Journal of Modern Research in English Language Studies ISSN: 2383-4242

Iranian EFL Raters' Cognitive Processes in Rating IELTS Speaking Tasks: The Effect of Expertise

Rajab Esfandiari[1*], Payam Noor[2]

[1*] Assistant Professor of English Language Teaching, Imam Khomeini International University, Iran, *esfandiari@hum.ikiu.ac.ir*

[2] Ph.D. Candidate of English Language Teaching, Imam Khomeini International University, Iran, *payam_ns1914@yahoo.com*

Abstract

Variations in rating the EFL learners' oral performance are often attributed to the variations in the raters' cognitive processes. Han's (2016) 4-stage processing model was used to examine what cognitive processes expert and novice raters follow to rate a recorded response to the IELTS Speaking Task Two by using the IELTS rubrics. Novice and expert raters attended the 4-phase verbal protocol sessions in order to explore the cognitive processes underlying (a) their representations of IELTS speaking rubric, (b) qualitative assessment of a recorded sample response to IELTS Speaking Task Two, (c) quantitative assignment of ratings to the input and (d) revision of the assigned ratings. Qualitative data collection was followed by transcribing, segmenting, encoding, and analyzing the contents of the recorded verbal protocol reports. After content analysis, the four categories of (1) grammatical range and accuracy, (2) fluency and coherence, (3) lexical resources, and (4) pronunciation in IELTS speaking rubric were schemed into 80 themes. NVivo 8 and SPSS 19 were used to analyze the data qualitatively and quantitatively, respectively. Both qualitative and statistical findings showed that the L2 raters with a different range of expertise widely focus on different aspects of the spoken response input, have different interpretations, and apply different criteria when judging the verbal input. The findings of the present study may carry implications for rater training and validity of ratings. Expertise, as the findings of the study show, can exert an influence on the reliability of the ratings.

***Keywords**:* Cognitive, Expertise, IELTS, Rater, Speaking Task

Received 27 December 2018 Accepted 12 May 2019
Available online 25 June 2019 DOI: 10.30479/jmrels.2019.9383.1248

1. Introduction

As a salient factor in oral language assessment, the rating procedure plays a significant role, especially, in high-stakes tests (Tosuncuoglu, 2018). Concerning such rating processes, Han (2016) counted several factors which make the final result of testing speaking unreliable. One of these criteria is personnel, or what may be commonly recognized in literature as *raters* (Bridgeman et al., 2011). Human raters are normally involved in the evaluation of verbal responses that the examinees produce in L2 speaking assessment (Berns, 2005). Therefore, they provide scores for a variety of speaking products, and based on those scores, stakeholders make inferences about examinees' oral language proficiency in opportunities like employment, education, and immigration (Lazaraton, 2005). As Myford and Wolfe (2003) stated, raters as human beings are susceptible to biased, inaccurate, and inconsistent patterns of judgment, which in educational assessment are known as *rater effects*.

Numerous attributes of raters, rater types, and their role in international examinations have been studied, restricted to learners' writing ability (Erdosy, 2004; Eckes, 2008; Wolfe, Matthews, & Vickers, 2010), yet researchers have paid little attention to test takers' speaking abilities. Other attributes such as rater severity, consistency, and interaction with other aspects of rating have been thoroughly and experimentally studied in L2 speaking assessment by some researchers such as Hsieh (2011). Rater cognition and rating process, however, were explored in only a limited number of qualitative studies in L2 speaking assessment (Eckes, 2012). These studies can be classified into two types of exploring the procedure and the cause (Han, 2016).

Han (2016) also stated that one group of studies examined how L2 raters agree, or disagree, in their cognitive processes and/or their rating conducts. These studies looked into the most frequently explored aspects of rating L2 speaking performance, including (1) L2 raters' focus and feature attention, (2) L2 raters' approaches to rating, and (3) L2 raters' representation of the scoring criteria for aspects of performance (Han, 2016). According to Purpura (2014), none of those studies have taken a cognitive approach to investigate the rater judgment, or understanding the processes and strategies which are exploited while raters are trying to internalize the response input.

To fill part of this gap, through qualitative studies, researchers carried out studies to figure out the patterns that show how raters would act while rating a language product and what discrepancies may be witnessed (see Barkaoui, 2010). In these studies, the level of raters' proficiency as well as the amount of training and experience raters have could reinforce the existing rating differences (Davies, 2015).

In line with such gaps, this study attempted to examine the role of rater expertise in assessing speaking. The rationale behind this attempt was to comparatively explore the impact of expertise on the quality of rating behavior between novice and expert raters. Specifically, the cognitive strategies they adopt, while rating a language performance, are qualitatively traced in order to identify noticeable themes affecting a more consistent and systematic rating behavior of Iranian raters in an EFL context. The significant difference between novice and expert raters due to the existence of such important themes is another focal area of the present researchers' interest.

2. Literature Review

2.1. Rater Cognition

According to Bejar (2012), rater cognition is a source of hypotheses about the rating procedure, which together with facts and details on the evaluation, feed right into a validity argument. That interpretive argument should explicitly include assumptions about a rater's variability or a quality-manipulate mechanism, which is actively tracking the scoring manner. Moreover, the validity argument desires to preserve for the scores, which can be acquired beneath extremely variable situations and on an ongoing foundation instead of one factor at a time (Brennan, 2013; Chapelle, 2012; Kane, 2013; Roever, 2011).

Since the early 1990s, research on rater variability has been witnessing a cognitive shift. More studies began to focus on the decision-making processes involved in scoring (Barkaoui, 2010; Lumley, 2002). Typically employing a qualitative and process-oriented approach, in particular verbal protocol analysis, researchers identified a number of decision-making strategies or reading styles utilized by raters when evaluating essays. Cumming's studies (Cumming, Kantor & Powers, 2002) put forward a descriptive framework as the most comprehensive taxonomies for analyzing rater's decision-making strategies. Cumming, et, al. (2002) conducted impressionistic and statistical analysis on raters' thinking aloud verbal reports and identified 27 decision making strategies used by raters to interpret and evaluate L2 compositions (Table 1).

Different from previous studies, Cumming, et, al. (2002) classified the raters' decision- making strategies according to two dimensions—focus and strategy—which were then further divided into three major foci (self-monitoring, rhetorical and ideational, and language) and two types of strategies (interpretation and judgment of a text). In addition, they differentiated raters' meta-cognitive strategies for self-monitoring purposes from the cognitive strategies adopted to process essay features. Thus, this model describes raters' micro-processing strategies in more fine-grained

manner by integrating focus and strategy and by differentiating cognitive from meta-cognitive processes.

Table 1

The Descriptive Framework of Raters' Decision-Making Strategies

Strategies\focus	Meta-cognitive processes Self-monitoring focus	Cognitive processes	
		Rhetorical and Ideational focus	Language focus
Interpretation strategies	Read or interpret prompt or task input or both; Read/reread composition; Envision personal situation of writer.	Discern rhetorical structure; Summarize ideas or propositions; Scan whole composition or observe layout.	Classify errors into types; Interpret or edit ambiguous or unclear phrases.
Judgment strategies	Decide on macro-strategy for reading and rating; Consider own personal response or biases; Define and/or revise own criteria; Articulate general expression; Articulate or revise scoring decision.	Assess reasoning, logic or topic development; Assess task completion or relevance; Assess coherence and identify redundancies; Assess interest, originality, and creativity; Assess text organization, style, register, or genre; Consider use and understanding of source material; Rate ideas and rhetoric.	Assess quantity of total written production; Assess comprehensibility and fluency; Consider frequency and gravity of errors; Consider lexis; Consider syntax and morphology; Consider spelling and punctuation; Rate language overall.

Through analyzing raters' verbal protocols, researchers (e.g., May, 2011) have identified some of the main (meta)cognitive categories and features of interactional competence that raters put emphasis on, such as non-verbal interpersonal communication (e.g., gestures, gaze, laughter), interactional listening comprehension, or interactional management (e.g., topic change and turn organization) (Ducasse, 2010). Not only do those identified features have implications for redefining the construct of oral proficiency in interactional speaking contexts and operationalizing it in rating scales, they also enrich our understanding of rater cognition by demonstrating the variations in raters' focus and feature attention while scoring a different L2 speaking task type.

2.2. The L2 Rater Cognitive Processing Models

A thorough understanding of rater characteristics is fundamental to grasp a better image of raters' behavior, or decision-making processes; such knowledge will serve to explain why and how the raters assign scores and what attributes or elements they still need to improve in their rating performance (Kim, 2015). Among all such attributes, rater language background (Wei & Llosa, 2015), rater experience (Kim, 2015), and rater training (Davis, 2015) have been most frequently studied for examining their effects on raters' cognitive processes and rating behaviors in L2 assessment.

Raters' cognitive processes are among other rater dimensions, which may affect rating behavior. "Raters' cognitive processes mainly pertain to the architecture of human information processing" (Han, 2016, p. 3), and to the various strategies that the raters deploy during a successful rating (Purpura, 2012). According to Han (2016), the architecture of human information processing can effectively "explain the underlying structure and processes (e.g., working and long-term memories) involved in the encoding, storage and retrieval of information during rating" (Han, 2016, p.3). A few cognitive processing models that can thoroughly conceptualize those processes undertaken in rating L2 speaking and writing assessment have been reported in the literature. Han (2016) suggested one of the most influential cognitive processing models.

Based on Purpura's (2014) model of the architecture of human information processing, Han (2016) proposed a tentative, unified model for hypothesizing rater cognition in the context of L2 speaking assessment. As shown in Figure 1, Han's model interfaces the process of rating L2 speaking with the components of the architecture of human information processing (Purpura, 2014). It also incorporates a wide range of (meta)cognitive, (meta)affective, and (meta)socio-cultural-interactional strategies that raters may employ during the rating process. This model delineates how the assessment input (i.e., the rubrics, the exemplars, and the L2 spoken responses) might be picked up by the raters' sensory receptors and selectively attended to and initially processed in short-term memory (STM). It also explains how the working memory (WM) de/encodes the assessment input information, retrieves and activates different types of knowledge from long-term memory (LTM), so that all types of information can be reorganized and mental representations of both the rubric and the L2 responses can be formed (Han, 2016).

Han's model describes how those mental representations are compared and contrasted to one to another to produce tentative scores in working memory (WM) (Baddeley, Eysenck, & Anderson, 2009). Finally, how the scores are reviewed or revised through an iterative scoring process by means

of the cognitive components (i.e., sensory memory, STM, WM, and LTM) are presented in this model (Davis, 2015). These cognitive steps are regulated and empowered by the range of cognitive strategies (e.g., attending and monitoring) (Purpura, 2012). Many elements of rater cognition, discussed so far, have the capability to serve as explanatory variables for rater conduct that negates the assumption that raters are scoring appropriately and functioning as exchangeable agents. That is, in a standardized evaluation, it does not have to matter to test takers and rating recipients who happens to score the responses or while or where the scoring certainly occurs (Kim, 2015).

Proofs of variability in rater consistency and severity amongst raters (Esfandiari & Myford, 2013) implies that it could be counted who happens to

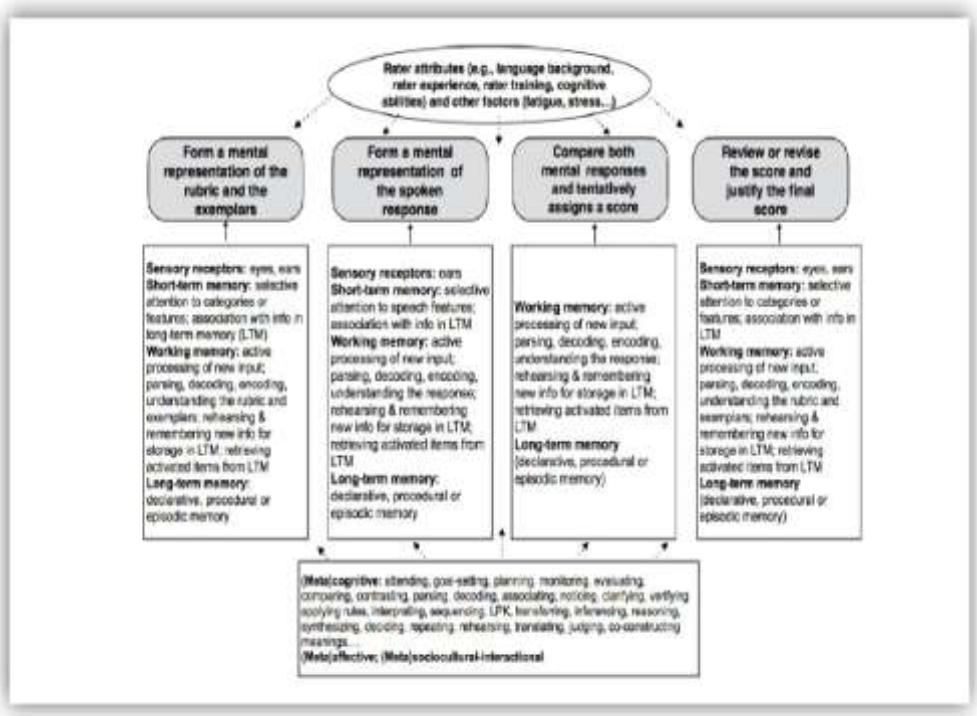

Figure 1. A Tentative, Hypothesized Cognitive Processing Model of Rating L2 Speaking (Han, 2016, p.5)

be the scorer of a given response, particularly whilst rankings are based totally on a single rater and there are not any methods of satisfactorily controlling every rating (Davis, 2015). Laming (2003) believed that the precision and consistency of scores given by raters are promoted over time. This is because of the fact that as the raters rate more language performances, they get acquainted with a larger set of examinee performances, which helps them to draw a contrastive pattern between different performances before

they decide on what score they must give to a specific performance. Controlling raters' rating behavior, however, can be done using various cognitive techniques and strategies like task-specific performance, as outlined in Ericsson (2006). Such task-based performances are among the factors that enforce the inexperienced raters to adopt such enhanced rating behavior as expert raters typically do. Ericsson (2006) defined the term as the skill one may gain in a specific task domain. Further, patterns of similarity and dissimilarity among clusters of raters would also suggest that it matters who happens to score a given response (Davis, 2015).

Considering scoring variability, Mislevy (2010) believed that novice raters are incapable of handling some obstacles that are easily handled by expert raters. He drew a distinction between two groups of such constraints: those due to lack of rating procedure and those due to lack of rating knowledge. In his division, he viewed the former as individuals' deficiencies regarding manipulating and processing information as well as justifying and using logical patterns to have a more consistent scoring. The latter, on the other hand, implies a well-conducted analysis of the data, distinguishing different sets of data from one another.

It was the aim of the present study to investigate the role of rating expertise (novice vs. expert) and the variability in their cognitive processes underlying rating L2 speaking processes. This study was founded on Han's (2016) prototype cognitive processing model of L2 raters and Cumming et al.'s (2002) descriptive framework of raters' decision-making strategies. Therefore, the following research questions were used to achieve the goals of the present study.

1. What between-group differences exist between the expert and novice raters' cognitive processing regarding the cognitive representation of IELTS speaking rubrics?

2. What between-group differences exist between the expert and novice raters' cognitive processing regarding the qualitative assessment of IELTS speaking tasks?

3. What between-group differences exist between the expert and novice raters' cognitive processing regarding the quantitative assessment of IELTS speaking tasks?

4. What between-group differences exist between the expert and novice raters' cognitive processing regarding the revising/finalizing their assessment of IELTS speaking tasks?

3. Method

3.1. Participants

In this research, the participants included 10 expert and 7 novice Iranian raters who were selected with a purposeful sampling technique. The major qualifications of these participants were their knowledge of L2 speaking assessment and their level of English language proficiency, which was assumed high as a general requirement for IETLS raters or examiners. The participants' personal attributes such as their age, sex, and educational background were neutralized for their possible confounding effects in this study. Table 2 reports the demographic information of the participants.

Relying on Govaerts, Schuwirth, Van der Vleuten, and Muijtjens' (2011) definition of expertise, through snowball sampling technique, raters with the range of at least 7 years of experience were selected as expert raters. These raters had also attended IELTS examination interview sessions and gained scores above band 8, and they had at least seven years of rating IELTS test or its mock version.

Table 2
The Expert and Novice Raters' Demographic Information

Raters	Gender	Degree	Specialty	Years of Experience
AA	Male	MA	TEFL	6
AP	Male	BA	English & Communication	7.5
FS	Male	MA	TEFL	5
HA	Female	MA	MBA	6
PM	Male	MA	TEFL	10
LM	Female	PhD candidate	English Literature	5
RM	Male	MA	TEFL	4
HS	Male	PhD candidate	TEFL	7
KA	Male	PhD	Medical sciences	7
AZ	Male	BA	Translation	7
MM	Female	MA	TEFL	1
VP	Male	MA	TEFL	1
ME	Male	BA	English Literature	1
KJ	Male	PhD Candidate	English Literature	1
CN	Female	PhD candidate	TEFL	1
PP	Male	BA	TEFL	1
VN	Female	BA	TEFL	1

Note. For anonymity purposes, raters' initials have been used.

3.2 Instruments

In this research, (a) an IELTS speaking scoring rubric and (b) an Iranian mock IELTS candidate's response to Task Two (The cue card) in the

IELTS Interview were incorporated as the inputs to collect data from the participants in terms of verbal protocols of their rating process in the light of cognitive processing model.

3.2.1. IELTS Speaking Scoring Rubric (Public Version)

The IELTS speaking marking scheme is on a scale of nine. The criteria in IELTS speaking section consist of (a) fluency and coherence, (b) lexical resource, (c) grammatical range and accuracy, and (d) pronunciation. The participants' mental representation of the IELTS speaking scoring rubric was investigated in Phase 1 of data collection.

3.2.2. L2 Speaking Assessment Task

The participants were required to rate an Iranian mock IELTS candidate's recorded response to Task Two in the IELTS Speaking Interview according to the IELTS 9-band scale. In this task, a candidate responded to the following questions for 5 minutes.

- *Let's talk about your home town. What kind of a place is it? What's the most interesting part of your town? Would you say it is a good place to live? Why?*

The candidate's response was recorded by the researchers to elicit the raters' rating processes in verbal protocols phases 2, 3, and 4 of data collection.

3.2.3. QSR NVivo Version 8

NVivo is a computer programme that analyses qualitative data. It is designed to help the researchers organize, analyze, and find insights in unstructured or qualitative data such as interviews, open-ended survey responses, articles, social media, and web contents. NVivo is used predominantly by academic, governmental, and commercial researchers across a diverse range of fields, including social sciences such as anthropology, psychology, communication, sociology, as well as fields such as forensics, tourism, criminology and marketing. The first NVivo software was developed by Richards (1999). Originally called NUD*IST, it contained tools for fine, detailed analysis and qualitative modeling (McNiff, 2016).

In order to encode and analyze the themes distribution in four categories of IELTS Speaking rubric, the raw data were inserted into NVivo 8, which assisted the researchers with counting thematic frequencies and cross-examination of the individual rater participants' verbal protocols.

3.2.4 Think Aloud (Verbal) Protocol

Think-aloud (verbal) protocol as method of data collection requires the research participants to think aloud while they are performing a set of specified tasks (i.e., introspective) or afterwards (i.e., retrospective) (Lumley,

2005). The participants are asked to say whatever comes to mind impromptu or they remember later after performing a task. This offers the observer/examiner insights into the participant's cognitive or meta-cognitive processes (rather than only their final product), and makes thought processes as explicit as possible during task performance. In a formal verbal protocol, all verbalizations are recorded, transcribed, and then analyzed.

3.3 Procedures

In this study, 17 Iranian IELTS speaking raters were purposefully selected. Due to the researchers' limited access to a large enough sample of qualified IELTS raters, some of them were selected through snowball sampling as a non-probability sampling technique. This purposeful sampling of the participants was done among the small community of IELTS raters in Karaj.

The verbal protocol sessions lasted about 30-45 minutes for each individual rater. Before the verbal protocol meetings, the researchers recorded an EFL learner's voice that provided a response to Task Two in the IELTS interview as a mock IELTS test. This recorded response was used as the input to elicit the raters' verbalized rating processes. The verbal protocol sessions were conducted in four consecutive phases. After a short warm-up and briefing on the research objective and the verbal protocol rudiments, the procedure was conducted as the following:

Phase I: Eliciting the mental representation of the IELTS speaking rubric. The IELTS raters were required to elaborate on the IETLS speaking rating rubric while their voices were recorded by the researchers. To refresh their memory, the participants were provided with a printed copy of the rubric. This phase lasted between 20 to 30 minutes.

Phase 2: Qualitatively assessing the exemplar response. The IELTS raters had to listen to the input and verbalize their stream of thoughts in their rating process while their voices were recorded by the researchers.

Phase 3: Quantitatively assessing the exemplar response. Immediately after listening to the recorded response, the IELTS raters were asked to rate the input holistically while they had to verbalize how they reached their assigned score. Their voice was recorded by the researchers.

Phase 4: Revising/assigning the final score. The terminal phase in the verbal protocol was replaying the recorded input, which the IELTS raters were required to listen to again. They were asked to revise their assigned score or finalize it while the researchers recorded their verbalized rating processes.

3.4 Coding System

The IELTS speaking rubric consists of four criteria including (a) fluency and coherence, (b) grammatical range and accuracy, (c) lexical resources, and (d) pronunciation. Therefore, the raters' recorded verbal protocols were segmented, encoded, and analyzed following these categories in the IELTS speaking rubric. The researchers made use of QSR NVivo 8 to encode the transcribed verbal reports as well as the IELTS speaking rubric. The coding schemes are summarized in Table 3, Table 4, Table 5, Table 6, and Table 7. In every table, a recorded example is provided for the encoded themes. The reference is made to the rater's initials and the number of the verbal report phase (e.g., AZ2 means AZ's record in Phase II).

Table 3
Coding Scheme 1: Fluency and Coherence (n = 29)

Theme	Example	Frequency
Pauses	She speaks fluently except some parts especially in the beginning of her performance she had pauses. AZ2	77
Expressing and expanding ideas	She is talking about the weather of Ahvaz irrelevantly. AZ4	59
Communication	This may and can lead to misunderstanding and blockage of message. FS1	58
Speaking and Conversation	At the first moments, the speech was harmonious. AP4	30
Discourse markers	Discourse markers are seen more and accurate. Self-correction is less than band 6. HA1	25
Memorization	She has memorized them. She is not natural. AA4	21
Language-related hesitations	She has got some pauses but the pauses are logical philosophical pause. LM4	18
Pre-fabrication	She has been told to use gap-fillers, adverbs and many pre-fabricated words. AA4	16
Fillers	But a band 8 has the total awareness how to logically and naturally fill the empty spaces. FS1	15
Connectors	The sentence becomes bigger and the candidate uses more connectors. FS1	14
Reasoning	She is making logic well. FS2	13
Length	Therefore, they can produce noticeably long and relevant sentences. FS1	11

Paraphrase	She used paraphrase strategy because she didn't have a proper word for that. AZ4	10
Loophole	She is trying to get out of the loop. KA2	9
Flow of speech	For scale 7, in Fluency and coherence, the person has all natural flow of speech. AA1	9
Over usage	He or she should try not to repeat these cohesive devices. PM1	7
Familiar Topics	A person who deserves band score 4 is a person who is able to talk about familiar topics. LM1	6
Unfamiliar Topics	Sometimes they may be able to lead speech about *unfamiliar topics.* HA1	6
Linking words	She used a little bit of complex grammars like "because of" which was good. PM4	6
Signpost	She mentioned a signpost to determine what she is going to talk about. AZ2	5
Fixed phrases	There were some fixed phrases. KA2	4
Mind map	I urge them to use my mind map applications also. KA4	4
Anthropomorphism	She thinks government is a person, is a big male person. KA2	3
Linking Sentences	She has limited ability to link simple sentences. LM1	3
Forward movements	They don't have forward movement. LM1	2
Backward movements	Instead, they have backward movement. LM1	2
Circumlocutions	Again circumnavigation (circumlocution) always going on. KA2	2
Cliché usage	So, over usage of the cliché advantages disadvantages another negative bias. KA4	2
Full range	Band score 9 is given to a person who uses a full range of structures naturally and appropriately. LM1	2
Scenarios	My students have at least 3 different scenarios in their sleeves. KA4*	1

According to the summarized data in Table 3, the category of *Fluency and coherence* in the IELTS Speaking rubric is encoded into 29 themes after

content analysis of the raters' verbal protocols. Apparently, the theme of *Pauses* (f = 77) is the most frequent one in this category.

Table 4
Coding Scheme 2: Lexical Resources (n = 23)

Theme	Example	Frequency
Basic Structures	The range of vocab is really simple.	48
Lexicon	This running short of vocabulary and lacking vocabulary means is meaningless for band 8. FS1	33
Wide range	In band 4 we may have a *wider range* of the words AH1	24
Idiomatic language	I didn't notice any idiomatic language AZ4	15
Lexical primes	These are usually as we may call it lexical primes but… KA2	13
Surface words	She can use synonyms instead of a lot of good and bad. PM2	12
Chunking	She doesn't use chunks. LM4	10
Contextualization	Once in a while she tries to use some high range vocabulary but they may not be contextualized. LM3	9
Collocations	In lexical resources we may have idiomatic structures and collocations are more accurate and vast. HA1	7
Adjectives	In that case, I usually focus on the use of adjectives, and adverbs mostly for above 6 or 7. KA1	7
Familiar topics	I can say in band 5 the range of vocabulary is wider and the speaker is capable of conducting a speech about the familiar topics. HA1	6
Unfamiliar topics	Sometimes they may be able to lead a speech about *unfamiliar topics.* HA1	6
Synonyms and antonyms	It means he shouldn't repeat for several times and try to substitute some appropriate synonyms. PM1	4
Adverbs	Something negative is the repetition of the adverbs really which make her speech boring!!! HA4	4
Nouns	She recognizes singular and plural nouns. FS2	3
Phrasal verbs	She didn't use phrasal verbs. LM3	3
Rhyming	For example, she had this *rhyming* priming that we also have. KA2	3
Catch phrases	We use some of our catch phrases or catch words or we	2

	make one. KA2	
Prepositions	She can recognize the prepositions. FS2	2
Orthography	They make mistakes for example in those word clusters having the same *orthography* in their nouns and verbs. FS1	1
Proper nouns	She is using some proper names in that case. KA4	1
Proverbs	He uses some less common and idiomatic vocabularies some proverbs some slangs. LM1	1
Verb	They make mistakes for example in those word clusters having the same orthography in their nouns and verbs. FS1	1

As Table 4 displays, the category of *Lexical resources* in IELTS Speaking rubric is encoded into 23 themes after content analysis of the raters' verbal protocols. Accordingly, the theme of *basic structures* (f = 48) is the most frequent one under this category.

Table 5
Coding Scheme 3: Grammatical Range and Accuracy (n = 13)

Theme	Example	Frequency
Accuracy	I can say his words are *correct*. The circle of words he uses is vast and the meanings are *correct*. HS1	59
Basic Structures	In band 5, the performance is focused mostly on *basic* tenses. HA1	48
Structures	She doesn't know how to use passive *structures*; she says "everything damaged there" AA2	39
Wide Range	In band 4, we may have *a wider range* of the words. HA1	24
Tense	Do they recognize what specific *tense* they must take advantage of? FS1	23
Passiveness	This person knows *passives* and can talk about past. FS2	13
Relative structures	She doesn't use reduced *relative clauses* or cleft structures. LM4	6
Clauses	Compound and complex structures, reduced *clauses*, conditionals, and the passive sentences and these are really effective in the score they achieve. AZ1	4
Conditional Sentences	She didn't use *conditional* structures. LM3	2

Full range	Band score 9 is given to a person who uses a *full range* of structures naturally and appropriately. LM1	2
Verb clauses	And *verb clauses* they come later. KA4	2
Infinitive	An *infinitive to* was missed as a grammatical point. AZ2	1
Noun clauses	In *noun clause* style usually using a lot of noun *clauses.* KA4	1

As it can be seen in Table 5, the category of *Grammatical Range and Accuracy* in the IELTS Speaking rubric is encoded into 13 themes after content analysis of the raters' verbal protocols. Apparently, the theme of *Accuracy* (f = 59) is the most frequent one in this category.

Table 6

Coding Scheme 4: Pronunciation (n=15)

Theme	Example	Frequency
Accuracy	Accuracy is not of an alterable concept because it has been amassed. AP1	59
Intonation	They may have problems with intonations occasionally which would not lead to misunderstanding. FS1	33
Accent	Her pronunciation and accent are not natural yet. She is thinking too much about the structures. AA4	22
Stress	She knows how to differentiate the stress in the first pattern in the first syllable and the second syllable. LM1	20
Occasional	Her fluency is not bad but occasionally she has pauses like her vocabulary range is limited and she looks for words. AZ2	19
Stress patterns	The stress on the wrong syllabus ignoring the fact that maybe the one cannot pronounce consonant clusters. FS1	11
Pronunciation features	Among *the features* that increase a candidate's score, only the word stress is important. AZ1	8
Persian accent	And for pronouncing a couple of words wrong and a Persianized accent. AZ4	6
Tone	Usually I focus on enunciation of the words, not the whole *tone* or their accent. KA1	3
Vowel and Consonant clusters	Many of the consonant clusters are difficult to handle by such a person. They insert an extra sound. FS1	3
Rising intonation	…because when you are talking it is completely crystal	2

	clear that you know when to go up ok? PM1	
Falling intonation	The rise and fall is crystal clear. PM1	2
American accent	You know Iranian accent is OK. It doesn't have to be American or British. KA1	1
Listening	On the side of the listening, that is the most important thing that I focus on, you know some you know Iranian accent is OK. KA1	1
British accent	You know some you know Iranian accent is OK. It doesn't have to be American or British. KA1	1

As Table 6 summarizes, the category of *Pronunciation* in the IELTS Speaking rubric is encoded into 15 themes after content analysis of the raters' verbal protocols. Accordingly, similar to grammatical range and accuracy, the theme of *Accuracy* (f = 59) is the most frequent one in this category.

Prior to investigating the research questions in this study, since the raters' cognitive representations in individual verbal protocol phases were cross-sectioned with one another, as well as with the IELTS Speaking rubric, the researchers decided to segment and encode the IELTS rubric band scores similar to the raters' provided input in the next coding scheme. Table 7 illustrates the coding scheme 5 in this study.

Table 7

Coding Scheme 5: IELTS Speaking Rubric (n = 25)

Theme	Examples	Frequency
Communication	Produces simple speech fluently, but more complex *communication* causes fluency problems.	11
Fluency and coherence	Speaks *fluently* with only rare repetition or self-correction.	11
Basic	Uses *simple* vocabulary to convey personal information.	10
Self-correction	Speaks fluently with only rare repetition or *self-correction*.	9
Appropriacy	Speaks coherently with fully appropriate cohesive features.	9
Errors	Uses a limited range of more complex structures, but these usually contain *errors* and may cause some comprehension problems.	8
Flexibility	Uses a range of connectives and discourse markers with some *flexibility.*	8
Lexical resources	Uses *vocabulary* with full flexibility and precision in all topics.	8
Accuracy	Uses idiomatic language naturally and *accurately*.	8

Pauses	Any *hesitation* is content-related rather than to find words or grammar.	7
Speaking and conversation?	*Speaks* coherently with fully appropriate cohesive features.	7
Structures	Uses a full range of *structures* naturally and appropriately.	7
Comprehension	May make frequent mistakes with complex structures though these rarely cause *comprehension* problems.	6
Discourse markers	Speaks coherently with fully appropriate *cohesive* features.	6
Repetition	Speaks fluently with only rare *repetition* or self-correction;	6
Authenticity	Uses a full range of structures *naturally and appropriately.*	3
Collocation	Shows some awareness of *collocations* and style.	3
Content-related hesitation	Any hesitation is *content-related* rather than to find words or grammar.	3
Grammatical range and accuracy	Any hesitation is content-related rather than to find words or *grammar.*	3
Pronunciation	Uses a full range of *pronunciation* features with precision and subtlety.	3
Memorization	Only produces isolated words or *memorized* utterances.	3
Accent	Is easy to understand throughout; L1 *accent* has minimal effect on intelligibility.	1
Language	No ratable language.	1
Language-related hesitation	May demonstrate *language-related* hesitation at times, or some repetition and/or self-correction.	1
Flow of speech	Usually maintains *flow of speech* but uses repetition.	1

Table 7 summarizes the 25 encoded themes inside the IELTS Speaking rubric after content analysis. As it can be seen, on the act of communication and the candidates' natural flow of speech, the themes of *Fluency and coherence* as well as *communication*, each with 11 occurrences inside the IELTS band scores are the cornerstones in the IELTS Speaking rubric.

4. Results and Discussion

4.1. Results

4.1.1. Investigating the First Research Question

To answer the first research question that asked if some meaningful between-group differences could be found between expert and novice raters' cognitive processing regarding the cognitive representation of IELTS speaking rubrics, we used QSR NVivo 8 which provided us with the following percentages as shown in the following figures.

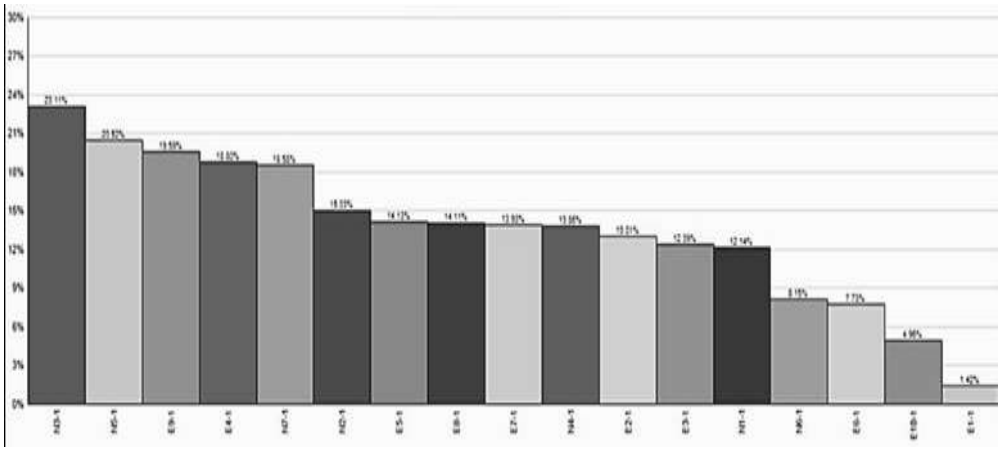

Figure 2. Fluency and Coherence: Phase 1 for Expert and Novice Rater

As it can be seen in Figure 2, the highest amount of cognitive representation of IELTS Rubric related to fluency and coherence belongs to Novice rater 3 with 23.11%, while the lowest amount belongs to Expert rater 1 with 1.42%. It should be noted that the second highest amount of cognitive representation again belongs to a Novice rater. Compared to IELTS Speaking rubric, the fluency and coherence load was 34.22%. All raters were below the rubric amount when it came to fluency and coherence in Phase 1.

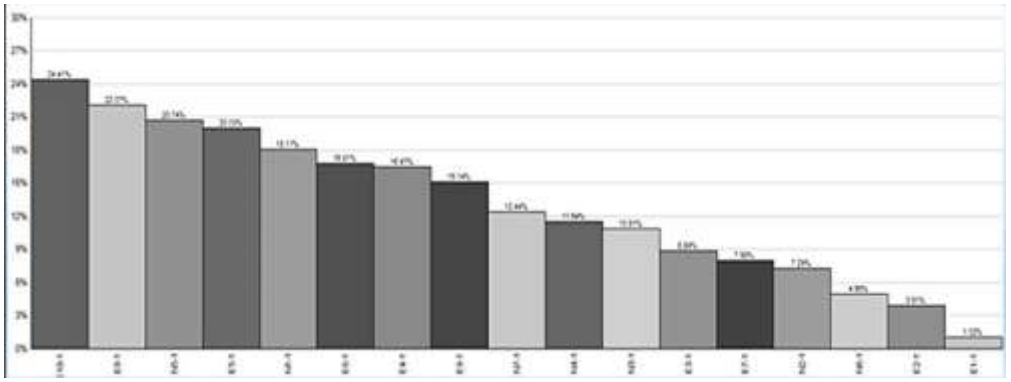

Figure 3. Grammatical Range and Accuracy: Phase 1 for Expert and Novice Raters

As Figure 3 represents, the cognitive representations of the raters for grammatical range and accuracy in Phase 1 is displayed. Among 17 Novice and Expert raters, 13 exceeded the amount of grammatical range and accuracy detected in IELTS Speaking Rubric (7.64%). Their records ranged from 24.41% for Expert rater 10, and the lowest range of 1.02% for Expert rater 1.

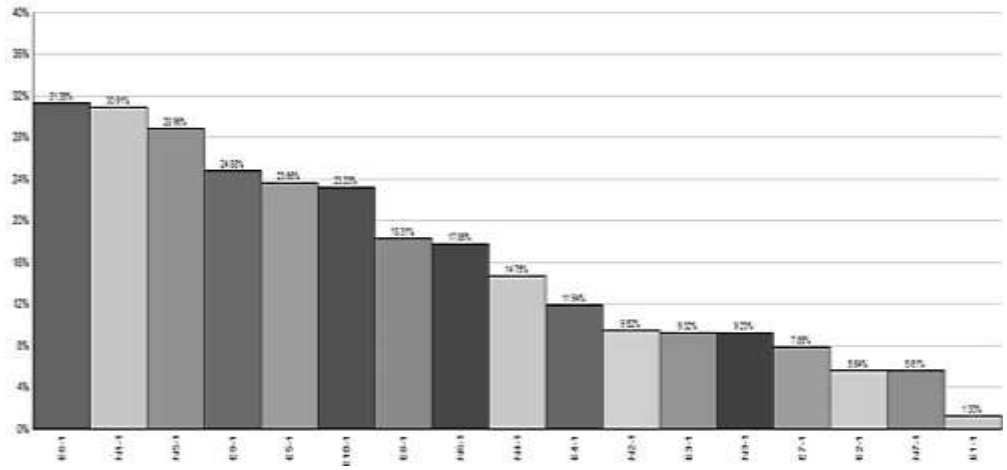

Figure 4. Lexical Resources: Phase 1 for Expert and Novice Raters

In Figure 4, it can be seen that in Phase 1 of verbal protocols, the frequency for lexical resources had a range from 1.30% to 31.35% for Expert rater 1 and Expert rater 6, respectively. Such a low frequency for this criterion compared to IELTS speaking rubric (43.85%) shows a sort of oddity since the importance of vocabulary as a central component of the production of speech is always emphasized.

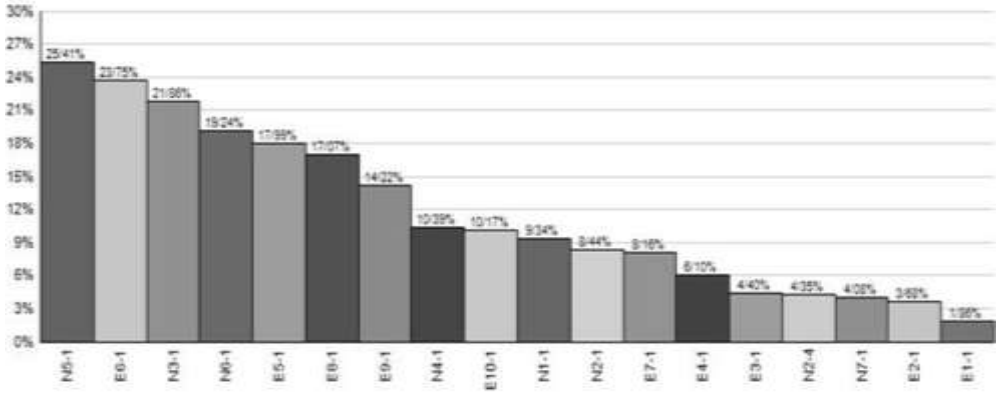

Figure 5. Pronunciation: Phase 1 for Expert and Novice Raters

As shown in Figure 5, in Phase 1 of verbal protocols related to the frequency counts of pronunciation, the rate was 14.29% in the content analysis of the IELTS speaking rubric. However, only six raters have exceeded the rubric rate. The lowest degree belongs to Expert rater 1 with 1.86% attention to pronunciation features. The highest, however, belongs to Novice rater 5 who had 25.41% of her attention to pronunciation features.

A 2 x 4 group-independence Chi-square test was performed to assess the relationship between the nature of expertise and the four criteria of IELTS speaking rubric in Phase 1. A contingency table for these data is shown in Table 8.

Table 8
Contingency table for Expert vs. Novice Raters and IELTS Speaking Rubric Criteria in Phase 1

			Criteria				Total
			Fluency & Coherence	Lexical Resources	Grammatical Range & Accuracy	Pronunciation	
Rater	Novice	Count	30	27	23	25	105
		Expected Count	28.0	29.0	23.0	23.0	105.0
		% within Criteria	23.0%	20.0%	21.0%	24.0%	22.0%
	Expert	Count	97	102	83	78	360
		Expected Count	98.0	99.0	82.1	79.0	360.0
		% within Criteria	76.0%	79.1%	78.0%	75.0%	77.0%
Total		Count	127	129	106	103	465
		Expected Count	127.0	129.0	106.0	103.0	465.0
		% within Criteria	100.0%	100.0%	100.0%	100.0%	100.0%

As Table 9 shows, there was a statistically significant relationship between the raters' expertise and the criteria they used, while verbalizing their cognitive representation of the IELTS speaking rubric in Phase 1 [χ^2= 0.000$_{, (3, 465),}$ p = 0.000, *Cramer's V* = .033].

Table 9
Chi-square test between Raters and Criteria in Phase 1

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.000	3	.000
Likelihood Ratio	.000	3	.000
Linear-by-Linear Association	.015	1	.000
Cramer's V	.033		.000

4.1.2. Investigating the Second Research Question

To examine the second research question on the presence of between-group differences between the expert and novice raters' cognitive processing

regarding the qualitative assessment of IELTS speaking tasks, we used QSR NVivo 8 to analyze the collected data. The results are shown in the following four figures.

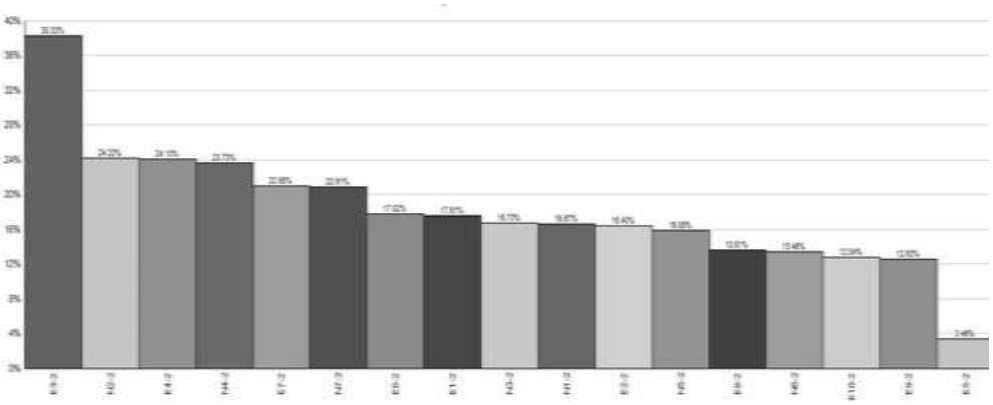

Figure 6. Fluency and Coherence: Phase 2 for Expert and Novice Rater

In Phase 2 of the verbal protocols, as Figure 6 represents, the highest amount of frequency loads for fluency and coherence belong to Expert rater 3 with 38.33%, which is close to the IELTS Speaking Rubric amount (34.22%). The lowest frequency load belongs to Expert rater 5 as well with 3.46%. As it can be seen, only one rater's frequency load exceeded the IELTS Rubric figure.

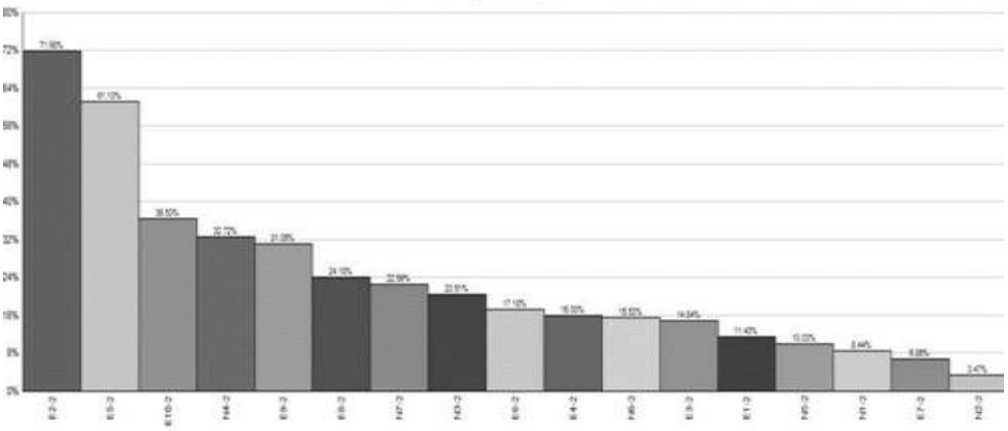

Figure 7. Grammatical Range and Accuracy: Phase 2 for Expert and Novice Rater

As Figure 7 illustrates, attention paid to Grammatical Range and Accuracy increased up to 74.98% in Phase 2 by Expert rater 2, followed by Expert rater 5 with 61.13% which was different from other raters by an

increased amount around twice more. Novice rater 2 has paid the least attention to grammatical soundness of the input with the amount of 3.47%.

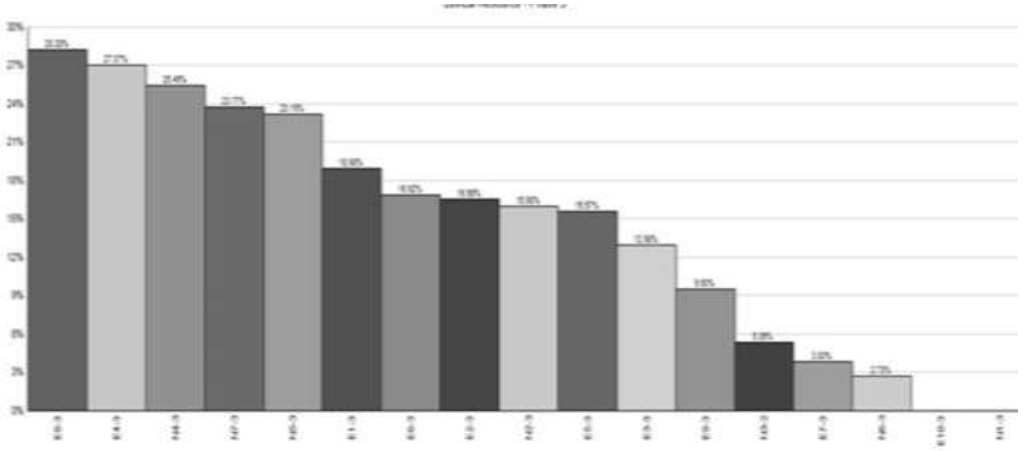

Figure 8. Lexical Resources: Phase 2 for Expert and Novice Raters

As Figure 8 shows, in Phase 2 of verbal protocols, the highest rate for Lexical Resources belongs to Expert rater 8 (28.26%) and the lowest rate belongs to Novice rater 1 (0.00%) and Expert rater 10 (0.00%).

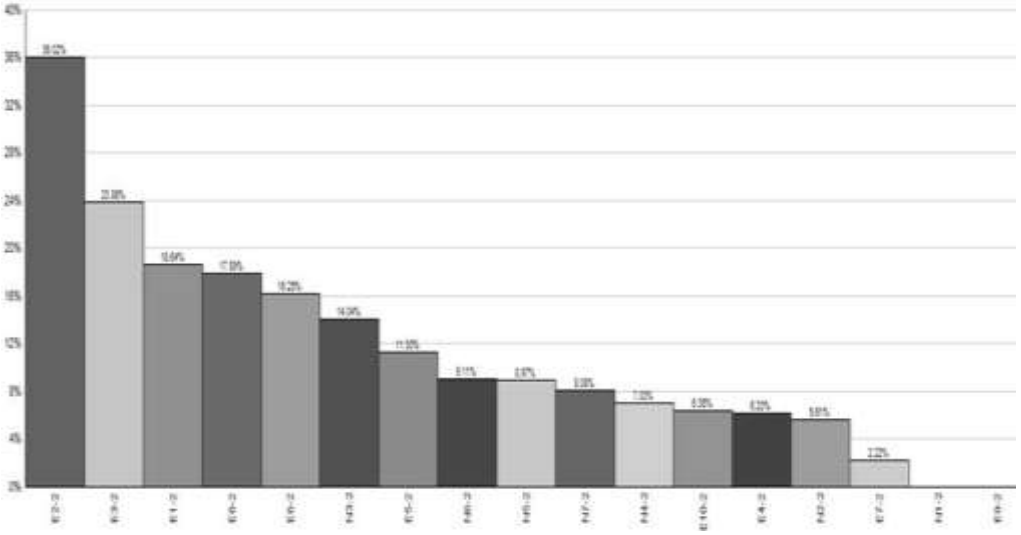

Figure 9. Pronunciation: Phase 2 for Expert and Novice Raters

As displayed in Figure 9, in Phase 2 of verbal protocols, the highest degree of attention to Pronunciation belongs to Expert rater 2 with 54.02% and the lowest degree goes with Novice rater 1 and Expert rater 9 with 0% attention to Pronunciation. It should be noted that five top raters were Experts

in this phase and all of these raters have exceeded the pronunciation range of 14.29% in the IELTS speaking rubric.

A 2 x 4 group-independence Chi-square test was performed to assess the relationship between the nature of expertise and the four criteria of IELTS speaking rubric in Phase 2. A contingency table for these data is shown in Table 10.

Table 10
Contingency table for Expert vs. Novice Raters and IELTS Speaking Rubric Criteria in Phase 2

			Criteria				Total
			Fluency & Coherence	Lexical Resources	Grammatical Range & Accuracy	Pronunciation	
Rater	Novice	Count	29	14	26	16	85
		Expected Count	27.0	23.0	21.0	12.0	85.0
		% within Criteria	28.0%	15.0%	31.0%	33.0%	26.0%
	Expert	Count	74	76	57	32	239
		Expected Count	75.0	66.0	61.0	35.0	239.0
		% within Criteria	71.0%	84.0%	68.0%	66.0%	73.0%
Total		Count	103	90	83	48	324
		Expected Count	103.0	90.0	83.0	48.0	324.0
		% within Criteria	100.0%	100.0%	100.0%	100.0%	100.0%

As Table 11 shows, there was a statistically significant relationship between the raters' expertise and the types of criteria they used, while qualitatively assessing the IELTS speaking input in Phase 2 [$\chi^2 = 7.000_{(3,\ 324)}$, $p = 0.049$, *Cramer's V* = .049].

Table 11
Chi-square test between Raters and Criteria in Phase 2

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7.000	3	.049
Likelihood Ratio	8.000	3	.039
Linear-by-Linear Association	1.000	1	.000
Cramer's V	.000		.049

4.1.3. Investigating the Third Research Question

To examine the third research question on the presence of between-group differences between the expert and novice (meta)cognitive processing

regarding the quantitative assessment of IELTS speaking tasks, we used QSR NVivo 8 to analyze the collected data.

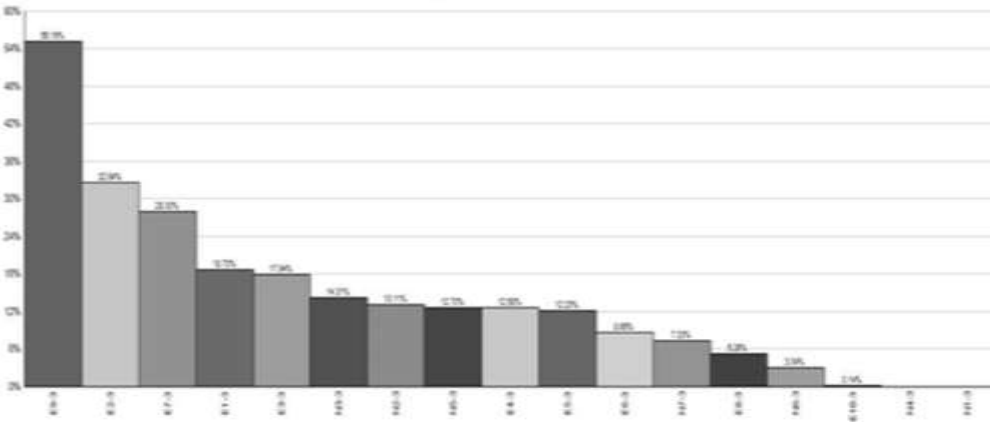

Figure 10. Fluency and Coherence: Phase 3 for Expert and Novice Raters

In Figure 10, Phase 3 of the verbal protocols is illustrated in detail for both Expert and Novice raters. Similar to Phases 1 and 2, Expert rater 4 has an exceeding frequency load of 55.19% for Fluency and Coherence while Novice raters 1 and 4 showed no attention (0.00%) to the Fluency and Coherence of IELTS speaking performance.

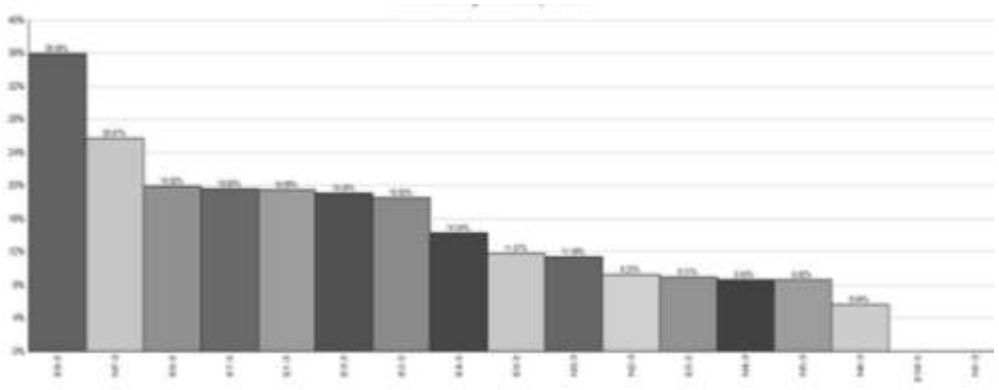

Figure 11. Grammatical Range and Accuracy: Phase 3 for Expert and Novice Raters

As illustrated in Figure 11, a relative decline of attention to Grammatical Range and Accuracy is shown in Phase 3, with Expert rater 8 showing the highest frequency load of 35.95%, while Expert rater 10 and Novice rater 1 with zero attention to Grammaticality of IELTS speaking input stood at the lowest position.

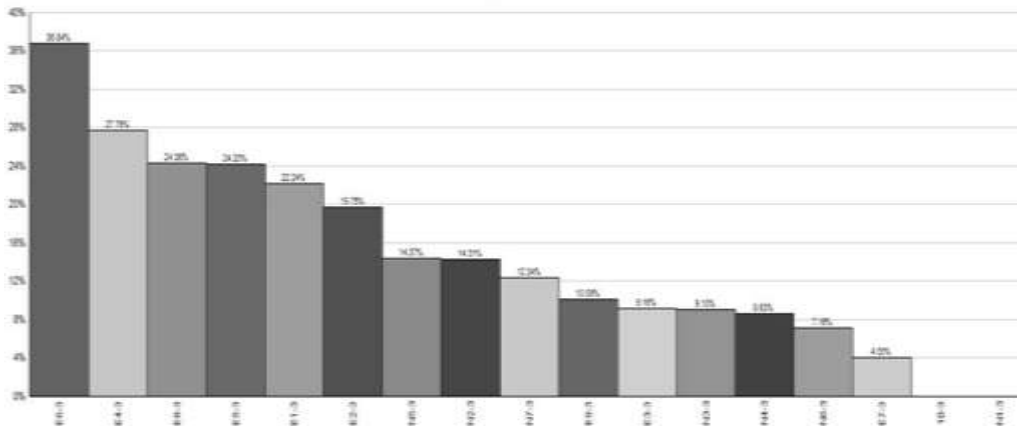

Figure 12. Lexical Resource: Phase 3 for Expert and Novice Raters

As illustrated in Figure 12, in Phase 3 of verbal protocols, due to the quantitative nature of the data, a natural decrease in the percentages is revealed and two expert and novice raters have paid no attention to Lexical resource. The highest amount belongs to Expert rater 8 (28.28%) while the Expert rater 10 and Native rater 1 showed minimum attention to the lexical choice of the IELTS candidate.

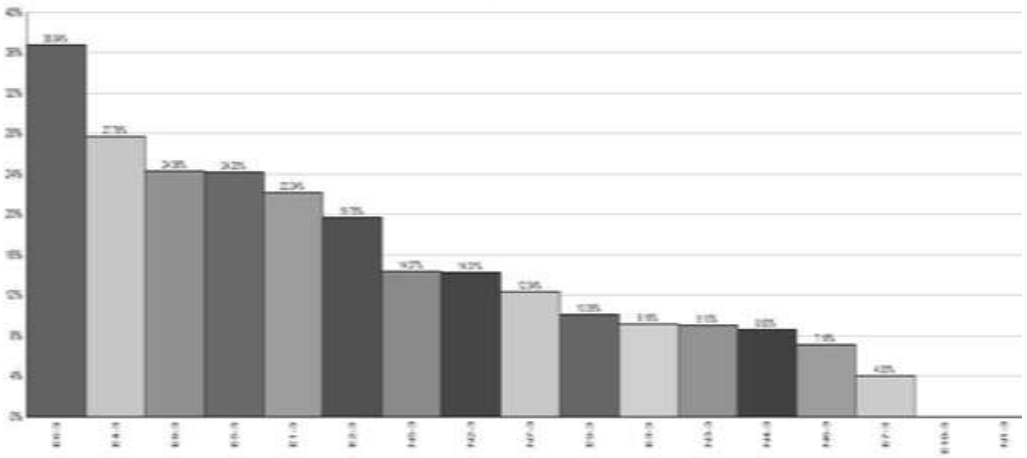

Figure 13. Pronunciation: Phase 3 for Expert and Novice Raters

In Figure 13, while Expert rater 10 and Novice rater 1 showed no attention to Pronunciation features (0.00%), the highest degree of attention belongs to Expert rater 6 with 36.84%. This time, the number of raters exceeding the IELTS Speaking Rubric pronunciation range 14.29% has increased to 8 out of 17 raters.

A 2 x 4 group-independence Chi-square test was performed to assess the relationship between the nature of expertise and the four categories of IELTS speaking rubric in Phase 3. A contingency table for these data is shown in Table 12.

Table 12
Contingency table for Expert vs. Novice Raters and IELTS Speaking Rubric Criteria in Phase 3

			Criteria				Total
			Fluency & Coherence	Lexical Resources	Grammatical Range & Accuracy	Pronunciation	
Rater	Novice	Count	10	8	10	11	39
		Expected Count	8.0	9.1	11.0	9.0	39.0
		% within Criteria	43.0%	33.0%	32.0%	44.0%	37.0%
	Expert	Count	13	16	21	14	64
		Expected Count	14.0	14.0	19.0	15.0	64.0
		% within Criteria	56.0%	66.0%	67.0%	56.0%	62.0%
Total		Count	23	24	31	25	103
		Expected Count	23.0	24.0	31.0	25.0	103.0
		% within Criteria	100.0%	100.0%	100.0%	100.0%	100.0%

As Table 13 shows, there was a statistically significant relationship between the raters' expertise and the types of criteria they used, while quantitatively assessing the IELTS speaking input in Phase 3 [χ^2= 1.000, $_{(3, 103)}$, p = 0.000, *Cramer's V* = .000].

Table 13
Chi-square test between Raters and Criteria in Phase 3

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.000[a]	3	.000
Likelihood Ratio	1.000	3	.000
Linear-by-Linear Association	.000	1	.000
Cramer's V	.000		.000

4.1.4. Investigating the Fourth Research Question

To examine the research question 4 on the presence of between-group differences between the expert and novice raters' cognitive processing regarding the revising/finalizing their assessment of IELTS speaking tasks, we used QSR NVivo 8 to analyze the collected data.

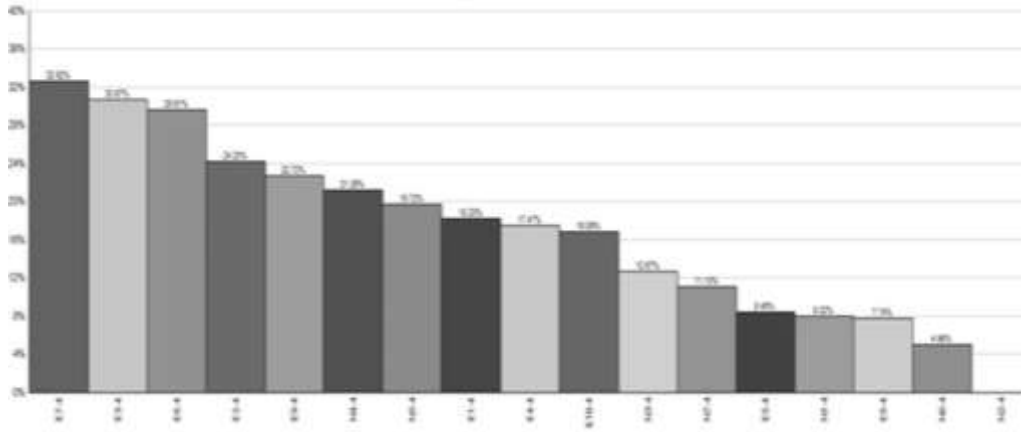

Figure 14. Fluency and Coherence: Phase 4 for Expert and Novice Raters

In Figure 14, six raters have exceeded the IELTS speaking rubric for Fluency and Coherence (14.29%). The lowest degree belonged to Expert rater 1 with 1.86% attention to Fluency and Coherence features. The highest, however, belonged to Novice rater 5, who has scored 25.41%.

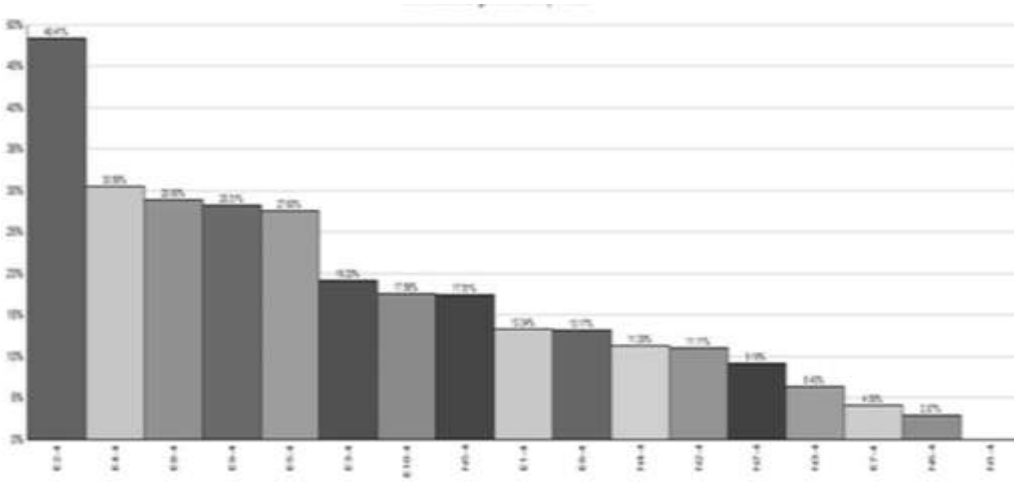

Figure 15. Grammatical Range and Accuracy: Phase 4 for Expert and Novice Raters

As Figure 15 shows, the highest degree of attention to Grammar belonged to Expert rater 2 with 36.02% and the lowest degree was for Novice rater 2 and Expert rater 9 both with .000% attention to the Grammatical Range and Accuracy.

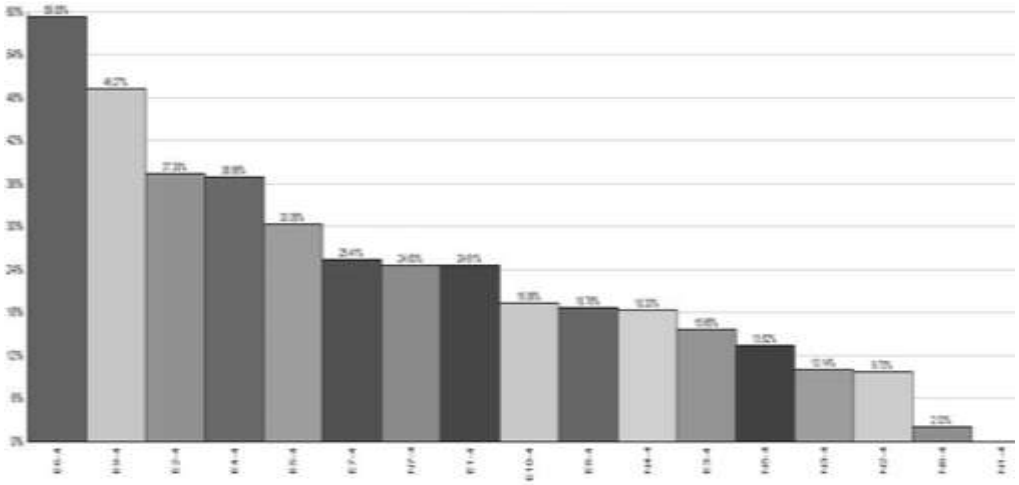

Figure 16. Lexical Resources: Phase 4 for Expert and Novice Raters

As can be seen in Figure 16, Expert raters 6 and 9 outperformed others with 59.35% and 49.27%, respectively. The lowest frequency counts of Lexical Resources belonged to Novice rater 1 with 0.00%.

Finally, as presented in Figure 17, 4 raters (Expert rater 10 and Novice raters 5, 1 and 4) showed no attention to pronunciation features (.000%). The highest amount increased up to over two times more than the IELTS speaking rubric pronunciation range (14.29%), that is, 41.51% by Expert rater 2.

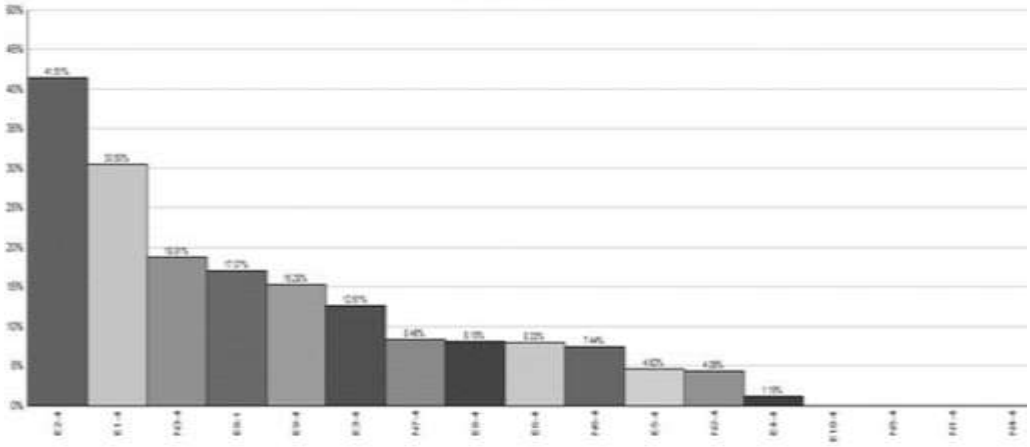

Figure 17. Pronunciation: Phase 4 for Expert and Novice Raters

A 2 x 4 group-independence Chi-square test was performed to assess the relationship between the nature of expertise and the four criteria of IELTS speaking rubric in Phase 4. A contingency table for these data is shown in Table 14.

Table 14
Contingency table for Expert vs. Novice Raters and IELTS Speaking Rubric Criteria in Phase 4

			Criteria				Total
			Fluency & Coherence	Lexical Resources	Grammatical Range & Accuracy	Pronunciation	
Rater	Novice	Count	16	20	13	10	59
		Expected Count	19.1	18.0	13.0	7.0	59.0
		% within Criteria	21.1%	27.0%	23.0%	33.0%	25.0%
	Expert	Count	60	54	42	20	176
		Expected Count	56.0	55.0	41.0	22.0	176.0
		% within Criteria	78.0%	72.0%	76.0%	66.0%	74.0%
Total		Count	76	74	55	30	235
		Expected Count	76.0	74.0	55.0	30.0	235.0
		% within Criteria	100.0%	100.0%	100.0%	100.0%	100.0%

As Table 15 shows, there was a statistically significant relationship between the raters' expertise and the types of criteria they used, while revising their assigned scores to the IELTS speaking input in Phase 4 [χ^2= 1.000$_{(3, 232)}$, p = 0.000, *Cramer's V* = .091].

Table 15
Chi-square test between Raters and Criteria in Phase 4

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.000[a]	3	.000
Likelihood Ratio	1.000	3	.000
Linear-by-Linear Association	1.000	1	.000
Cramer's V	.091		.000

4.2. Discussion

The present study aimed at analyzing Iranian novice and expert raters' cognitive processes while rating the IELTS speaking Task Two. Four research questions were raised on the possible differences among the expert and novice raters' cognitive representation of (1) the IELTS Speaking rubric, (2) the qualitative assessment of the response to Task Two in the IELTS interview, (3) the assigned quantitative scores to the exemplar response, and finally (4) revising/finalizing the assigned scores.

Rating constructed responses is a cognitively complex decision-making process, which involves information processing. Myford and Wolfe (2003) described the possible processes going through raters' minds. However, their study and many others mostly focused on the thinking processes without linking raters' cognition with rating outcomes. Hence, it remains unclear whether the differences in raters' cognitive processes such as expertise influence the rating quality and, more importantly, whether some processes may yield better and more accurate scoring decisions.

To bridge the gap, in this study, the researchers' general observation was the presence of noticeable and vast differences among the IELTS Speaking novice and expert raters both at the levels of their knowledge of the IELTS Speaking rubric and its reliable application. Such discrepancy surprisingly existed in spite of more or less similar final judgements they reached after rating the recorded response input. The divergence in the expert and novice raters' cognitive representations was observable not only at every phase of verbal protocols, but also fluctuated within individual ratings as the raters moved forward through the successive phases of verbal protocols.

Such fluctuations, however, were not always reflected in the finalised scores as they only had a thin margin of ± 0.5 in this study. This might be one of the aspects of the term "play-it-safe" strategy stated by Wolfe, Chiu, and Myford (1999), who claim that raters usually try to give the conservative scores (moving around 5-6 in IELTS scale) to have a safer rating while being monitored. In other words, they try to give scores which are not extreme to avoid later criticism; they also stated that inability to make fine distinctions between and among the criteria is what taints the raters' scores to be central tendency effect.

As Lumley (2002) emphasized, the rater is at the center of the rating activity. One of the rater factors that seem to play an important role in the rating process is rating expertise. Cumming et al. (2002), for instance, argued that the expertise and knowledge that "raters bring to the rating task are essential for a reliable and valid rating" (p. 15). There is a relatively extensive literature on the effects of rater expertise on ESL essay rating

processes (e.g., Erdosy, 2004; Sakyi, 2003). Such studies indicate that experienced and novice raters employ qualitatively different rating processes. Cumming, et, al. (2002), for example, found that experienced teachers had a much fuller mental representation of the essay assessment task and used a large and varied number of criteria, self-control strategies, and knowledge sources to read and judge ESL essays. Novice raters, by contrast, tended to evaluate essays with only a few of these component skills and criteria, using skills that may derive from their general reading abilities or other knowledge they have acquired previously (e.g., editing).

However, there is no research on how raters with different levels of expertise approach essay rating with different types of rating scales. Cumming, et, al. (2002) hypothesized that novice raters, unlike experienced raters, may benefit from analytic scoring procedures to direct their attention to specific aspects of writing as well as appropriate evaluation strategies and criteria, whereas Kim (2015) hypothesized that analytic scoring is easier for inexperienced raters, as fewer unguided decisions (e.g., weighting different evaluation criteria) are required.

Novice raters were reported to use more interpretation strategies and self-monitoring focus than the experienced raters who used more judgment strategies and rhetorical and ideational focus. Generally, novice raters were more dependent on the rating scales for rating criteria and decisions than were the experienced raters. They tended to refer to the rating scales and rely on criteria listed in the scales more frequently when making their scoring decisions. In addition, they tended to focus on specific, local aspects of writing more often and to spend more time interpreting and/or editing text than the experienced raters did (Sakyi, 2003).

These tendencies seem to be due to the novice raters' lack of experience with ESL writing, which might have led them to focus on local linguistic features in order to understand the texts before they could evaluate other aspects. In addition, because they lack established criteria for judging writing quality and/or how to approach the rating task, these novice raters may have relied on the rating scale more heavily and/or based their score decisions and justifications on simple or easily discernable aspects of writing such as lexis, syntax, and punctuation (Han, 2016).

Experienced raters, by contrast, reported more judgment strategies and rhetorical and ideational focus and tended to allot more time reading and assessing the essays overall, particularly in terms of rhetoric and ideas, than the novice raters did (Cumming, et. al., 2002). In addition, the experienced raters tended to refer to other criteria than those mentioned in the rating scales (e.g. length, writer's situation) more frequently than did the novice raters.

Similar to the findings of this study, training courses, which typically lead to expertise, are reported to be influential to decline judgment biases to a great extent. This means that when raters' bias is reduced, less interpretation strategies such as selective attention and making prediction are adopted; rather, more objective judgment strategies are used by expert and well-trained raters (Hamilton, Reddel, & Spratt, 2001). Rating behavior is reported in a couple of other studies to be enriched after training courses (Furneaux & Rignall, 2007; Shaw, 2002). Other instances of similarity are seen in a study conducted by Eckes (2011) indicating that the quality of rating behavior would undoubtedly face an undeniable improvement. However, he asserted that inter-rater reliability is not influenced by the instructions. Similar results in a wider scope are even found in the literature such as Fahim and Bijani (2011). They believed that not only does expertise have a significant impact on the way raters' cognitive processes are conducted, but also training courses lead to much lower rating variances.

A few studies, however, contradicted the findings of the current research. For example, Lim (2011) believed that raters having received no training, and with less expertise, have shown satisfactory rating behaviors. The effect of training in case of being prolonged is negligible and minimal (Knoch, 2011; O'Sullivan & Rignall, 2007). Severity of the raters, resulting from interpretative nature of strategies (specifically those adopted by novice raters) was not influenced by expertise as Lim (2011) found.

There was some evidence of a differential effect of rating scales across rater groups. There was a general trend among the novice raters to attend to specific linguistic features (e.g., lexis, error frequency, syntax, spelling) more frequently. It is possible that, because the scale lists several specific linguistic features (grammar, vocabulary, spelling, etc.) without any indication of their importance relative to each other or to other criteria, it led the novice raters to treat these features as multiple categories that need to be considered (and perhaps weighted and scored) separately. By grouping these aspects under one heading in the scale, the analytic scale seems to have led the novice raters to treat these specific aspects as one component rather than multiple categories that need to be considered separately. It would, thus, seem that variation in novice raters' ratings is due to a subjective cognitive scale choice.

5. Conclusion and Implications

The L2 speaking assessment investigated within cognitive processing frameworks as well as cognitive processes which portray raters' mental patterns are not widely found in the literature (Purpura, 2014). In other words, very few researchers have taken a cognitive processing (Purpura, 2012) approach to conceptualize the rating process of the L2 speaking. To the

best knowledge of researchers, almost none of those studies have adopted a cognitive processing approach to examine the raters' judgment processes, exploring the underlying processes while the expert and novice raters are attempting to understand the response input, formulate a mental representation of the response, compare the response representation with the rubric, and evaluate the response accordingly.

In this study, it was shown that the expert and novice raters focused on different aspects of the recorded response input and had different interpretations or applied different criteria when judging the recorded input. This study is of great significance in terms of its implications for training IELTS raters and inspiring researchers in L2 assessment to investigate raters' cognition in the light of the current models of human information processing and the functionality of their brain for L2 speaking assessment.

Susceptibility to bias and inconsistent patterns of judgment are a caveat to the raters of high-stakes speaking tests such as IELTS, which demand the raters' meticulous re-examination of the validity of test scores and the inferences that are to be followed. A thorough understanding of the raters' personal preferences, professional qualifications and other background factors such as race, gender, and educational profile is fundamental to IELTS raters so that they could better analyze raters' behavior or decision-making processes. Such factors serve to explain why raters assign ratings the way they do and what attributes or strategies they still need to improve their rating performance validity.

Since raters' cognition is a complex matter and needs to be conceptualized based on further empirical data, the research on L2 speaking raters' cognition can highly inform our individuals' understanding of the exact nature of rating variability of speaking performance and help them tackle the practical problems regarding test score validation and training L2 speaking raters.

Further research in the future needs to be done to examine how raters of L2 speaking assessment form cognitive representations of the scoring rubric. Researchers may consider examining (1) how the L2 speaking raters compare the L2 speakers to assessment criteria or against each other during the rating process; (2) how and why the L2 speaking raters have dissimilar perceptions of the components of the L2 speaking proficiency, and place weighted emphasis on a limited set of aspects and features instead of consulting the full scoring rubric; (3) how and why the L2 speaking raters focus on certain self-generated features that are not explicitly included or explained in the scoring rubric; (4) how and why the L2 speaking raters draw on inferences about a candidate's personality, maturity, world knowledge, and educational background to justify their scoring patterns and decisions; and (5) how and

why the L2 speaking raters incorporate their personal preferences into their decision-making, especially with respect to their language background and personal attitudes toward the response or the speaker.

References

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory*. New York, NY: Psychological Press.

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.

Berns M (2005). Expanding on the Expanding Circle: Where do WE go from here? *World Englishes, 24(1),* 85–93.

Brennan, R. L. (2013). Commentary on "Validating the interpretations and uses of test scores". *Journal of Educational Measurement, 50,* 74–83.

Bridgeman, B., Powers, D., Stone, E., & Mollaun, P. (2011). TOEFL iBT speaking test scores as indicators of oral communicative language proficiency. *Language Testing, 29(1),* 91–108.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple. *Language Testing, 29,* 19–27.

Cumming, A., Kantor, R., & Powers, D. (2002). *Scoring of TOEFL Essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series N 22). Princeton, NJ: Educational Testing Service.

Davis, L. (2015). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33*(1), 117-135.

Ducasse, A. M. (2010). *Interaction in paired oral proficiency assessment in Spanish: Rater and candidate input into evidence based scale development and construct definition*. Frankfurt: Peter Lang.

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

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt: Peter Lang.

Eckes, T. (2012). Operational Rater Types in Writing Assessment: Linking Rater Cognition to Rater Behavior. *Language Assessment Quarterly, 9,* 270–292

Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL*

compositions (TOEFL Research Report No. RR-03-17). Princeton, NJ: Educational Testing Service.

Ericsson, K. A. (2006). The Influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 683–704). Cambridge: Cambridge University Press.

Esfandiari, R., & Myford, C. M. (2013). Severity differences among self-assessors, peer assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18(*2), 111-131.

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,* 1–16.

Furneaux, C., & Rignall, M. (2007). The effect of standardization-training on rater judgements for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment* (pp. 422–445). Cambridge: Cambridge University Press.

Govaerts, M. J. B., Schuwirth, L. W. T., Van der Vleuten, C. P. M., & Muijtjens, A. M. M. (2011). Workplace-based assessment: effects of rater expertise. *Advances in Health Sci Educ 16,* 151–165

Hamilton, J., Reddel, S., & Spratt, M. (2001). Teachers' perception of online rater training and monitoring . *System, 29,* 505-20.

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.

Hsieh, C. N. (2011). Rater effects in ITA testing: ESL teachers' versus American undergraduates' judgments of accentedness, comprehensibility, and oral proficiency. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 9*, 47-74.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing, 28,* 179–200.

Kim, H. J. (2015). *Investigating raters' development of rating ability on a second language speaking test* (Unpublished doctoral dissertation). Teachers College, Columbia University, New York, NY.

Lazaraton, A. (2005). *Non-native speakers as language assessors: Recent research and implications for assessment practice.* Paper presented at the BAAL, Bristol.

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

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19,* 246–76.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective.* New York: Peter Lang.

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127-145.

Mislevy, R. J. (2010). Some implications of expertise research for educational assessment. *Research papers in education, 25,* 253-270

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'Sullivan, B., & Rignall, M. (2007). Assessing the value of bias analysis feedback to raters for the IELTS writing module. In L. Taylor & P. Falvey (Eds.), *IELTS Collected Papers: Research in speaking and writing assessment* (pp. 446–478). Cambridge: Cambridge University Press.

Purpura, J. E. (2012). *What is the role of strategic competence in a processing account of L2 learning or use?* Paper presented at the American Association for Applied Linguistics Conference, Boston, MA.

Purpura, J. E. (2014). Cognition and language assessment. In A. J. Kunnan (Ed.), *the companion to language assessment* (pp.1452–1476). Boston, MA: John Wiley & Sons, Inc.

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing, 28,* 463–481.

Sakyi, A. A. (2003). *A study of the holistic scoring behaviors of experienced and novice ESL instructors* (Unpublished doctoral dissertation). Toronto: University of Toronto.

Tosuncuoglu, I. (2018). Importance of Assessment in ELT. *Journal of Education and Training Studies, 6(9)*, 163-167

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.

Wolfe, E.W., Chiu, C.W. T., & Myford, C. M. (1999). *The manifestation of common rater effects in multi-faceted Rasch analyses.* Princeton, NJ: Educational Testing Service, Center for Performance Assessment.

Wolfe, E. W., Matthews, S., & Vickers, D. (2010). The effectiveness and efficiency of distributed online, regional online, and regional face-to-face training for writing assessment raters. *The Journal of Technology, Learning and Assessment, 10(1).* 1-22.