Applied Research on English Language

V. 11 N. 2 2022
pp: 1-24
http://jare.ui.ac.ir

DOI: 10.22108/ARE.2021.129400.1744
Document Type: Research Paper

Factors Affecting Untrained Raters' Cognition in Rating Oral Proficiency of English Language Learners: Insights from Iran

Nasimeh Nouhi Jadesi [1*], **Alireza Ahmadi** [2], **Seyyed Ayatollah Razmjoo** [3]

[1] *Assistant Professor of TEFL, Department of English Language Teaching, Faculty of Literature and Humanities, Salman Farsi University of Kazerun, Shiraz, Iran*
[2,3] *Professor of TEFL, Department of Foreign Languages, Faculty of Literature and Humanities, Shiraz University, Shiraz, Iran*

Received: 2021/07/04 **Accepted: 2021/11/28**

Abstract: Score decision-making is largely an undocumented process in performance assessment. To conduct a more in-depth cognitive study in scoring, one must ask if these underlying processes can be identified efficiently and objectively. To this end, the present study attempted to shed some light on how Iranian teachers as untrained raters rate speech samples of learners and how their cognition functions in the decision-making process in terms of the scores they assign. A series of monologues were obtained from a group of language learners; afterward, English language teachers were asked to rate them. The raters were asked both to assign a score and provide comments regarding why they assigned a specific score. Having rated the samples, the raters were individually interviewed. The results of the recorded interviews and the comments they had provided on scores were subjected to qualitative analysis like coding and extracting both idiosyncratic and shared features of the raters' cognition. The results revealed that some of the factors attended to by the raters were both linguistic and relevant to speaking proficiency construct like fluency, accuracy, and complexity. Some other factors influencing the raters while rating were non-linguistic and not directly related to speaking proficiency construct like the tone of voice, personality feature of the testee, etc. It seemed that the untrained raters did not have a clear definition of oral proficiency construct. The implications of the study for rater training programs have been discussed.

Keywords: Untrained Raters, Rater Cognition, Speaking Proficiency, Monologue, Iran.

* **Corresponding Author.**

Authors' Email Address:
[1] **Nasimeh Nouhi Jadesi** (nnoohij1985@gmail.com), [2] **Alireza Ahmadi** (ar.ahmadi55@gmail.com), [3] **Seyyed Ayatollah Razmjoo** (arazmjoo@rose.shirazu.ac.ir)

Introduction

The scores provided by the raters on productive skills (speaking and writing) may be used for high stake decisions in areas like education, career opportunities, etc. Nonetheless, being human, raters may inevitably be biased, inaccurate, or inconsistent in their ratings due to a large number of factors. This has been referred to as rater effects in assessment literature. Based on Wolfe and McVay (2012), rater effects can be defined as "patterns of ratings that contain measurement errors" and can thus lead to issues regarding validity in human scores (p. 32).

Although rater effects do influence the raters' judgments, they cannot be easily detected in the scores granted (Orr, 2002). The raters may give the same score on the same speaking task using the same criteria and still have completely different interpretations and perceptions of the performance they have rated. Hence, it is vital to explore the validity of such scores due to such latent variability in the raters' cognition. To better understand rater effects, and hopefully, to tackle them, describing rater cognition and cognitive processes of the raters may sound fruitful. A few rater cognition definitions have been proposed by scholars. Davis (2016) has defined it as "the mental processes occurring during scoring, at either a conscious or unconscious level" (p. 9). Having reviewed the studies on rater cognition, Bejar (2012) proposes that two aspects of rater cognition have been investigated: "the attributes of the raters that assign scores to student performances, and their mental processes in doing so" (p. 2). Mental processes entail the architecture of human information processing (e.g., short-term, working, and long-term memory) and the various (meta)cognitive strategies (e.g., attention, reasoning, judgment, planning, monitoring) the raters take (Han, 2016).

Rater cognition has been extensively studied in L2 writing assessment (e.g., Bachman & Barkaoui, 2010). Rater cognition in L2 speaking assessment, however, is still under-researched. Some studies have touched on some features that are deemed to affect the cognitive processes and rating behavior of the raters in L2 speaking assessment. The areas of investigation include investigating raters' language background (Wei & Llosa, 2015; Zhang & Elder, 2011, 2014), studying rater experience (Davis, 2016; Kim, 2015), and exploring rater training (Davis, 2016; Kim, 2015). The focus of this piece of research is, however, different. While most studies on raters and rating have focused on trained or experienced raters and their different features, this study has highlighted the untrained raters' rating cognition. The justification is that only a few language teachers have had formal training on rating (not

considering a few courses on testing passed at university). Hence, teachers raters as untrained raters outnumber formally trained ones to a great extent. It sounds logical and also promising to investigate untrained raters' rating cognition so as to see what factors, in effect, they are influenced by while rating oral proficiency of test-takers.

Statement of the Problem

The study was carried out in Iran where, as a partial requirement of the assessment process in public schools and language institutes, the teachers have to rate the productive skills of the learners. However, the language teachers have no formal training in rating. Most are graduates of English translation or literature, and a few are graduates of the English teaching discipline (TEFL). Although English language teachers normally have passed courses like Language Testing at the university, the number of such courses is very negligible (4 credits normally and at most), they are not professional assessment experts and they are not formally trained like international tests (IELTS, TOEFL, etc.) examiners. Hence, they are considered untrained raters here. Normally, they have to resort to their experience as the guiding source. Consequently, their cognitions as raters play some role in the decisions they make in terms of the scores they assign. This may endanger the reliability and validity of the scores assigned and also cast doubt on the value of the certificate of language proficiency granted to the language learners.

As a requirement for the training programs, in-depth data-driven studies that tap on the untrained raters' cognition in terms of the actual features that they attend to in oral assessment are essential. Trying to fill this void in the literature, this study attempts to identify factors that influence and account for raters' performance in rating oral proficiency of learners; that is, in assigning scores which factors they attend to and which features of the speech sample influence or impress them. Ignorance of such factors may lead to a limited and limiting description of untrained raters' cognition. This inadequacy may be reflected in inefficient rater training programs. As such, the main objective of this study is to identify the factors that may have been underrepresented in the literature which might in effect influence the raters in inflation or deflation of the scores they assign. A qualitative approach to data collection and analysis may serve this purpose. This understanding can be beneficial to the testing organizations, test developers, and test users.

The findings can help rating program training designers or implementers to have a perception of the actual mental processes and perceptions and work on them accordingly. A

rating training designer might already be familiar with trained raters, however, they might lack familiarity with untrained raters' rating cognition. Having an elaborate understanding of untrained raters' rating cognition can help in rater training sessions, where the trainer can not only train them on what they should heed but also caution them against what they should not pay attention to or be distracted by. To this end, a clear and deep perception of what mental processes language teachers as untrained raters go through can assist in providing better rating training programs. Hence, a focus on their rating experience, expertise, and perceptions can contribute to the rating literature. This is also in line with the professional development of teachers, with a focus on the current status of their rating cognition.

Review of the Related Literature

As mentioned earlier, the lion share of studies carried out on rating speaking samples has been done with a focus on experienced or trained raters (Brown, Iwashita, & McNamara, 2005; Sato, 2012, Ang-Aw & Chuen Meng Goh, 2011). However, a comparison between inexperienced, less experienced, novice, and nonprofessional raters has also been the subject of investigation and their performance has been compared with that of experienced, professionally trained raters with regard to their rating and applying rating strategies in several studies (Attali, 2016; Fahim & Bijani, 2011; Davis, 2016; Winke, Gass, & Myford, 2012). Different dichotomies have been used in the literature to refer to those raters who are not formally trained but are supposed to rate productive skills of learners as part of their professional obligations, namely, trained vs. untrained, novice vs. expert, professional vs. non-professional. For the purpose of this study, the term trained and untrained might do justice, which is used hereafter. In this section, some of the most recent studies on untrained rater rating are touched on.

Davis (2016) explored the differences between experienced teachers of English scores on a TOEFL iBT speaking test before and following training. The obtained scores were analyzed using multifaceted Rasch measurement and traditional measures of rater reliability and agreement. Moreover, the frequency with which exemplar responses were viewed was measured. The results of the study indicated that training resulted in both an increased inter-rater correlation and agreement and improved agreement with established reference scores. An interesting point was that additional experience gained after training seemed to have a little further effect on raters' scoring consistency, although the level of agreement with reference scores continued to increase. The author believed that these results raise questions regarding the relative contribution of scoring aids such as exemplars and scoring rubrics to desirable scoring patterns. In the same line, applying an empirical approach, Duijm, Schoonen, and

Hulstijn (2018) tried to identify professional and non-professional raters' differences in terms of relative responsiveness to fluency and linguistic accuracy in an occupational context. A Dutch L2 actor read 17 responses to a Dutch L2 exam which had been converted into four different versions manipulated for morphosyntactical accuracy and/or fluency. The obtained 68 stimuli were rated by both professional and non-professional raters. The findings revealed that the linguistically non-trained raters appeared to take advantage of the fluency improvement compared to linguistically trained raters. On the other hand, the linguistically trained raters praised morpho-syntactical improvement relatively higher than the non-trained raters. The overall findings indicated that raters with linguistic expertise tend to pay more attention to accuracy while non-trained raters tended to be attentive to fluency.

Some studies have been carried out in the context of Iran with the same focus. Bijani (2018) explored how 20 experienced and inexperienced raters who rated the oral stimuli obtained from 200 language learners differed prior to and after a training program. The findings suggested that training contributed to higher measures of inter-rater consistency and mitigates biases towards using rating scale categories. The author also argued that as rater variability is impossible to be eradicated even with the help of training, rater training procedure can be applied with the purpose of making raters more self-consistent (intra-rater reliability) rather than consistent with each other (inter-rater reliability). The inexperienced and experienced raters' quality improved after training, with inexperienced raters experiencing a higher consistency and less bias. It does not stand to logic, the author argued, to exclude inexperienced raters from rating solely because of their lack of adequate experience. The author suggested that it might be practical to recruit inexperienced raters and avoid bulky budgets on experienced raters and instead invest the same budget on inexperienced raters. Another study done in the context of Iran is the one carried out by Tajeddin, Alemi, and Pashmforoosh (2011) who investigated the performance of non-native EFL teachers' rating in terms of the criteria they consider in the rating of L2 learners' speaking performance. They also attempted to estimate the impact of a rater training program on raters' rating criteria. The findings of their study indicated that 10 common rating criteria which ranged from fluency to communicative effectiveness were used by the non-native raters. Nonetheless, the raters reconsidered the significance they previously attached to some criteria after the training sessions. The overall findings suggested that the raters were initially influenced by the traditional skills-and-components-based perspective on language proficiency, which made them lose sight of macro-level, higher-order components like fluency and organization. Both these two studies have

focused on untrained raters' performance prior to and after treatment, which is different from the focus of this study in which the status of untrained raters' rating cognition per se. with no particular training program is explored and described.

Taking writing as another productive skill into the investigation, Attali (2016) compared the performance of a group of newly-trained raters to the performance of 16 expert raters on a writing task. The findings showed a small difference in the performance of these two groups in terms of measurement properties (mean and variability of scores, reliability, and various validity coefficients, and underlying factor structure). Moreover, the results indicated that rater performance is less influenced by actual experience in rating responses, which was the main difference between the groups, but is more influenced by what raters learn during initial training and also the abilities acquired before training.

So far, the studies mentioned above have all focused on the performance of untrained raters before and after training in either speaking or writing. The present study intends to explore the untrained raters' cognition in rating oral proficiency with no specific training and as their current status. In line with this objective, the following research question guides this study:

What features do the untrained EFL raters consider in rating the oral proficiency of learners?

Methods

Design

Since the objective of this study is to deeply explore the cognition of the raters in rating the speaking proficiency of learners, the study enjoys a qualitative design. It is hoped that the qualitative approach provides a deeper and richer understanding of the phenomenon under investigation.

Participants

This study needed two different sets of participants. The first group of the participants of this study was composed of 16 BA TEFL students studying at the Salman Farsi University of Kazerun. Language learners of both genders were selected. It was attempted that students of varying proficiency levels be selected. As such, their professors' impressionistic judgment on their overall proficiency and their GPA were used as criteria for their selection, with their ages ranging from 19 to 23. Since the language learners were required to take time to perform the task, they were selected based on their volunteer participation. To secure the participants' right

to be informed of the judgment, the results of the ratings were delivered to those participants who were interested.

The second group of the participants was 32 Iranian English language teachers as untrained raters. For the purpose of this study, teachers who had formal training in scoring and rating other than their courses at the university were selected as untrained raters, that is, English language teachers who were not certified raters or examiners. The participants of this group were both male and female (16 male and 16 female) ranging from 20 to 50 years old and 3 to 24 years of experience. They were also selected based on their volunteer participation. Since the rating could be quite time-consuming, the teachers were paid to participate in the study and rate the speech samples.

Instruments

Interview

A series of in-depth semi-structured interviews were carried out with the raters about the rating in retrospect with the intention of identifying the underlying factors that affected their judgments.

The interviews were held in Persian to make sure that the raters could easily convey their opinions and that no language barrier hindered the negotiation of ideas. The researchers themselves did the interviews and recorded them. The interviews normally took about twenty to thirty minutes and were held immediately after the rating to ensure that they thoroughly remembered the experience. The researchers tried to obtain as much information as possible from the raters about the rating they had done.

Rating Sheets

The raters filled rating sheets designed for rating speaking proficiency of the speech samples obtained from the language learners. They were asked to both provide a score and comment on each monologue they hear, justifying the scores they assigned, and expressing the factors they took into account for assigning the scores.

Observation

The researchers were present in all the rating sessions, taking an observer as participant stance. This is defined by Ary, Jacobs, and Sorensen (2010) as: "researcher may interact with the subjects long enough to establish rapport but do not become really involved in the behaviors or activities of the groups" (p. 433). The researchers took notice of how actually the raters

approached and did the ratings. They also attended to some factors like, if the raters could make decisions about the scores quickly or they were hesitant, if they were attentive or distracted, their body language and, generally, if they had any problems and how they tackled them.

Assessment Task

The task used in this study was a monologue. The language learners performed a monologue on 'early marriage vs. late marriage'. This topic was selected by the researchers because it was thought to be of a general nature, familiar to the Iranians, and also interesting to the learners' discussion rising. Since the language learners were unaware of the topic before doing the task and had to improvise, they were granted five minutes to think about the topic and organize their thoughts and were also provided paper and pencil to jot down notes, if needed. They were also encouraged to freely express their ideas. No prompts were given by the researchers during the assessment task to the language learners. Their speeches were audio-recorded to be rated later by the raters.

Procedure

This study intended to examine how untrained raters actually rate the oral ability of EFL learners. Specifically, it intended to find out what factors affected their rating and how they came to a conclusion while rating. What follows presents the steps taken in carrying out the study.

What Language Learners Did

The language learners were given a monologue task. They had to talk about the topic presented to them. They had five minutes to think and were supposed to give a ten-minute speech. Each participant individually did the monologue task and was audio-recorded. The results were 16 audio files.

Having presented their monologue, the learners were kept out of contact with other learners, so that the topic was not known by other learners. The samples were recorded using a high-quality recording device to ensure that the speech samples recorded were clear enough.

What Raters Did

The untrained Iranian EFL raters were asked to rate the speech samples the learners had produced. The raters were asked to rate the participants by listening to the audio files. No scale or analytical framework was presented to them. They were supposed to rate and give a score

of one to six reflecting basic, elementary, intermediate, upper-intermediate, advanced, and mastery levels as proposed by the Common European Framework for Reference (CEFR). The raters were asked to individually rate the speech samples. The room where the raters listened to the recordings was quiet and headphones were available to those who needed them.

After listening to each audio file, the raters were given time to make decisions about the scores. The researchers were present in the rating sessions; however, they did not get involved in the rating process in order not to influence or manipulate the raters unintentionally. The raters were also required to provide comments explaining the reason for assigning a specific score.

Although they were encouraged to provide as much detail as possible in their comments, they were advised not to be obsessed with grammar and lexicon while providing comments. This advice was given so that the raters not be hindered by their English language proficiency. The comments in Persian were translated and checked by the authors and another colleague who was a PhD holder of TEFL to mitigate translation bias. As fatigue might affect the raters, they were allowed to take a break whenever they needed and do the ratings in several sessions at their convenience. The ratings took different for different raters. Some needed that the sample should be replayed several times before assigning the scores. The recordings were played to raters randomly and in different orders; that is a speech sample might have been given as the first sample to one rater and the fifth sample to the other rater. As the rating process of the previous phase finished, with no time-lapse, each rater was interviewed by the researchers.

The interviews were carried out to know how the raters actually came to the scores they assigned and what factors they attended to while rating. The following questions were asked in the interviews:

1- Did you have any specific problem with the rating?

2- What factors did you attend to in rating?

3- Did you assign a holistic score or have some pre-specified criteria to stick to?

4- Were you consistent in rating or did your criteria change?

5- Were you certain or hesitant in the ratings?

6- Did you feel a need for any scale or training or collaboration with other raters?

However, the researchers were not limited to these questions. These questions had a guiding function, used to elicit as much information as possible from the raters.

Data Analysis

The results of the recorded interviews with the raters were subjected to qualitative analysis like coding and extracting both idiosyncratic and shared features of the raters that might account for how they rated. The actual process of rating, the raters' feelings, the degree of certainty, etc. were also investigated here. The coding was carried out by each researcher independently to ensure the corroboration of the findings. The disagreements in the coding done by the researchers were discussed and resolved.

The interviews conducted and recorded were transcribed. The Grounded Theory approach (Glaser & Strauss, 1967) was used to reduce and cluster the data. Hence, both the interviews and the comments provided by the raters were subject to the basic coding or open coding which entailed reading up through the transcripts several times and extracting similar language and content. The second step was selective coding which was done with the purpose of forming the initial thematic groupings that led to thirteen features of fluency, accuracy, complexity, comprehensibility, adequacy and content, speech organization, tone of voice, personality features of the learners, effort to talk, and communication strategies, as described and exemplified in detail below. Further reading and analysis narrowed down the themes, as described in the axial coding phase of grounded theory and led to the identification of two umbrella categories of linguistic and non-linguistic factors.

Results and Discussion

Generally, there were some factors of which the raters were consciously aware. That is, they could consciously pinpoint such factors as the ones they attended to in oral assessment. These factors were directly mentioned by the raters in the interviews or comments for each score. Still, there were factors that did influence the raters but they were not consciously aware of. These factors needed a keen eye to read between the lines and extract them. Since the comments the raters provided for the learners' performance on each task were more detailed and related to specific speech samples, such factors were more identifiable in the comments. That is because they were in the actual act of rating and their mind was actively engaged in rating.

Table 1- The number and percentage of raters referring to each feature (the figures have been rounded off)

	fluency	accuracy	complexity	Adequacy an content	Comprehensibility	Speech organization	Tone of voice	Effort to talk	Communication strategy
The number of raters pointing to this feature	32	32	11	22	19	11	8	10	11
The percentage of raters pointing to this feature	~100%	~100%	~34%	~69%	~59%	~34%	~25%	~31%	~34%

Linguistic Factors

Linguistic features of the speech samples were among the very first and most features that were pointed to by nearly all the raters. Some of the features were easier for them to attend to and consider in rating and some others were less accessible to them. The most commonly referred-to features were fluency, accuracy, and complexity.

Fluency

Fluency was among the first features referred to by all the raters. They were deeply impressed by the fluency or dysfluency of the learners. They were also impressed by the rate of speech of the learners. They used the general term 'fluency' to describe a smooth flow of speech uttered at an acceptable rate. Pauses, as well, were repeatedly pointed to by the raters as a sign of dysfluency. This is in line with the definition provided for fluency in the literature. Fluency is usually measured by the rate of speech and quantity of unfilled pauses, which have been found to be significant markers of fluency (Leaper & Riazi, 2014, p.185). As some comments delineate this:

(Phrases and sentences in italic have been either mentioned by the raters in the interviews or written in the comments on the scores they assigned)

- *He doesn't speak naturally, makes so many pauses.*

- *Maybe he is s searching for ideas or… maybe he is searching for words… he makes so many pauses.*

- *He speaks like a robot.*

- *She makes so many pauses, the listener does not like to listen to her anymore*.

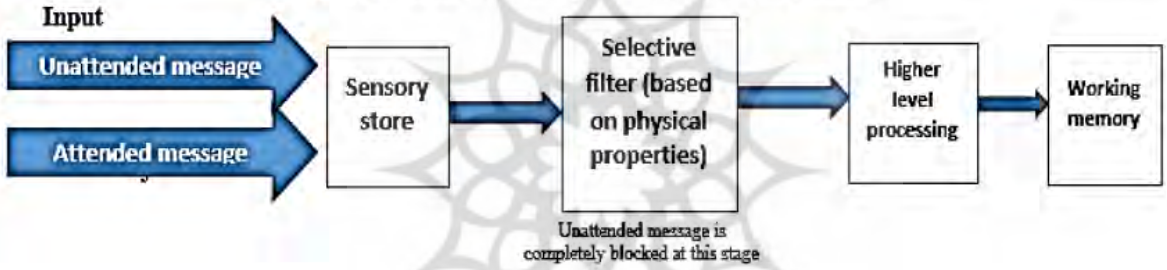
Related to the acoustic feature of speech, fluency can be described as the first feature of an oral production perceived by a listener. It has a critical role in keeping or losing the attention

or interest of a listener. A learner can put an air of proficiency by appearing fluent with the help of a smooth flow of speech and a small number of pauses even if they are not using accurate or elaborate language. That is why all the raters repeatedly pointed to fluency as a factor that they attended to in rating.

This can also be in line with the early filtering hypothesis proposed by Broadbent (1958). He argued that environmental information is filtered out of awareness if it is identified as being irrelevant to a person's current goals.

Filtering appeared to be based on superficial physical features (e.g., pitch, loudness, location, voice type, speaker rate, and tone). All higher levels of processing, such as the extraction of meaning, happen post-filter. This filtering may also take place in a rater's mind, hearing the speech samples and being primarily (if not only) impressed by the phonological features of the speech samples produced by the learners.

Fig. 1 Broadbent's (1958, p.42) filter model of attention



Accuracy

Accuracy was another feature that all the raters pointed to as what they attended to while rating the samples. Different aspects of accuracy like syntax, lexicon, pronunciation, etc. errors were attended to by the raters. However, among these, most of the comments concerning accuracy were related to pronunciation errors. Intonation, stress patterns, and pronunciation of individual sounds were factors that nearly all raters referred to.

- Repeatedly pronounces /d/ for /ð/ or /s/ for /ɵ/

- Pronunciation errors like 'advantageous' instead of 'advantages' that make a problem for meaning

- Farsi intonation

- Stress pattern is of no concern to her. She just pronounces the words the way she liked

Some of the raters were too strict with pronunciation, reducing scores for very delicate pronunciation points.

For instance, a rater went to the extreme assigning a low score for not using linking words together, since she deemed this as important.

Grammatical errors held second place in the comments on accuracy.

- Good choice of words but grammar problems

- She speaks so much and that makes her inadequacy of grammar show up. The tenses is [sic.] specifically her weak point.

- She just mixes up everything.

- I appreciate her correct use of conditionals. But, she overuses it, using it several times.

-what is he saying? Does he know anything about grammar? … 'the best important', 'a lot of confidence'

However, this attention to accuracy was not uniform across all grammatical points. That is, some errors were more severely frowned upon. The raters attended to the gravity of the error. For instance, the utterance 'getting marriage' instead of the correct form 'getting married' led to a larger score loss than misuse of a preposition.

Lexical errors were also common in the comments.

- Persian expressions and idioms translated into English like 'man of living' or 'see the empty side of the glass'

- he only uses 'good' and 'bad'. …doesn't know any other adjective.

-he has a good command of phrasal verbs. This distinguishes him from others.

The fact that pronunciation errors compared to other errors held first in attracting the attention of the raters can also be attributed to the selective attention theory explained above. Since the sounds and the acoustic features are the first features that a listener hears, intentionally or unintentionally they may attend to them more readily. Hence, the raters may be more sensitive to pronunciation errors and pinpoint this kind of error more easily.

Complexity

Of the three linguistic factors, complexity was the least referred to by the raters. That is, it was less easily accessible to them and not readily mentioned. Not all raters cared about complexity as long as the sentences were accurately and fluently uttered. Since untrained, the raters may not have been familiar with the precise definition of complexity; however, they were 'impressed' by a speech sample including more complex structures like embedding, conjunction, etc. They expressed this as:

- She used beautiful sentences not just simple ones.

- *He uses just basic elementary sentences. ….can't even connect two sentences using connectors.*

- *'And' is the only conjunction in his mind.*

This was more attended to by the raters who themselves were excessively form-focused and were concerned with using various sentence forms and complex structures, and also the ones who were either highly educated or proficient. Apparently, raters were more concerned with what they liked or were proficient in. This can be related to the mental model a rater has developed. When the environment becomes truly complex, decision-makers fail to respond appropriately by constructing new mental models. Instead, they seem to revert to older, simpler models.

Our mental models are limited, internally inconsistent, and unreliable. Our ability to understand the unfolding impacts of our decisions is poor. We take actions that make sense from our short-term and parochial perspectives, but due to our imperfect appreciation of complexity, these decisions often return to hurt us in the long run. Where the world is dynamic, evolving, and interconnected, we tend to make decisions using mental models that are static, narrow, and reductionist (Sterman, 2011, p. 11).

Non-linguistic Factors

The other group of features can be categorized as non-linguistic factors since they deal with features that are beyond the features of the very sample produced. Below, some of these features are explained and exemplified.

Comprehensibility

'Clarity and comprehensibility' was also a factor that the raters attended to. The ability to exemplify and rephrase where it was needed could denote this skill as reflected in the following comments:

- *I could easily follow her. Especially when she gave examples of her own life or her relatives.*

- *He spoke rationally and had good analysis.*

Adequacy and Content

For an utterance to have an effect, it is not enough to be accurately and fluently uttered and desirably complex. An utterance can be accurate, fluently uttered, and complex but it may not make sense. Moreover, the ideas conveyed should also be interesting to the raters. The raters

did actually attend to this feature. They assigned high scores to productions that were not just error free, but also made sense:

- Can't keep the conversation interesting

- His use of quotes is a strong point for her.

- Some cliché ideas

Still, some raters went to the extreme in taking the ideas conveyed rather than speaking ability into account in assigning scores. The excerpt of the oral data uttered by the learners (underlined part) and comments mentioned by the raters below each excerpt show this:

***Azar**: …we in our life we we should try to marriage and we should try to choose a good person*

em person that provide [sic.] every facility we want.

-Childish idea. Anyone will laugh if someone tells you that you should try to marry someone who should provide everything for you.

The excerpt below is a further example:

***Sepideh**: It depends on on you yourself, but your attitude to life. Every person wants some specific thing in life. For me the best important is my life money and work not love not marriage not anything else, just work and money*

-I would have assigned -1 if I could for her thoughts. Her view is so superficial.

This shows the fact that the untrained raters listened to the speech samples not just as a rater but as an ordinary human who takes part in daily conversations and criticizes the ideas presented by others. They may not be able to stay impartial. They might get biased by the worldviews expressed by the examinees.

Speech Organization

The raters were sensitive to the overall organization of the speech. The following comments reflect this sensitivity:

- She doesn't know what she wants to say. Someone who doesn't know what to talk about goes directly to advantages and disadvantages (of marriage).

- I liked that he first gave an introduction and talked about his own city.

- Her mind is distracted. I did not understand what her opinion was. She talked about so many things.

- The start of the conversation was too cliché. He just asked what your opinion about early marriage and late marriage

is. He could have started the conversation in a more interesting way

- Sara said: to break the ice, I start. When you actually want to break the ice, you don't mention the phrase. It's not

natural. You should just start the conversation in this way.

The raters expected a logical development and sequencing of ideas which is what they have experienced in a normal conversation in their first language. That can be attributed to the fact that they were sensitive to the need for coherence in speech.

Tone of Voice

Unconsciously, the raters were impressed and influenced by the quality of voice of the candidates. The following comments represent this issue:

- She tried to make her voice attractive

- Unclear voice

- She speaks so slowly and not energetically. I barely hear what she said. I got distracted easily several times.

- Her clear and pleasant sound made one interested to listen to what she said.

Hearing a voice that was too low, not clear, or lacked energy, distracted the attention of the raters. Such voice qualities may hinder a rater from paying attention to delicate features like complexity or lexical choice. That is, a rater first should be interested to continue listening attentively to a voice. Dominant voices express leadership, assertiveness, and security which may impress the raters. Submissive voices express uncertainty, passiveness, and doubt in oneself which may negatively influence the raters. There are some studies corroborating this, showing that raters attend to non-relevant criteria in their assessment of performance, such as the voice quality of the test takers (Brown, 2000; Orr, 2002; Sato, 2012). This was the case with a learner called Negar whose speech sample showed acceptable features; however, received low scores due to her voice which was low, unenthusiastic, and unclear. On the contrary, the learner named Tarlan received higher scores compared to Negar in spite of her relatively moderate features of speech samples and most probably due to her loud clear, and energetic voice.

Personality Traits of the Learners

The raters were inadvertently influenced by the personality traits of the learners they rated. Two of such features which were more repeatedly mentioned by the raters are explained below. These two features are both conveyed through the tone of voice and the content of the speech.

Assertiveness and confidence

Assertiveness was expressed by both the tone of voice and the content. In terms of the tone of voice, a voice that is energetic is more successful in attracting the attention of the rater. The raters were influenced and impressed by assertiveness as reflected in comments like:

- *not confident*
- *She seems uncertain.*
- *It seems that she herself is dubious about what she is saying.*
- *his justifications for late marriage didn't convince me*

Enthusiasm

As related to the above category, the raters cared about whether learners put energy in both their voice and also the way they talked.

- *He just wants to get rid of the situation.*
- *She doesn't show energy to attract the audience.*
- *It was obvious that he liked English.*
- *Her speaking is boring.*

There were other comments reflecting different personality features sporadically pointed to by the raters. However, they were not repeatedly mentioned to be taken as a distinct category.

- *Hasan had a good sense of humor.*
-*she seems so snobbish.*

As stated before, this attention to personality features shows that the untrained raters cannot stay impartial.

Effort to Talk

Whether a learner relinquished the effort to talk in case of failure to talk or tried to tap on every available resource and ability to be communicatively successful was also among the factors that attracted the attention of the raters.

- *She tries to take turns although she doesn't know English much.*
- *She does not stay silent, every now and then she says something.*

Communication Strategies

The raters considered the learners' using communication strategies to compensate for problems in their speaking positively. This may be in line with 'effort to talk.'

- He corrected his mistakes, that's good. (self-correction)

- I know that she couldn't find the correct word but she should have found a synonym or something, not turn to Farsi,

it shows she doesn't feel confident in her English.

Observation Results

As mentioned earlier, the researchers observed how the raters rate the samples, taking notice of factors like the speed with which they rate, their hesitance or certainty, etc. Although different raters approached the task of rating differently, some common trends were also evident. The researchers took notice of both common and idiosyncratic behaviors while the raters did the ratings. The following are the results of these observations.

1- Initially, the raters rated more intuitively. Rating the first speech samples, they were more holistic in rating but gradually they developed some specific criteria. This was evident in the speed with which they made a decision about a score. Initially, unable to handle so much information or so many criteria, they needed much time to come to a score. Even in some cases, the raters felt afraid that they might not be consistent in rating and felt that they needed to listen to all samples once before they started rating. Or sometimes they felt that they needed to go back and change a score previously assigned. However, gradually they could rate the learners with more ease. It seemed that some criteria were being shaped in their minds.

2- The raters differed in degrees that they were attentive. Although the ten-minute monologue, for instance, was fully played, some raters made their decisions after they had listened to the first two or three minutes. Only in cases where a learner spoke too little, the raters looked for instances of his or her speech to assign a score. The proficiency level of the learner seemed to influence their attentiveness. The raters were more willing to rate more proficient learners and did that more attentively. However, they also rated the least proficient learners hastily, giving them no further chance. This was the case with Hashem, a learner with the lowest score assigned by all raters. Listening to him for less than 10 seconds, some raters assigned the score. An opposite case was with Shima, one of the high scorers. One of the raters who was apparently impressed by her choice of word as reflected in the starting sentences of

her monologue with 'investigating early marriage and late marriage, one can come to the conclusion that….' said:

- *I know even by this first sentence how proficient she is.*

3-This rater -who was experienced- did not listen to the whole monologue to come across the long pauses Shima had in the middle of her speech, while she was searching for ideas to talk about.

It was quite common for the raters to come to a conclusion very rapidly. They were so confident of their evaluation that they did not seek further evidence to confirm or disconfirm their judgment.

This may do injustice to some learners who may be able to prove themselves proficient in later stages of their production.

4- The scores the raters assigned were much influenced by the raters' proficiency or their perception of proficiency.

Raters who themselves were form-focused paid attention to form. The comments provided by a rater who, as stated by herself, cares about fluency in her own speaking were limited to fluency and pronunciation. In the same vein, if a rater was not proficient in idiomatic use of language, for instance, he would take the expression 'you can say that again' literally and hence would fail to appreciate the high proficiency of the learner in this respect and just be distracted by a large number of pauses and his heavy local accent. By contrast, if a rater was proficient enough and sensitive to the precision of word choice, he could assign a high or low score in this respect.

- *…to me fluency is very important. I myself try to speak with a native-like accent.*

The same rater was so concerned with the phonological aspect of speaking that assigned a lower score to the test takers for not using connected speech and linking the sounds. Maybe either this rater herself has been a fluent speaker of the English language (as expressed by herself) or has had some special training or studies on fluency and pronunciation.

5- The raters somehow justified the performance of the learner and tried to come to the underpinning reasons of performance. For instance, commenting on Tarlan (with a high rate of speech), some raters rightly pointed to the fact that since she produced more sentences, her errors were noticeable compared to a learner of the same proficiency level who produces fewer sentences. Another instance was the raters attributed different justifications for the pauses in the learners' speeches they rated:

- *His pauses are well placed because he is searching for ideas.*

- …. Lots of pauses, maybe he can't find the words he wants or maybe it's his style of speaking.

- When they make long pauses, it shows that their mind is distracted and coherence may be in danger.

- Parisa speaks slowly because she wants to buy time to think. This helps her make fewer mistakes.

6- The inexperienced raters were in some cases more concerned and distracted by going to the extreme in finding faults with their language use. As an example, this was the case with a male inexperienced rater who found fault with Shima (with the mean score of 4.08 out of 6 and among one of the high scorers) as:

She spoke with a feminine voice to be attractive, that's different from having an attractive English accent. I assign a low score to her.

The raters did attend to different factors in rating performances. Some of the features were linguistic like fluency, accuracy, etc. and some other features were mostly non-linguistic. They pointed to some of these factors more frequently and some others less frequently. The factors pinpointed can help understand the diversity of the factors that the raters either attentively or inattentively consider in their rating. As indicated by the frequency of speaking criteria, the teachers mentioned linguistic features more than other rating categories. This is in correspondence with the findings of Kim (2009), Brown et al. (2005), Plough, Briggs, and Van Bonn (2010), and Zhang and Elder (2011) who found that teachers were more critically oriented toward certain features of spoken production in their ratings such as pronunciation, specific grammar use, and accuracy.

Hence, it can be concluded that the raters are influenced by both construct-relevant and construct-irrelevant factors which may have resulted from the lack of a clear definition of the speaking proficiency construct in the raters' cognition. This is referred to as construct representation (Messick, 1989). Two threats to this representation are construct underrepresentation and construct-irrelevant variance. The first is observable in the performance of some of the raters of this study in terms of attention to a limited set of factors and leaving some others unattended. The raters may have an idiosyncratic (mis)understanding of the speaking proficiency construct in their minds; which can be represented as their tendency to attach different definitions and descriptions to the constructs or their components.

Another threat is construct-irrelevant variance which was represented as irrelevant factors influencing their judgment. Some of these raters overemphasized factors irrelevant to the oral proficiency construct like the ideas expressed by the speakers or the personality features of the speakers. This may result in raters' not being able, in some respects, to stay impartial and were biased by such construct-irrelevant factors. Trained raters may also be influenced by these factors, but the difference may lie in the awareness of such biases and the extent to which they play a role in rating. Both the raters and the test developers and users should be aware of these factors.

It seemed that the untrained raters did not have a clear definition of oral proficiency construct, making it difficult to disentangle their perception of oral proficiency from contaminating factors. Raters -as human beings- approach any evaluation task based on their personal judgment and are limited by their cognitions. Since they might not even be aware of the factors they are influenced by, the key factor to ensure the accurate and reliable rating is exerting control on their cognition. As long as the rating cognitions of the raters, which are their actual guide in ratings, are not recognized and defined, no guideline or guide by itself can be of any help.

Conclusion

This piece of research just scratched the surface of untrained raters' cognition in oral assessment.

However, grounded in the actual data, the findings can be helpful in terms of identifying factors that the untrained raters actually take into account in assigning a score.

Normally, training programs may be more focused on what the raters should consider in their rating and not what the raters bring into the rating prior to the training. However, if the trainer becomes aware of the current status of the untrained rater before the actual training, they can better design the training path and implement it. If the raters are familiar with what construct relevant and irrelevant factors the untrained raters are already attentive to, this can help him/her better approach the training undertaking.

To be more specific, the results of this study and other studies can provide the necessary evidence for a description of untrained raters' cognition. The finding that the untrained raters are prone to be influenced by the tone, or voice quality of a speaker can help the training program developer and the trainer to devise activities to both make the raters aware of such construct-irrelevant factors and also help them monitor and control their rating habits.

Alternatively, an untrained rater may be impressed and influenced by the personality feature of the test-taker in a way that this influence is reflected in the score assigned and can be made aware of his/her bias, and be offered techniques to avoid such faulty impressions. Being aware of such tendency on the part of some of the untrained raters, the rating trainer can first caution the untrained raters against such pitfalls and design techniques to help them not get entangled and mistakenly influenced by such factors and not just start the training program with no awareness or knowledge of the actual mental processes going in the minds of their trainee subjects.

It means that the rating cognition developed as the result of experience should be scanned and constantly monitored by the trainer to detect how the raters perceive different factors, the weighing assigned to them, whether the factors they attend to are relevant, and what factors they miss. The key difference is awareness-raising. A rater may already do the task of rating at an acceptable level even if he is not consciously aware of his mental processing while rating. He may be a better rater if he becomes aware and may have control over his rating cognition. This can be brought about by intensive training programs which may hopefully result in the raters' experiential knowledge turning to expertise.

As ideas for future research, more variables like gender, experience, and proficiency and education level of the raters can lend themselves to a more in-depth investigation, enabling one to come to a broader picture of the factors influencing Iranian untrained raters. Trained raters can also be employed to have a broader picture. Employing the trained raters may provide a benchmark to make a comparison between the trained and untrained raters feasible. Moreover, video files can be used instead of audio files to come to richer data of the learners. Using video files can help obtain data like body language, facial expression, etc. of the learners.

References

Ang-Aw, H. T., & Chuen Meng Goh, C. (2011). Understanding discrepancies in rater judgment on national-level oral examination tasks. *RELC Journal*, *42*(1), 31-51.

Ary, D., Jacobs, L. C., & Sorensen, C. K. (2010). *Introduction to research in education.* Eighth Edition. Belmot, CA: Wadsworth/Thomson Learning.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99-115.

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

Bejar, I. I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice, 31*(3), 2-9.

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education, 5*(1), 1460901.

Broadbent, D. (1958). *Perception and communication.* London: Pergamon Press.

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

Brown, A., Iwashita, N., & McNamara, T. F. (2005). *An examination of rater orientations and test-taker performance on English-for-academic-purpose speaking tasks* (TOEFL Monograph No. MS-29). Princeton, NJ: Educational Testing Service.

Davis L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

Duijm, K., Schoonen, R., Hulstijn, J. H. (2018). Professional and non-professional raters' responsiveness to fluency and accuracy in L2 speech: An experimental approach. *Language Testing, 35*(4), 501-527.

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

Glaser, B. G., & Strauss, A. L. (1967*). The discovery of grounded theory: Strategies for qualitative research.* Hawthorne, NY: Aldine de Gruyter.

Han, Q. (2016). Rater cognition in L2 speaking assessment: A review of the literature. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics, 16*(1), 1-24.

Kim, H. J. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*(3), 239-261.

Kim, Y. H. (2009). An investigation into native and non-native teachers' judgments of oral English performance: A mixed methods approach. *Language Testing, 26*(2), 187-217.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing, 31*(2), 177-204.

Messick, S. (1989). *Validity*. In R.L. Linn (Ed.), *Educational measurement*. Third Edition. New York: MacMillan.

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

Plough, I. C., Briggs, S. L., & Van Bonn, S. (2010). A multi-method analysis of evaluation

criteria used to assess the speaking proficiency of graduate student instructors. *Language Testing*, *27*(2), 235-260.

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223-241.

Sterman, J. (2011). Communicating climate change risk in a skeptical world. *Climatic Change*, *108*(4), 811–826.

Tajeddin, Z., Alemi, M., & Pashmforoosh, R. (2011). Non-native teachers' rating criteria for L2 speaking: Does a rater training program make a difference?. *Teaching English Language*, *5*(1), 125-153.

Wei, J., & Llosa, L. (2015). Investigating differences between American and Indian raters in assessing TOEFL iBT speaking tasks. *Language Assessment Quarterly*, *12*(3), 283-304.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, 30(2), 231-252.

Wolfe, E. W., & McVay, A. (2012). Application of latent trait models to identifying substantively interesting raters. *Educational Measurement: Issues and Practice*, *31*(3), 31-37.

Zhang, Y., & Elder, C. (2011). Judgments of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary constructs? *Language Testing*, *28*(1), 31-50.

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgments of oral proficiency in the college English test-spoken English test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306-325.