can be assigned is 1 point while the highest score is 4 points. As a *Total Score,* the raters can assign 5 points as the lowest score to a very poor performing student while the students with a successful performance can receive up to 20 points.

Two grading sheets developed by the researchers were used by the raters while assigning scores in the pre- and post-tests. Although the same information was provided in the two forms (i.e., students' pseudo IDs, the tasks they performed, and the categories of the rating scale), the proficiency levels of students were only presented in the grading sheet used for the post-test. Moreover, in order to investigate whether the raters were familiar with any of the students, a section that asks whether the raters taught or knew the students was included in both sheets. The data gathered from those raters familiar with any of the students were not included in the data analysis.

2.4. Procedure

Once the participants were informed about the study, for standardization, two pre-selected video recordings which were not the ones used in the actual study were rated by the participants using the analytic rubric. No training was provided about the rubric since the raters were already familiar with it; yet, the components and the descriptors of the rubric were discussed very briefly. Once the raters assigned scores for Video #1, they were asked to reveal their scores for each student in relation to the five components of the rubric and the *Total Score*. The scores were presented on the board in order to show the inconsistencies among the raters, and to explain the possible reasons for the inconsistencies. The same procedure was followed for scoring Video #2.

In the pre-test scoring session, first, the raters were informed about think-aloud protocols, and they practiced scoring and providing verbal reports for one-preselected video which was not one of the six videos used as rating materials. After this practice session, the raters, first, watched one video, and then, provided verbal reports while scoring the students' performances. The same procedure was followed for each of six videos. One of the researchers was present from the beginning to the end of the procedure as an observer. The researcher did not interfere with any part of the verbal reports unless there was a long pause or the raters were likely to assign scores without verbalizing what they were thinking. The raters were not allowed to go back to the videos, rewind or forward it due to the fact that they are not able to go back to the speech samples of students during an actual oral performance assessment. The order of the videos was assigned randomly for each rater in order to prevent future problems such as raters' discussing about the videos with other participants although they were requested not to, and the order of the videos presented to the same rater was different in the pre- and post-tests in order to minimize the possible recall effect. The same procedure was followed in the post-test scoring session which was conducted with at least a five-week interval. In the post-test, the proficiency level of each student was written in the post-test grading sheet, and the raters were told that some raters asked for this information as this information was provided to the raters on the exam sheet in actual assessments in that institution.

2.5. Data Analysis

First, the data collected via ratings were analyzed in Statistical Package for Social Sciences (SPSS, version 21). Wilcoxon Signed Ranks Test was run for each rater's assigned scores in the pre- and post-test in order to determine if there was a statistically significant difference between the scores assigned without the knowledge of students' proficiency levels (pre-test) and with that knowledge (post-test). The scores assigned by each rater were analyzed separately to see whether there was a significant difference between their pre- and post-test scores in the aspects of five categories of the rubric which are *Fluency and Pronunciation, Vocabulary, Grammatical Range and Accuracy, Task Completion, and Comprehension*, as well as in the *Total Scores*. Further analysis was also carried out with the rating data to investigate if the raters had a bias towards students with a specific proficiency level. Second, the qualitative data gathered from think-aloud protocols were transcribed verbatim and analyzed with content analysis by using the existing framework of the rubric as well as the themes that emerged from the data such as proficiency. While assessing the students' performances and providing verbal reports for why they were assigning those scores, the raters had the tendency to follow the order of the components in the rubric. Thus, transcribing the data using the framework of the rubric and focusing on their references to theme under scrutiny, i.e., proficiency levels of the students, were completed successfully.

3. FINDINGS

The results of Wilcoxon matched pairs signed rank test indicated a statistical difference between the scores assigned by eight raters (#1, #3, #4, #6, #8, #11, #14, and #15). As shown in Table 2, each rater behaved differently while assigning scores to the different components of the rubric. While some raters assigned different scores only in one component of the rubric, some raters assigned higher or lower scores in more than one. The *Vocabulary* and *Grammatical*

Range and Accuracy were the components of the rubric in which the raters showed significant differences the most frequently while only one rater (Rater #1) behaved differently in the *Total Scores* component.

Table 2: The components of the rubric showing a statistically significant difference

Raters	Fluency & Pronunciation	Vocabulary	Grammatical Range & Accuracy	Task Completion	Comprehension	Total Score
#1	.025*	.007*			.038*	.011*
#3			.046*			
#4	.025*	.034*				
#6				.020*		
#8			.025*		.046*	
#11			.025*			
#14		.025*				
#15				.008*		

*p<.05

Although the results indicated a significant difference only in one rater's pre- and post-test *Total Scores*, when the descriptives of the pre- and post-test *Total Scores* assigned to individual students were analyzed, it was observed that the majority of the scores assigned by the 15 raters, including Raters #2, #5, #7, #9, #10, #12, and #13, changed in the post-test as higher or lower *Total Scores* (See Table 3).

Table 3: Comparison between the pre- and post-test for the total scores

Raters	Negative Ranks*	Positive Ranks**	Ties***	Scorings Included	Scorings Excluded
#1	0	8	1	9	3
#2	6	4	2	12	0
#3	5	5	2	12	0
#4	6	2	2	10	2
#5	2	3	5	10	2
#6	2	4	5	11	1
#7	4	3	4	11	1
#8	6	1	4	11	1
#9	6	3	3	12	0
#10	4	5	3	12	0
#11	8	2	2	12	0
#12	6	4	2	12	0
#13	5	5	1	11	1
#14	6	1	4	11	1
#15	2	8	2	12	0
TOTAL	68	58	42	168	12

* post-test scores < pre-test scores

** post-test scores > pre-test scores

*** post-test scores = pre-test scores

As Table 3 shows, while negative ranks demonstrate that there was a decrease in the assigned scores, positive ranks show that the raters assigned higher scores in the post-test. Although 42 *Total Scores* (25 %) did not change in the post-test, 58 scores increased (35 %) while 68 scores

(40 %) decreased. In other words, 75 % of the *Total Scores* assigned by these 15 raters ranked lower or higher in the post-test, ranging from one point difference to more than 10 points. As discussed by Myford and Wolfe (2000), one point may not seem like or be considered as a large difference, but it can have an important effect for the test takers whose scores are around borderline/pass score.

Figure 2 below presents the results about the raters’ behavior in terms of (a) whether there was a statistically significant difference between their pre- and post-test scores, and (b) whether they referred to the proficiency levels of the students in their think aloud protocols.

Rater #	***Significant Difference***	***Reference to the levels***
1	YES	YES
2	▨	YES
3	YES	YES
4	YES	YES
5	▨	▨
6	YES	YES
7	▨	YES
8	YES	YES
9	▨	YES
10	▨	YES
11	YES	▨
12	▨	YES
13	▨	▨
14	YES	YES
15	YES	▨

Figure 2. The existence of a significant difference in raters’ scorings and/or reference to the proficiency levels in their verbal reports.

For more in-depth analysis, the verbal reports of the raters in relation to the scores they assigned will be presented in the next section.

3.1. Raters with Statistically Significant Difference between their Scorings

The results indicated that eight raters, Raters #1, #3, #4, #6, #8, #11, #14, and #15, did not show consistent scoring behaviors within themselves in different sections of the rubric. When the think aloud protocols were analyzed, it was seen that among these eight raters, six of them (Raters #1, #3, #4, #6, #8, #14) referred to the proficiency levels of the students, while Raters #11 and #15 did not refer to the levels of the students in their verbal reports.

In terms of the leniency/severity towards the students within the same proficiency level, each of these six raters who referred to the proficiency levels of the students behaved differently. Most of the scores assigned in the post-test ranked higher or lower, but more severity was observed in B level students’ *Total Scores*. It is also interesting that while some raters changed their scores when they referred to the levels of students, some raters were consistent in their scorings and comments (See Figure 3).

Raters	**Student No**	**Student Level**	**Partner's No & Level**	**Component of the rubric**	**Pre-test score & comment**	**Post-test score & comment**	**Pre-test *Total Score***	**Post-test *Total Score***
#1	# 2	C	#1 - D	*Vocabulary*	(2) *The student used very limited vocabulary*	(3) *It is clear that the student is a C level student and her vocabulary use/range was not bad*	(12)	(15)
#3	#4	B	#3 - B	*Fluency and Pronunciation*	(4) *The student had some problems with the pronunciation of some words, but she was good, she had no hesitations in terms of fluency.*	(3) *She had some hesitations, wrong pronunciation for some vocabulary, but in general, she could deliver the message if we consider they are B level students.*	(20)	(18)
#4	#12	D	#11- B	*Total Score*	(15) *She was better in the first task. They are also influenced by each other, by the structures and the vocabulary they used.*	(11) *I did not see a big difference between them. D level student needs to practice a lot.*	(15)	(11)
#6	#12	D	#11 - B	*Vocabulary*	(3) *Cough, headache. Good appropriate vocabulary.*	(4) *Good vocabulary such as get stressed, cough, it was good considering she is a D Level student.*	(15)	(19)
#8	#11	B	#12 - D	*Total Score*	(16) *His partner was a little better than him especially in the first task, so she got 2 points higher than him.*	(18) *He was more fluent, enthusiastic. We should also consider that this student is a B level student, and the other one is a D level student.*	(16)	(18)
# 14	#12	D	#11 – B	*Total Score*	(17) *They were both good, they were not very fluent, they did not speak comprehensively, but they are in the production phase.*	(18) *She was successful considering she is a D level student. She had good sentences and used appropriate vocabulary. They were not bad, they were fairly average students.*	(17)	(18)

Figure 3. Examples from raters #1, #3, #4, #6, #8, #14.

As seen in Figure 3, raters' scores and perceptions about the students' performance changed depending on their expectations from a student with a certain level language proficiency. It is interesting to see that most of the changes were observed in the scores of Students #11 and #12 who were paired with a high-level student and a low-level student, respectively.

Although significant differences were observed in their scorings, Raters #11 and #15 did not refer to the proficiency levels of the students while assigning scores. In terms of the leniency/severity towards the students within the same proficiency level, similar to the six raters whose data were discussed, Raters #11 and #15 behaved differently, and most of the scores assigned by these raters also ranked higher or lower in the post-test. Figure 4 shows some extracts from the pre- and post-test verbal reports of Raters #11 and #15 in relation to the scores they assigned.

Raters	Student No	Student Level	Partner's No & Level	Component of the rubric	Pre-test score & comment	Post-test score & comment	Pre-test Total Score	Post-test Total Score
#11	#11	B	#12 - D	*Grammatical Range and Accuracy*	(4) *Good. No big mistakes, some minor errors.*	(3) *Some minor errors, but they did not obscure meaning*	(18)	(16)
#15	#2	C	#1 - D	*Task Completion*	(2) *The first task was difficult. She started appropriately, but could not continue. In the second task, she usually continued the dialog, but while asking questions, he did not ask relevant questions.*	(3) *Especially the topic of the first task was difficult. Although she did not deal with the topic comprehensively, she did her best. In the second task, she tried to interact, communicate, but her partner was not active, enthusiastic, so she had some problems here.*	(13)	(15)

*Figure 4. Examples from raters #11 and #15***.**

As seen in Figure 4, although Raters #11 and #15 did not refer to the proficiency levels of the students, differences were observed in their post-test scores and perceptions about the success of the students' performance. Moreover, task difficulty and the performance of the candidate's partner were the themes that emerged frequently in Rater #15's verbal reports.

3.2. Raters with No Statistically Significant Difference between their Scorings

Seven raters, #2, #5, #7 #9, #10 #12, and #13, showed no statistically significant differences in their post-test scores, yet five of them referred to the proficiency levels of the students in their verbal reports. Although the results indicated no significant difference for these five raters (#2, #7, #9, #10, and #12), a majority of the students received lower *Total Scores*. Also, the number of positive ranks was greater than the equal scores. Further analysis was conducted in order to see how different rankings each proficiency level of students received in their *Total Scores* assigned by these five raters. It was found that while some raters were more lenient towards the higher level students, some were more severe towards lower level students. In general, most of the lower level students received more severe scorings. While Rater #2 assigned both negative and positive ranks for C and B level students, she was more severe in her scorings for D level students. Rater #7 was slightly more severe towards C level students; however, there was no strong pattern in the scores he assigned in terms of leniency/severity towards a specific level. Rater #9 mostly assigned lower scores in the post-test. While Rater #10's scores for two out of three C level students increased, he assigned equally lower and higher *Total Scores* for the other levels. However, D and B level students received more negative ranks rather than positive ranks from Rater #12, while, out of three C level students, the *Total Scores* of two students ranked higher. When all the scores were considered, the results indicated that the number of lower, equal and higher scores assigned to B level students were almost the same, but the scores of C level students changed the most in terms of negative or positive ranks. Out of 15 scorings assigned for C level students, only two did not change. Moreover, half of the scorings assigned to the D level students ranked lower in the post-test. The qualitative analysis of the verbal reports by these five raters revealed that they referred to the proficiency levels of some students while assessing the oral performances of the students (See Figure 5).

Rater No	**Student No**	**Student Level**	**Partner's No & Level**	**Component of the rubric**	**Pre-test score & comment**	**Post-test score & comment**	**Pre-test Total Score**	**Post-test Total Score**
#2	#2	C	#1 - D	*Vocabulary*	(4) *The student was very excited in the first task. The second task was very good, she asked all the questions and used all the necessary words. She used connectors such as unfortunately.*	(2) *Although she is a C level student, she was very excited and had limited vocabulary range, the word "unfortunately" is the only the word range we can see.*	(18)	(8)
#7	#2	C	#1 - D	*Vocabulary*	(2) *She could tell her ideas only by using adjectives.*	(4) *The student had adequate range for this level of the student.*	(10)	(15)
#9	#10	B	#9 - D	*Total Score*	(13) *Her partner was better. She was less successful compared to her partner, but in pair work, it was obvious that this was a pair work, they asked questions to each other.*	(10) *Her partner continued the conversation although he was a D level student. She was passive although she was a B level student, she performed less successfully than her partner.*	(13)	(10)
#10	#12	D	#11 - B	*Vocabulary*	(3) *She did not use sophisticated words, but did not have errors.*	(3) *She used basic words, but she could accomplish what was expected of her. She used words appropriate to her level. She had problems in grammar, her vocabulary use was not very bad.*	(14)	(13)
#12	#11	B	#12 – D	*Vocabulary*	(4) *No problem, very good. He used the connectors effectively.*	(4) *He used appropriate words according to his level.*	(19)	(17)

Figure 5. Examples from raters #2, #7, #9, #10, and #12.

As seen in Figure 5, these raters referred to the levels of the students and assigned lower or higher scores in the post-test. As discussed before, there was no pattern about how different attention each proficiency level received from the raters, but most of the raters referred to the proficiency levels of the same two students, Student #11 and #12 who were a B and a D level matched-pair. In other words, the highest proficiency level and the lowest proficiency level matched-pair received utmost attention from the raters.

4. DISCUSSION

When the pre- and post-test *Total Scores* assigned by the 11 raters were investigated in terms of their degree of leniency/severity towards lower and higher proficiency level students, it was observed that the raters behaved differently when the information about students' proficiency levels was provided in the post-test. While Raters #2, #8, #9, and #12 assigned lower *Total Scores* for D level students, Rater #1 was more severe in her scorings. For C levels, while Raters #4, #8, and #14 were more severe, Raters #1, #3, and #6, assigned more favorable scores in the post-test. B level students received harsher scorings from Raters #3, #4, #9, #12, #14 while Raters #1 and #13 were more lenient towards B level students. Since no reference to the levels was found in four raters' verbal reports, the results are inconclusive for these raters either because the measurement error was random or there was "incompleteness due to synchronization problems" (Van Someren et al., 1994, p. 33). In other words, the variable in the post-test, raters' knowledge of students' proficiency levels did not affect their scorings or these raters may not have verbalized what they thought exactly, so there may be some missing data in their reports due to the difference between the pace they thought and they spoke (Van Someren et. al, 1994). On the other hand, 11 raters who referred to the proficiency levels of the students assigned higher or lower post-test *Total Scores* to individual students when the information of the students' proficiency levels was provided in the post-test. The raters' comments presented in Figure 3 and 5 suggest that they assigned scores to the students' performances by considering their proficiency levels. Some raters also assessed the performances by referring to what each level could achieve in terms of the curriculum they were taught (e.g., *Figure 5,* Rater #7's comments for Student #2).

There may be several reasons for why each rater perceived the performances of the students differently in the pre- and post-test and so differed in their interpretations of the students' performances and degree of severity by assigning lower or higher post-test scores. The types of rater effects on scores such as halo effect, central tendency, restriction of range, and leniency/severity (Saal, Downey, & Lahey, 1980) could explain the rater variance observed in the present study.

First, the knowledge of the students' proficiency levels might have caused a halo effect. In other words, the raters may have assigned scores with a global impression of each test-taker rather than distinguishing his/her different level of performances in different aspects of the assessment (Saal et al., 1980). For example, for Student #10, Rater #9 assigned scores to the components of the rubric from both lower and higher bends and a *Total Score* of 13. His comments were *"The student was less successful compared to her partner, but in pair work, it was obvious that this was a pair work, they asked questions to each other."* However, in the post-test, he mostly assigned scores from lower bends adding up to 10 points as a *Total Score*. His post-test comments were *"Her partner continued the conversation although he was a D level student, but this student was passive although she was a B level student, and she performed less successfully than her partner."* As seen in the example, when the information about the student's proficiency level was available in the post-test, the rater had higher expectations from a B level student and assigned lower scores for *Task Completion* and *Comprehension* in the post-test. In short, the students' poor or better performance in one aspect may have affected the judgment of the raters if they considered the proficiency levels of the students while assigning scores.

Second, the central tendency which refers to "raters' reluctance to make extreme judgments" (Saal et al., 1980, p. 417), and the restriction of range, that is, raters' overusing certain bends in each category of the rubric (Myford & Wolfe, 2003) may have an effect on the differences in their scorings. Raters might have considered the students' levels and what scores other raters would assign. Although they did not report such considerations verbally, novice raters or raters who did not want to stand out may have yielded to the effect of central tendency and the restriction of range. For example, for Student #10, Rater #1 assigned the lowest point possible (5 points) as a *Total Score* in the pre-test and commented: *"The student's performance was very bad, she could not speak at all."* However, in the post-test, the rater assigned 13 points as a *Total Score* stating *"Although the student is a B level student, she could not speak and could not do the task."* As seen in the example, the rater assigned the lowest score in the pre-test, but her score in the post-test was around the midpoint which might be the effect of rater's considering that the student might receive higher scorings from other raters because she is a B level student. Since the raters were aware that the data provided from all raters would be analyzed, there is a possibility that, even if they used the lowest or the highest bends in the pre-test, they assigned scores around midpoint in the post-test in order not to differ from the other raters' in terms of their degree of leniency/severity. Also, raters may have avoided assigning scores from the highest bends for lower levels and scores from the lowest bends for higher levels considering the proficiency levels of the students and what scores other raters might assign for these students.

Variations have been observed also in the leniency/severity of the raters in their post-test ratings, a phenomenon that can be explained by criterion based assessment. In other words, the raters might have assessed the performances according to the curriculum taught during the year. Although all the students took the same proficiency exam, the content of the instruction provided in the institution differs for lower levels and higher levels. This may have affected the raters' judgments in two ways; first, appreciating the efforts of lower-level students, and second, due to their higher expectations from a higher level student, disgracing their performances when compared to lower level students. For a C level student, Student #2, Rater #7 assigned 2 points for the pre-test *Vocabulary* saying "*She could tell her ideas only by using adjectives," and* 10 points as the *Total Score* reporting "*The student was nervous in the first task, so she could not speak much. In the second task, although she had errors in her sentences, she told her ideas."* However, a favorable judgment was observed in the post-test. The rater assigned 4 points for *Vocabulary* pointing out *"The student had adequate vocabulary range for this level of student,"* and 15 points as the *Total Score* commenting *"The student tried, but her sentence constructions were problematic, so even if she had a better performing partner, I don't think she can express herself well, still she completed her tasks."* Yet, the reverse was also observed for higher proficiency levels because of the higher expectations as shown in Rater #2 where she showed severity in her scorings for Student #2 since she considered that C level is a higher level than her partner's D level. In the pre-test, she assigned 18 points as a *Total Score* for Student #2 reporting *"She was excited in the first task, but she could formulate some sentences. It could be better. The second task was very successful, she asked all the questions and used all the necessary words. She initiated the conversation and it was very effective."* Yet, a great degree of severity was observed in her post-test scorings and verbal reports when the information about the students' levels was provided. The rater considered the level of the student as a higher level compared to her partner, and she assigned 8 points as a *Total Score* commenting *"Although she was a C level, the student was very excited. She had limited vocabulary range and grammar errors even in simple sentences which obscured the meaning. She had lots of pauses, so she had problems in fluency."* As a result, raters cannot be directly compared in terms of the degree of severity they exercise when scoring, but the knowledge of the students' proficiency levels seems to have affected each rater's degree of leniency/severity to some extent.

In this study, the students were paired randomly without considering their proficiency levels, and although very few took the exam with a same-proficiency-level student, most of the pairs included students with different proficiency levels. The analysis of verbal reports revealed that some raters compared the performances of the two candidates taking the exam together as pairs by referring to their levels and assigning scores accordingly. This comparison might have an effect on the changes of the scores because some raters assigned scores in the post-test considering the performances and the proficiency levels of the candidates and their partners. For instance, Rater #8’s scorings and verbal reports for a pair, Students #11 (D level) and #12 (B level) indicate that Rater #8 assigned 16 points as a *Total Score* for Student #11 and 18 points for Student #12 commenting *“Student #11’s partner was a little better than him, especially in the first task, so she got 2 points higher than him.”* However, when the information about students’ proficiency levels was provided in the post-test, Rater #8 assigned 18 points for Student #11 and 16 points for Student #12 stating *“Student #12’s partner was more fluent, enthusiastic. Generally, female students are more excited. They were successful. We should also consider that this student is a D level student, and the other one is a B level student.”* As a result, even though the proficiency level might not be a variable on its own, when combined with pairs from different levels, it does seem to influence raters’ judgments.

In conclusion, the findings of the present study concur with the previous studies by confirming that raters may be affected by factors other than the actual performance of the test-takers (e.g., Chalhoub-Deville, 1995; Chalhoub-Deville & Wigglesworth, 2005; Lumley & McNamara, 1995; Myford & Wolfe, 2000; Winke & Gass, 2012). Whether random or systematic, similar to the other studies, measurement error was observed in this study underlining the influential factors that may cause disagreement within and/or among the raters’ judgments in oral performance assessments. In light of the findings of the present study, it can be argued that the raters’ prior knowledge of students’ proficiency levels could be an important factor that may cloud raters’ judgments and affect their scoring behaviors during oral interview assessment especially in proficiency exams; thus, jeopardize the assurance of the two important qualities of a good test: reliability and fairness (Bachman & Palmer, 1996; Kunnan, 2000).

5. IMPLICATIONS, LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

The results of this study revealed that the knowledge of students’ proficiency levels is one of the factors that may impact the results of the assessment, the reliability of the institutions, and academic and personal lives of the students. For this reason, some recommendations can be made for the institutions to minimize the effects of the construct irrelevant factors on the scorings. There are some implications already suggested in the literature to avoid rater bias. First of all, the most commonly accepted suggestions to increase rater reliability and fairness include rater training (e.g., Brown, 2004; Hughes, 2003; Lumley & McNamara, 1995; Myford & Wolfe, 2000), using multiple raters (Council of Europe, 2001; Hughes, 2003), using a validated appropriate rubric (Hughes, 2003), introducing the rubric to the raters in detail (Bachman, 1990), and providing the same explicit and thorough instruction for all raters on how to assess the students’ performances in terms of what to expect and what to focus on. As for the implications that the present study suggests, in light of the assessment behaviors of the raters both during the norming sessions and in the exams, first, rater profiles should be created in order to investigate whether the raters are severe or lenient assessors by nature and to inform the raters about their scoring performances. Then, since using multiple raters as assessors is highly suggested in the literature (Council of Europe, 2001; Hughes, 2003), raters should be paired according to their profiles created. In terms of fairness, it is better to match a severe rater with a lenient one instead of having two severe or lenient assessors for the same test-taker. Since paired interviews are widely used, the candidates may be asked to interact with a professional interlocutor rather than with a fellow candidate, but the advantages and disadvantages of using this format should be considered thoroughly (Hughes, 2003). More importantly, any information about the candidates that can lead to subjective scorings should not be provided to the raters either by the candidates or the institutions (Hughes, 2003), and the raters should only base their judgments on the performances of the test takers and the rubric they use (Council of Europe, 2001).

There are several limitations to this study suggesting that the findings should be treated with caution. Initially, although great care was taken in order to create similar assessment conditions, there is a chance that the raters may not have behaved in the way they usually assign scores since they were aware that their scorings and verbal reports would be analyzed by the researcher. However, this conscience may also have led them to try to be more cautious and objective while assigning scores. Furthermore, although all the raters have had teaching and assessment experience of oral skills for at least one year, they did not receive any professional training for oral assessment and they were not certified raters. However, despite the limitations, this study augmented the literature on rater effects by revealing that raters' prior knowledge of students' proficiency levels in speaking exams can serve as a construct-irrelevant factor that can cloud raters' judgments and affect their scores.

Some suggestions can be made for further research. To begin with, this study can be replicated in another setting or with participants from different institutions and backgrounds to reach more generalizable findings. The number of the raters who assign scores and the number of students whose performance are assessed can be increased. Secondly, the study can also be replicated with a treatment and a control group. While the information about the students' proficiency levels can be provided to the treatment group in the post-test, no information can be given to the control group in order to analyze if there is a significant difference between their scorings.

6. CONCLUSION

The findings of the study are in accordance with the literature which suggests that the construct-irrelevant factors can influence the assessment of the raters and the scores of the test-takers in oral interviews (e.g., Chalhoub-Deville & Wigglesworth, 2005; Myford & Wolfe, 2000; O'Loughlin, 2002; O'Sullivan, 2000; Winke & Gass, 2012; Winke et al., 2011). Several factors that affect raters' scorings in oral interviews have been studied in the literature; however, to the knowledge of the researchers, no study has been conducted to investigate the effects of the raters' prior knowledge of the students' proficiency levels on their scoring behaviors during proficiency exams oral interviews. Therefore, this study might augment the literature by revealing another source of rater effects in oral interviews assessment. To conclude, it is hoped that the findings of the study and the pedagogical implications discussed above will help all the stakeholders gain insight into the importance of minimizing any external factor that may jeopardize the reliability and the fairness of the scorings assigned for the test takers.

7. REFERENCES

Bachman, L. F. (1990). *Fundamental considerations in language testing*. New York: Oxford University Press.

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

Brown, A. (2000). An investigation of the rating process in the IELTS oral interview. *IELTS Research Reports*, *3*, 49-84.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Longman.

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

Caban, H. L. (2003). Rater group bias in the speaking assessment of four L1 Japanese ESL students. *Second*

Language Studies, 21(2), 1-44.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing, 28*(2), 201-219.

Chalhoub-Deville, M. (1995). Deriving oral assessment scales across different tests and rater groups. *Language Testing, 12*(1), 16-33.

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes, 24*(3), 383-391.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Elder, C. (1998). What counts as bias in language testing? *Papers in Language Testing and Assessment, 7,* 1- 42.

Ellis, R. O. D., Johnson, K. E., & Papajohn, D. (2002). Concept mapping for rater training. *TESOL Quarterly, 36*(2), 219-233.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment.* London, New York: Routledge.

Hughes, A. (2003). *Testing for language teachers* (2nd Ed.). Cambridge: Cambridge University Press.

Joe, J. N. (2008). *Using verbal reports to explore rater perceptual processes in scoring: An application to oral communication assessment* (Unpublished doctoral dissertation). James Madison University, VA, USA.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. *Assessment in Education: Principles, Policy & Practice, 18*(3), 239-258.

Joughin, G. (1998). Dimensions of oral assessment. *Assessment & Evaluation in Higher Education, 23*(4), 367-378.

Kang, O. (2012). Impact of rater characteristics and prosodic features of speaker accentedness on ratings of international teaching assistants' oral performance. *Language Assessment Quarterly, 9*(3), 249-269.

Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida* (pp. 1-14). Cambridge: Cambridge University Press.

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71.

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-satings of second language proficiency: The role of ranguage anxiety. *Language Learning, 47*(2), 265-287.

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. MA & Oxford: Blackwell.

Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English Assessment System. TOEFL Research Report 65*.NJ: TOEFL Research Program, Educational Testing Service.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

O'Loughlin, K. (2002). The impact of gender in oral proficiency testing. *Language Testing, 19*(2), 169-192.

Orr, M. (2002). The FCE speaking test: using rater reports to help interpret test scores. *System, 30*(2), 143-154.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373-386.

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413-428.

Stoynoff, S. (2012). Research agenda: Priorities for future research in second language assessment. *Language Teaching, 45*(02), 234-249.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes*. London: Academic Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. New York: Palgrave Macmillan.

Winke, P., & Gass, S. (2012). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. *TESOL Quarterly, 47*(4), 762-789.

Winke, P., Gass, S., & Myford, C. (2011). The relationship between raters' prior language study and the evaluation of foreign language speech samples. *TOEFL iBT® Research Report*. NJ: Educational Testing Services.

Uzun Özet

İngilizce dili öğretiminde iletişimsel yöntemlerin vurgulanmasıyla beraber konuşma ve yazma becerilerinin ölçme ve değerlendirilmesi daha da önem kazanmıştır. Ancak bu becerilerde aynı yazma/konuşma performansının değerlendirmesinde notlandıranlar arası farklılıklar olduğu gözlemlendiği için, notlandıran olarak insan faktörünün varlığının sınavların geçerliği ve güvenirliği açısından büyük bir risk oluşturduğu ileri sürülmektedir. Yazma sınavları değerlendirilmesinde, öğrenci ismi kapalı notlandırma gibi tedbirlerin alınması mümkünken, literatürde, konuşma sınavlarında birçok faktörün adayın performansından bağımsız bir şekilde not verenlerin değerlendirmesini etkilediğini ortaya konmuştur.

Adayların eşleştirilerek ikili görüşme şeklinde yürütülen sözlü mülakatlar akademik amaçla konuşma becerisinin ölçme ve değerlendirilmesinde yaygın olarak kullanılmasına rağmen, notlandıranların belli bir derece sübjektif not verme eğilimi (Caban, 2003) bu sınav formatının kullanılıp kullanılmaması konusunda geçerlik-güvenirlik tartışmalarına sebep olmaktadır (Joughin, 1998).

Notlandıran etkisi üzerine yapılmış çeşitli çalışmalar göstermiştir ki not verenler farklı sebeplerden dolayı not verme davranışlarında değişiklik sergilemektedir. Not verenin eğitim geçmişi ve iş deneyimi, not verenin veya adayın uyruğu ve anadili, not verenlerin notlandırma eğitimi alıp almadığı, aldıysa hangi ortamda, nasıl içerikte, ne kadar süreyle eğitim aldığı, not verenin ya da adayın cinsiyeti gibi faktörlerin adayın performansından bağımsız bir şekilde not verenin değerlendirmesini etkileyebileceği farklı çalışmalarda ileri sürülmüştür (Chalhoub-Deville, 1995; Chalhoub-Deville ve Wigglesworth, 2005; Lumley ve McNamara, 1995; O'Loughlin, 2002; O'Sullivan, 2000; Winke ve Gass, 2012; Winke, Gass, ve Myford, 2011). Ancak, konuşma sınavlarının ölçme ve değerlendirme süreçlerini, notlandıranların not verme davranışlarını ve notları etkileyen performanstan bağımsız tüm faktörlerin ortaya çıkarılması henüz araştırma aşamasındadır (Stoynoff, 2012). Bu tür faktörlerin, adayların sınav sonuçlarını ve dolayısıyla akademik hayatlarını ve geleceğini de etkilediği gerçeğinden yola çıkarak, bu alanda daha çok çalışma yapılması gerekmektedir. Bu alanda yapılan çalışmaların az olma sebebi çoğunlukla not verenlerin not verme esnasında karar verme süreçlerini gözlemleme imkânının olmamasıdır. Bu sebeple, notlandıranların notlandırma esnasında ne düşündüğünü ortaya çıkarmak adına sesli-düşünme protokollerinin kullanıldığı çalışmaların azlığı (örn., Joe, 2008; Joe ve diğerleri, 2011; Orr, 2002) son zamanlarda literatürde önemle vurgulanmaktadır.

Bu yarı deneysel çalışma, dil yeterlilik sınavlarında sözlü mülakatların değerlendirilmesinde, olası notlandıran önyargısını ve notlandıranların öğrencilerin dil yeterlilik seviyelerini önceden biliyor olmasının verdikleri notlar üzerinde var ise etkilerini araştırmayı amaçlamıştır. Bu amaçla, çalışmanın uygulandığı Türkiye'deki bir devlet üniversitesinin Yabancı Diller Yüksekokulu'nda, yabancı dil olarak İngilizce öğreten ve aynı üniversitede sözlü sınavlarda notlandıran olarak görev alan 15 okutman ile ön ve son test olarak iki oturumda yürütülmüştür. Çalışmanın veri toplama sürecinde arşiv kayıtları ve çeşitli materyaller kullanılmıştır; araştırmacı, aynı üniversitede 2011-2012 akademik yılı muafiyet sınavı esnasında kaydedilmiş altı videoyu notlandırma materyali olarak seçmiştir. Bu videoların her biri ikili olarak eşleştirilmiş öğrencilerin sözlü performansını içermektedir. Toplamda 3 farklı seviyeden 12 öğrencinin kaydı notlandırma için kullanılmıştır. Veri toplama, notlandıranların iki ekstra videoda kayıtlı dört öğrencinin performansına verdikleri notların standardizasyon için tartışıldığı norm belirleme oturumu ile başlamıştır. Norm belirleme oturumundan sonra, katılımcılar analitik bir ölçek kullanarak arasında en az beş hafta olan ön test ve son testte bireysel olarak notlandıran görevini üstlenmişlerdir. Hem ön hem de son testte, 12 öğrencinin performansını içeren aynı 6 video kaydı kullanılmıştır. Hem ön hem son teste, notlandıranlardan üç farklı seviyeden bu 12 öğrenci için performanslarını gösteren video kayıtlarını izleyerek not vermelerini ve aynı zamanda not verirken ne düşündükleri ile ilgili sesli düşünme protokolleri ile sözlü bildirimde bulunmaları istenmiştir. Öğrencilerin dil yeterlilik seviyeleri ile ilgili ön testte herhangi bir bilgi verilmezken, notlandıranlar öğrencilerin seviyeleri konusunda son testte sözlü ve yazılı olarak bilgilendirilmiştir. Veri analizi için notlandıranların verdikleri notlar dosyalanmış, sesli-düşünme protokolleri video kaydına alınmıştır.

Sonuç olarak, ön ve son test notlarının nicel veri analizi, sekiz notlandıranın, kullanılan ölçeğin *Kelime, Anlama,* ya da her öğrencinin aldığı son notu temsil eden *Toplam Not* gibi farklı bölümlerinde verdikleri ön ve son test notları arasında istatistiksel olarak anlamlı bir fark olduğunu göstermiştir. 15 notlandıran tarafından her bir öğrenci için verilen *Toplam Notların* daha detaylı incelenmesi, ön test notlarına kıyasla, notlandıranlar tarafından verilen *Toplam Notların* % 75'inin, bir puandan 10 puandan fazlaya kadar çeşitlilik göstererek, son testte düştüğü veya yükseldiği, fakat % 25'inin değişmediği saptanmıştır. Tüm notlandıranların sesli düşünme protokolleri-sözlü bildirimleri, verdikleri notlar ve öğrencilerin dil yeterlilik seviyelerine değinmeleri ile bağlantılı tematik olarak incelendiğinde, 11 notlandıranın son testte not verirken öğrencilerin dil yeterlilik seviyelerine değindiği gözlemlenmiştir. Ayrıca, her biri farklı bir dil yeterlilik seviyesinden oluşan her bir öğrenci grubu için verilmiş Toplam Notlar incelenmiş ve sonuçlar notlandıranların düşük veya yüksek dil yeterlilik seviyesi öğrencileri için not verirken, hoşgörü ve katılık derecesi açısından farklılık gösterdiğini ortaya çıkarmıştır.

Dil yeterlilik sınavlarının sözlü mülakatlarında not veren etkisinin incelendiği bu çalışmada, notlandırma ortamı ve materyalleri her ne kadar gerçeğe eş yaratılmaya çalışılsa da notlandıranların, notlarının ve sözlerinin analiz edileceğini bilmesinin not verenlerin normalden farklı notlandırma davranışları sergilemesine sebep olmuş olabilir. Ancak, bu bilincin onları aynı zamanda olabildiğince dikkatli ve objektif not verme eğilimine yöneltmiş olması da muhtemel. Ayrıca, bu çalışmadaki not veren okutmanlar her ne kadar kendi kurumlarında notlandıran olarak aktif rol alsa da hiçbiri sertifikalı notlandıran değil.

Ancak, bu çalışmanın sınırlılıklarına rağmen, dil yeterlilik sınavlarında sözlü mülakatların değerlendirilmesinde, not verenlerin adayın yeterlilik seviyesini önceden biliyor olmasını notu etkileyen performanstan bağımsız bir faktör olarak ortaya çıkarması açısından bu çalışma notlandıran etkisi literatüre katkı sağlamıştır. Konuşma sınavlarının notlandırılması esnasında sesli-düşünme protokollerinin uygulanarak benzer çalışmalar yapılması önemle tavsiye edilmektedir.