COLLABORATION IN WRITING ASSESSMENT: INSIGHTS INTO NOVICE AND EXPERIENCED IRANIAN EFL RATERS' CRITICAL THINKING AND CRITERIA

KAZEM JAHANSHAHI[1]
ALLAMEH TABATABA UNIVERSITY (TEHRAN, IRAN)

ABSTRACT

There is a growing body of literature on effective techniques through which critical thinking (CT) can be fostered in an educational setting. In a way that agrees with teaching as an interactive process, considerable debate has taken place over enhancing teachers' CT abilities in recent years. Concerning the positive effects of collaboration on CT, the present study examines whether the frequency of CT components among novice and experienced Iranian EFL raters is different while assessing a passage individually and collaboratively (in pairs). For this end, a mixed-method approach was adopted. Writing assessments made by two groups of novice and experienced raters were coded in search of statements classified as one of the five components of CT (Stapleton, 2001). The data obtained included recorded think-aloud protocols and written comments made by raters while assessing 3 passages written by IELTS examinees. Then, novice and experienced raters were compared based on the total frequency of CT components used while assessing individually and collaboratively. Afterwards, their criteria for scoring a passage (without using a rubric)were attempted to be established. Findings revealed that collaboration in writing assessment improves raters' CT skills irrespective of their experience level, while CT indices were significantly higher for experienced collaborative pairs. Although, novice and experienced raters' CT skills were not significantly different while assessing individually.With regard to the criteria applied by novice raters, qualitative analysis revealedcentral concerns about the text surface level. Actually, they were believed to be in the process of forming their personal strategies and establishing their own criteria as raters. While experienced raters went one step ahead of syntax into semantics.

Key words:*Critical Thinking (CT); Novice Raters; Experienced Raters; Individual Assessment.*

RESUMEN

Hay una creciente literatura sobre técnicas eficaces a través de las cuales el pensamiento crítico (PC) puede ser fomentado en un entorno educativo. De una manera que concuerda con la enseñanza como un proceso interactivo, existe un debate considerable sobre la mejora de las capacidades de PC de los profesores en los últimos años. En cuanto a los efectos positivos de la colaboración en el PC, el presente studio analiza si la frecuencia de los componentes del PC entre evaluadores noveles y experimentados de EFL ironies diferente mientras se evalúa un pasaje de forma individual y colaborativa (en parejas). Para ello, se adoptó un enfoque de métodos mixtos. Las evaluaciones escritas realizadas por dos grupos de evaluadores noveles y experimentados fueron codificadas en busca de declaraciones clasificadas como uno de los cinco componentes del PC (Stapleton, 2001). Los datos obtenidos incluyeron protocolos de pensamiento en voz alta y comentarios escritos hechos por evaluadores mientras que evaluaron 3 pasajes escritos por examinados de IELTS. A continuación, los evaluadores principiantes y experimentados se compararon en base a la frecuencia total de los componentes de PC utilizados al evaluar individualmente y en colaboración. Posteriormente, se intentó establecer su criterio para anotar un pasaje (sin usar una rúbrica). Los hallazgos revelaron que la colaboración en la evaluación de la escritura mejora las habilidades de los evaluadores de PC independientemente de su nivel de experiencia, mientras que los índices de PC fueron significativamente mayores en parejas experimentadas. Aunque, las habilidades de PC de los evaluadores principiantes y experimentados no fueron significativamente diferentes al evaluar individualmente. Con respecto a los criterios aplicados por los evaluadores principiantes, el análisis cualitativo reveló preocupaciones importantes sobre el nivel superficial del texto. En realidad, se cree que están en la línea de formar sus estrategias personales y establecer sus propios criterios como evaluadores. Mientras que los evaluadores experimentados fueron más allá de la sintaxis en semántica.

[1] Email: jahanshahi2013@yahoo.com

Palabras clave: Pensamiento Crítico (TC); Evaluadores principiantes; Evaluadores experimentados; *Evaluación individual.*

1. Introduction

As is well known, education is of crucial importance for the development of a society.In this regard, although societies are developing at a surprising speed, and it is expected that the content of educational programs vary with those developments,it is not the case. Therefore, there is stilla long way to be taken to achieve our educational ideals.Together with thechangesoccurred in societies, teachers need to take other new roles.At the modern situation, they are required to change their mission as mere transmitters of knowledge and absolute authorities.They need toinvestigate the best methods and techniques to incorporate into their classrooms to improve autonomy, reflectivity, and CTskills among learners.For this end, although most of them are targeting on students,considerable debate has taken place over whether incorporation of different forms of CT in education is helpful.At the present situation, teachers' abilities would be of great importance, because they are playing an important role in preparing reflective students to take part in society. But how teachers' CT skills can be fostered?

Among various other methods, collaboration has always been praised for playing an important role in preparing reflective students.Regarding its effects on improving educational system quality, numerous studies have been conducted; although,nearly all invested on learners and there is little empirical evidence about whether it hasthe same effectiveness on teachers' success. To meet this issue and answer the raised question, a better understanding of the situation concerningthe relationship between teachers' experience level, collaboration in writing assessment and its effects on theirCT skills, and their rating criteriacan shed some light onmethodsthatmay improve reflectivity,CTand assessment criteria among raters.

2. Theoretical Background

2.1. Assessment

Traditionally, assessment could be described as a mere quantitative device used just for summative purposes, known as a motivated activity (Somervell, H. 1993). In such aperspective, great effort should be made to assure that the assessment procedure satisfiesthecriteria like reliability, validity and generalizability. Together with the changes of teaching and learning atmosphere in which students and their needs started to get more and more attention, and "as the goals of education have become more and more complex and the number of students have enormously increased" (Farhady, et al. p.1), evaluation has, accordingly, become much more difficult. To Poehner (2008), this new era of assessment is qualitatively different from how it was traditionally. In its dynamic sense, the new assessment is a critical and cooperativeactivity. Among the recently devised assessment methods, Somervell (1993) describes collaborative assessment as an alternative form of assessment which is contrary to its traditional form, and compares it with other methods in the following figure.

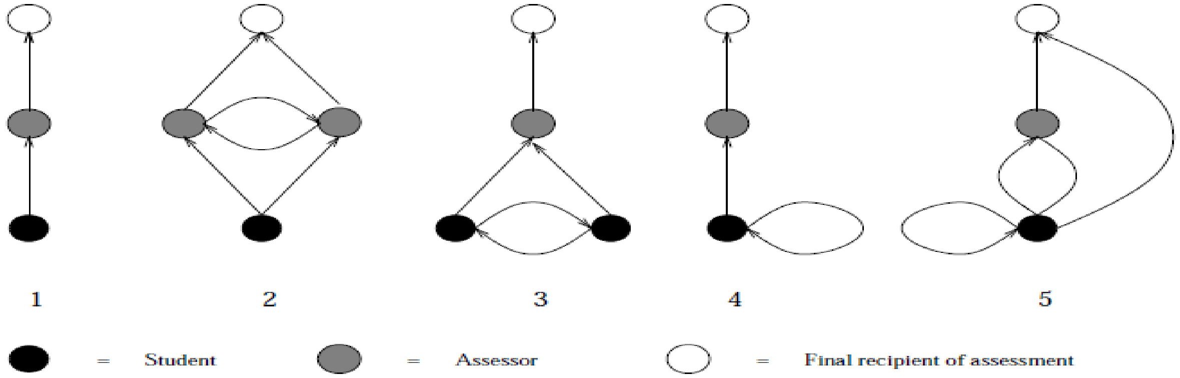

Figure 1. Different Assessment methods (Somervell, 1993)

Here, part one is the traditional form of assessment in which teacher is the only source of knowledge and power. The second collaborative method is explaining the type studied and worked upon in the present study. In this form, two assessors examine the same work collaboratively. The third and the forth diagrams can be referred to as peer–assessment and self–assessment, and the last one which is another form of collaborative assessment by a teacher and a student.

Instructorsgive studentswriting assignments for a variety of reasons. They canincreasethe mastery of a particular body of knowledge,or foster writing skills and intellectual growth such as CT skills and reflectivity. Somewriting assignments may involve all of these goals (Carroll, 2007). A large body of research exists (Carrol, 2007; Deremer, 1998; Ghafarsamar & Ahmadi, 2012; Knoch, 2009; Lumley, 2005) that explores the process of writing assessment by teachers as raters. Most of these studieshave occurred in classroom settings and investigated theeffects of different instructional tools on the assessmentsquality. One repeatedly reportedfindingis that rubrics, rating scales and training sessions can help raters to reach to a detailed understanding of the specific factors that improve the quality oftheirwriting assessment (Axelrod, et al., 1994). Thiscan developinsights that will result in the introduction of more tangible measures to manage potential areasoft lack within the context of writing assessment and help raters, and especially less experienced raters, to develop better rating strategies.

To find a systematic answer to the question of how raters actually go about scoring students' writings in practice, numerous investigations have been carried out over the past two decades. Those studies have, especially in the recent years, greatly improved our understanding ofthe knowledge, criteria affecting decisions, and thinking that raters use toassess and score writings. Different methods and formal elements also havebeendevised to help raters, but suchmaterials are only mechanicallyguide writing assessments. Therefore,overemphasizingand overusing themcan decline the value andcomplexity of students' writings and abilities.It also underestimates the raters' creativity in their job. However, an understanding of new methodsthat improve writing assessmentwill provideinsights for ratersandintroduce them more tangible measures to help find the potential areasof lack within the context of the standards in writing assessment.

2.2. Raters and the Concept of Experience

As stated by Hattie (2003), it should be asked where the major sources of variance in student's achievement lie, andconcentrate on enhancing them to truly make the difference.Those variance sources are divided into six categories; but, there is a strong consensus that among the variables improving student achievement, the teacher matters more thanany other single factor (accounting for about 30% of the variance)(Rowan, Correnti& Miller, 2002- cited in Mei, 2009). Taking a close look at the literature, it isquite evident that the focus of discussions are more around theinfluences of the students' home, structures of schools, and schoolbuildings. We also hear so much about the reduced class sizes, new examination methods, curricula and etc., while the teachers' role havesomehow not been invested on appropriately. We need to ensure that this greatest influence is working appropriately to havepowerful and positive effects on learners.

Writing assessment is an example of a subjective scoring task. Despite the existing standardized training procedures, there is not acriterion solution forwriting assessment. Typically, raters are trained to agree with each other byreaching to an agreement based on a scoringrubric and benchmark texts, without being able to use any standard or objective solutions. As such, Deremer (1988) reports that they are presented with a task environment (i.e., assess the texts using agiven rubric), but they are not given a ready-made representation of the rating activity. Given the importanceof writing assessment in any educational system, raters must develop their own plan of action, and this needs time and experience. To help with the experience issue, Lieberman & Miller (2008) suggest adopting such strategies as flexible schedulingas well as pairing novice raters with more experienced ones in co-teaching/ testing models that, according to the literature, are among the more successful approaches.

2.3. Critical Thinking

CT can be traced back to more than two thousand years ago when Socrates carefully questioned people from different dimensions and found that they couldn't always justify their claims (Wright 2002). Although, as one of the most pervasive academic literature terms, CT has not clearly been defined yet. For example, Paul (1990) simply describes CT as a unique and purposeful thinking which is systematic and habitual, while Facione (2011) writes more than one page to define it. However, nearly all definitions refer to the same concept and have their major points in common. No matter how they define it, nearly all successful teachers try to be involved in an ongoing reflection process, and approximately everyone agrees that CT has begun to play an outstanding role in education and has turned into one of its main goals.

In the same line with teaching as an interactive process between society and classroom, considerable debate has taken place over enhancing learners' CT abilities in EFL/ESL context in recent years. A number of researchers (e.g. Bataineh & Zghoul, 2006; Brown, 1994; Dinkelman, 1999; to name a few) claim that the classroom environment must provide modeling, rehearsal, and coaching for students to develop a capacity for informed judgments, for without CT systematically being integrated into instruction, learning is transitory and superficial. Accordingly, a body of research carried out claiming that various activities should be incorporated into classrooms to combine with reflective teaching and improve CT to enhance teaching quality. Findings revealed complementary and sometimes conflicting results in terms of the relationship between CT and such many other factors as assessment and evaluation (Ghafar Samar & Ahmadi, 2012; Stapleton, 2001), age and gender (Ramasamy, 2011; Soeherman, 2010), writing and speaking (Dantas- Whitney, 2002; Ghahramani- Ghajar & Mirhosseini, 2005), computer and technology (Burges, 2009; Tsang, 2011), assessment methods (Lynch, 2001), achievement (Birjandi & Bagher Kazemi, 2010), cognitive development (Facione, 2000), teachers' CT (Birjandi & Bagher Kazemi, 2010; Lishchinsky, 2010), disposition (Ramasamy, 2011), and teaching CT (Choy & Oo, 2012; King, 2010; Philips, 2010). This feature, the ability to be involved in an ongoing reflection process, then turned out to be a prominent characteristic of a successful educational system. Accordingly, some other studies (Ghahremani-Ghajar & Mirhosseini, 2005; Stapleton, 2001; Stout, 1993; Twardy, 2005) were conducted to improve teachers' CT skills, as those who are responsible to trigger and improve such students' abilities, as well. However, according to Ghafar Samar and Ahmadi (2012), "Because the priorities in today's classrooms include learners' CT abilities, little attention has been paid to this skill from the teachers' side as practitioners and mentors of these abilities" (p. 3).

2.4. Two Research Issues

Investigating the effects of collaborative assessment on enhancing Iranian EFL raters' CT skills by comparing the effectiveness of individual versus collaborative writing assessment was the first issue we focused on in the present study. Moreover, its potential influences on the CT of two different participant groups, novice and experienced raters, were investigated to find out the relationship between experience as a moderator variable and CT in a collaborative assessment context. The second point of focus here was studying different criteria and strategies applied by novice and experienced raters in rating sample writings without using any rubric or similar tools. Below, a more detailed image of the model through which the frequency of CT components of novice and experienced raters were measured is provided. Specifically, we set the following as our guiding research questions:

1. Is there any significant difference in the frequency of CT components among novice and experienced raters as a result of being engaged in individual and collaborative writing assessment?
2. What are the criteria applied by novice and experienced raters for assigning scores in a writing assessment and how they are comparable?

3. Methodology

The nature of the questions raised here requires a mixed-method approach and comprises both quantitative and qualitative procedures in collection and analysis of the data. For this end,2 groups of novice and experienced Iranian EFL raters were included, each containing 16 members with approximately the age range of 25 to 35. All participants were enrolling at teaching elementary to advanced levels of general English, at least 6 hours per week, at different English language institutes in Tehran, Iran. One of these groups consisted of novice raters with 1 to 2 years of experience in English language teaching, and the other comprised experienced raters with the minimum 5 years of teaching background. Both groups were given the same passages written by IELTS examinees to rate individually and collaboratively (in pairs). To make sure that they would participate in both phases of the experimentation– individual and collaborative assessment− attempts were made to choose pairs of raters who were working at the same place.

The sample writing papers applied for assessment included 3 passages written by general IELTS examinees. These passages rated individually and collaboratively by both groups of novice and experienced raters. After each passage, a blank space was provided to be filled by raters with a score and to clarify on their scoring criteria. To find their criteria, both groups of novice and experienced raters were asked to elaborate on their strategies for scoring each passage. Then, in accordance with the model proposed by Stapleton (2001), the collected data containing written and recorded think-aloud protocols were coded and scrutinized in search of CT components. The purpose here was to study the correlation between different assessment methods-individual and collaborative assessments-and the effects of those methods on novice and experience draters' CT skills.

3.1. The Research Procedure

Basically, the participants went through the following scoring steps. Here, all raters were engaged in a two-section writing assessment program. Firstly, the steps followed such as: introducing the IELTS writing texts, clarifying on individual and collaborative assessment methods, the manner of providing comments and criteria for scoring each passage, and verbalizing and recording thinking aloud while assessing each passage were clarified for raters. To prevent the negative effects of such factors as time restriction, fatigue, affective factors, and any other potential causes which may threat the reliability of the data, the raters were asked to submit their assessments after an acceptable time interval for both participants and there searcher. Because the participants here were two different groups of raters-novice and experienced-it was tried to choose pairs of novice and pairs of experienced raters who were working together in the same institute at the time of the data collection, so that they did not have any problem conducting the cooperative phase of the assessment. Therefore, the random assignment of the raters to pairs was not possible.

To collect the data, every rater took part in two assessment sessions. First, they were asked to score the passages individually while verbalizing and recording their thoughts. They were also asked to write their comments regarding each text and their criteria for assigning scores at the blank space provided after each passage. It was clarified that after assessing the sample writing papers individually, collaborative raters-pairs of novice raters and pairs of experienced raters-assess the same texts on another blank copy cooperatively while taking the same steps of assigning scores, writing their comments and criteria, and recording their thinking aloud protocols. Then, the data were analyzed in both novice and experienced groups looking for CT components and the raters criteria. The model applied here is what Stapleton (2001) has described as a logical way for measuring CT, with the purpose of discovering the level of raters 'CT while assessing a text individually and collaboratively. At this point, the participants were asked not to base their work on any scoring rubrics; and consequently, carry it out based on the same criteria they usually use while assessing compositions.

3.2. Coding the Data

The data was analyzed both quantitatively and qualitatively. First, the recorded think aloud protocols were carefully coded and transcribed in search of CT components based on the model dividing CT into the following five parts (Stapleton, 2001):

a. *Number of arguments:* which according to Stapleton, means either being "agree" or "disagree" with the passage itself or something mentioned in it. Itis usually being done by using claim markers such as "I think", "in my opinion", and assertions made by such modals as "should" and "must" followed by phrases such as "because", "due to", and "for this reason".

b. *Extent of evidence:* This part can be of different forms containing personal experiences, statistics, research studies, analogies, pointing out consequences, citing authorities, and precisely defining words.

c. *Recognition of opposition:* It includes recognizing opposite views in order torefuse or challenge them. For example, using expressions such as "even if....., ..." .

d. *Corresponding refutations:* This part refers to refusing a seemingly incorrect claim using conjunctive devices such as "although", "however", and "even though" among others.

e. *Number of fallacies:* As defined by Stapleton (2001), fallacies can be defined as different types of errors in reasoning. They usually occur when the reason for something does not adequately support the claim in one of a number of ways, such as oversimplifications, irrelevant assertions, wrong guesses, etc.

The collected data from both individual and collaborative groups were carefully coded and transcribed in search ofCT components. The decisions made upon analysis were a single expression or several utterances with a single aspect of the event as the focus. Such units may contain just one clause or many clauses focusing on a dominant component of CT. For further elaboration, consider the following examples:

3.3. Novice Raters' Comments

Rater A: *I think this passage is good, because grammar was a kind of ok. There were no serious grammatical mistakes and the concluding part was good. I think that if a text is grammatically ok, the majority of the work is done appropriately to reach to a good conclusion and get an acceptable score.* (Argument, individual assessment)

Collaborative Raters G & H: *I think here "radical" and "serious" have the same meaning and it's better to delete one of them, because it is unnecessary repetition*. (Argument)
-*I don't think so; I have seen some texts that writers prefer to use synonyms to make their sentence more beautiful and effective. Let's say to make the meaning rich.* (Extent of evidence)

Rater H: *Who says that no sensible person can deny the importance of breaking down barriers between countries? It doesn't make sense; we may have sensible persons who can break down the barriers between countries.* (Recognition of opposition, Individual assessment)

3.4. Experienced Raters' Comments

Rater C: *I think this paragraph which is supposed to do support the main idea is completely irrelevant with the concept, because it doesn't have anything to do with culture.* (Argument; individual assessment)

Rater G: "That brought with them pollution…" *that's a Farsi structure … in Persian we say 'kebakhodeshunaloudegiavordan…'* (Extent of evidence, Individual assessment)

Raters C & D: *We have a completely Farsi structure here… 'mesleinekebegimtarafofoqefekreshrobalaborde'…the writer has exactly translated a sentence according to Farsi structure. I'm dead sure that if an English speaker reads this sentence, he won't understand it.* (Fallacy)

After determining the frequency of CT components used by both groups, the criteria applied by novice and experienced raters were tried to be established by analyzing and comparing the comments provided at the end of each passage.

3.5. Findings

This part provides results of the analysis of the data gathered from think aloud protocols and written comments amongst 32 novice and experienced raters. Findings are presented in five principal sections corresponding to the research questions raised before. Briefly, those sections discuss the following aspects of the rating process:

- Does collaborative writing assessment improve novice raters’ CT skills?
- Does collaborative writing assessment improve experienced raters’ CT skills?
- How are individual novice and individual experienced raters comparable regarding the total number of CT components applied in writing assessment?
- How are collaborative novice and collaborative experienced raters comparable regarding the total number of CT components applied in writing assessment?
- What are the criteria applied by novice and experienced raters for scoring a passage?

To find any significant difference in the frequency of CT components made by novice raters in individual and collaborative assessments, the total number of components identified from individual (433 components in 125 minutes of discussion) and collaborative (285 components in 50 minutes of paired discussion) raters' assessments were collected. Table 1 shows the frequency of comments made with reference to the five major components of CT:

Table 1: Observed, Expected, Residuals and Percentage of CT Components Made by Novice Raters

Novice Raters’ CT Components		Observed N	Expected N	Residual	% of Total
Individual Assessment	Argument	70	83.8	-13.8	16.16
	Evidence	116	105.4	+10.6	26.78
	Opposition	118	125.8	-7.8 27.25	
	Refutation	129	115.0	+14.0	29.79
	Fallacy	0.0	3.0	-3.0	0.0
Novice Raters’ CT Components		Observed N	Expected N	Residual	% of Total
Collaborative	Argument	70	56.2	+13.8	24.55
	Evidence	60	70.6	-10.6	20.50
	Opposition	92	84.2	+7.8	32.28
	Refutation	63	77.0	-14.0	21.10

Assessment	Fallacy	5	2.0	+3.0	1.57

Here, collaborative raters exceeded the expected rate in employing such components of CT as *arguments*, *recognition of opposition* and *fallacy*, while individual raters outperformed conversely in *selection of evidence* and *refutation* choices. To probe any significant difference between novice raters' CT skills in individual and collaborative groups, Chi–square analyses were run. Although no significant discrepancies appeared regarding each individual component of CT, the total Chi–square observed value (x^2=21.27, *df*=4, p≥.000) was higher than the critical value of 6.48 at 4 degrees of freedom. Hence, there was a significant difference in the frequency of CT components among novice raters as a result of being engaged in individual and collaborative assessments. Accordingly, collaboration in writing assessment improved novice Iranian EFL raters' CT skills. The effect size of .17 for Chi–square observed value also supported the fact that the result was statistically significant and meaningful, as the effect size of .14 and above is considered strong based on the classification provided by Cohen (1998).

Table 2: Observed, Expected, Residuals and Percentage of CT Components Applied by Experienced Raters

Experienced Raters' CT Components		Observed N	Expected N	Residual	% of Total
Individual Assessment	Argument	104	185.1	-81.1	14.81
	Evidence	136	145.7	-09.7	23.46
	Opposition	185	162.0	+23.0	26.35
	Refutation	242	195.6	+46.4	34.47
	Fallacy	5	13.6	-08.6	0.71
Experienced Raters' CT Components		Observed N	Expected N	Residual	% of Total
Collaborative Assessment	Argument	249	167.9	+81.1	39.8
	Evidence	153	132.7	+20.3	17.58
	Opposition	124	147.0	-23.0	19.46
	Refutation	131	177.4	-46.4	20.56
	Fallacy	21	12.4	+08.6	3.29

Regarding experienced raters, the total number of CT components (672 cases out of 198 minutes by individual raters, and 678 ones out of 86 minutes by collaborative raters) were separately counted and displayed in table 2. Collaborative raters outperformed the expected rate in three components of *arguments*, *evidence* and *fallacy*. Chi–square observed value (x^2=111.221, *df*=4, p≥.000) here was higher than the critical value at 4 degrees of freedom. Accordingly, there was a meaningful difference in individual and collaborative performance. Therefore, there existed a significant difference in the frequency of CT components. The effect size of .30 for the Chi–square observed value also proved that the result was statistically significant and meaningful, supporting the fact that collaboration in writing assessment improved experienced raters CT abilities.

To answer the question regarding individual novice and individual experienced raters, Chi–square observed value (x^2=6.344, *df*=4, p>.05) was not higher than the critical value. Due to that, novice and experienced raters were not different regarding the total frequency of CT components applied while assessing a text individually, which is an indicative of the beneficial role of collaboration in assessment in improving raters' abilities. Accordingly, experience doesn't play any significant role on raters' critical reflection indices in an individual writing assessment. While for collaborative raters, the total amount of Chi–square observed value (x^2=28.718, *df*=4, p≥.000) was significantly higher than the critical value and then, novice and experienced raters performed differently. As a result, collaboration in writing assessment improved experienced raters CT skills more than their novice colleagues.

3.6 Assessment Criteria

To discover the criteria applied by novice and experienced raters while scoring a passage, since exploratory in nature, no inferential statistical analysis could be run. Instead, it was tried to find the answer by analyzing the written comments and coding the recorded think aloud protocols. To find any information regarding the raters' preferences in scoring and their criteria for assigning scores, scoring rubrics were provided neither for individuals nor for collaborative pairs. In contrast, they were requested not to base their assessments on any of such scales, but to construct their own criteria.

3.6.1. Novice Raters' Criteria

With relation to novice raters' criteria, a common specific strategy could not be found. However, the strong tendency to follow the single principle of a central concern about mechanics of writing including grammar and punctuation affected the frequency of the related rater behavior by novice raters. Although some other criteria were also applied once in a while, the main concentration was on the text surface level. For further elaboration, consider the following comments made by novice raters:

Rater A: *This passage is acceptable because grammar was a kind of ok, there were no serious grammatical mistakes, and the conclusion part was good. I think that if writing is grammatically ok, the majority of the work is done appropriately to get an acceptable score.*

Rater N: *Some serious grammatical mistakes and spelling errors, punctuation, and accuracy of the text are its major problems.*

Rater I: *Grammar was ok and there was no mistake on punctuation, I liked it.*

Rater G: *This part is fraught with grammatical mistakes. It's a kind of disaster, so many mistakes that we can't take care of them one by one. I'm not sure whether the writer's level is high enough to be an IELTS candidate, because the most simplistic structural rules are not stuck to here. I also think that a passage should be composed of simple, compound and complex sentences, something that we don't see here. This reasons decrease the score a lot.*

Indecision Phase

Another problem existing among some of the novice raters was an inability to make a final decision. Most of the novice raters based their scoring criteria on grammatical issues, but some knew at the end that something was wrong. There were some areas of obscurity which they failed to address and the reason is that they were in the process of forming their own criteria and needed to experience some new strategies, something that led to an indecision phase in the scoring process. An example is provided below where the rater cannot easily reach to a final conclusion to score a passage:

Rater J: *Grammar was not very good. We see some grammatical and punctuation mistakes. So I would give him perhaps 3 out of 5. But the content was interesting and I liked it. I would really like to increase it to 4 because of the content, but I don't think that it's entirely a 4, because it's full of mechanical mistakes. But I still like the passage. I don't know what to do.*

There were also a set of other specific personal strategies. For example, one of the raters widely believed in the use of complex vocabulary as an important feature of a good writing, saying that using sophisticated vocabulary convinced him to give a high score to the passage, regardless of its grammatical mistakes. Another was strongly in favor of the texts in which the main idea is being supported by various examples and explanations, no matter what kind of vocabulary be used. In another interesting case, participant 'G' had a very strong tendency to insert his personal attitudes to the text by changing all sentences, structures, vocabulary and collocations, and replacing them with his suggestions. He used the expressions such as "It is more interesting to use this structure, vocabulary, or sentence..." or "I think it's better to change this part to…" over and over again while rating each passage. He liked to change everything in a way that by inserting those changes, we would have a completely different text than the original one, at least in its surface level. It refers to the effect of the rater's personal attitudes on his rating strategy. Accordingly, a shared criterion used by all raters couldn't be met, except their common central concern for surface structure. Reading between the lines, they were in the process of forming their personal strategies and establishing their own criteria as raters.

Analyzing the data gathered here can provide us with the kind of feedback required to help improving novice raters' strategies for writing assessment. Hereby, we are able to give them more practical help to acquire the knowledge and skills of the experienced raters through proper training strategies. At a broader level, an awareness of the strategies applied by novice raters can help the developers of rubrics, scoring descriptors and training courses.

3.6.2. Experienced Raters' Criteria

Novice raters started their assessment from surface features and mostly focused on error identification and observable mechanical problems. In contrast, the methods applied by experienced raters were different and much more rule governed. Here, nearly all raters paid little attention to the text surface level and had a set of other priorities. They went one step ahead of syntax into semantics and treated writing as a process rather than a product. Mostly, they expressed concern about the main idea of the passage to be supported, discussed about their expectations from the writer, the text organization, task achievement, and finally the number and the order of paragraphs. Although, except for one expert IELTS rater, none of the other raters' assessment methods were as complete and to the point as a rubric based assessment; however, their criteria was very much more comprehensive while comparing to the novice raters. For further elaboration, consider the following examples:

Raters C&D: *There is no need to plural s here, but making such mistakes is not so important. Let's see what kind of information we can get from the passage.*

Rater O: *I actually examine each passage from four different dimensions of; content, grammar, organization and vocabulary. Then, I assign each one a score based on my observation of the text regarding each criterion.*

Rater I: *It just said that there will be tension and problem. Which tension? It is not to the point. It is beating around the bush. Go to the main idea and support it, bring examples. With using examples you can elaborate and explain.*

Raters A&B: *The biggest problem regarding this passage is its organization, because we don't have fluency of thought in it. You know, grammatical and vocabulary mistakes do not hinder understanding; they're not an issue here, but there is a lack of understanding, I don't know what the writer wants to say.*

As evident in the above discussions, the rating process is much more complex than it seems. There fore, raters need to be acutely conscious of the sources of variation and know the problematic areas of their assessment which can be addressed with practice and application of the experience of expert raters gained from extensive teaching and testing. These kinds of knowledge could be better acquired during rater training sessions and continuous exercise than by using motionless descriptors. Anyway, McNamara (1996, p.18) believes that "even with proper training, substantial differences between raters will persist with important consequences for the candidate…rater differences are reduced by training but (they still) do persist."

4. Discussion

In recent years, students' CT has occupied the minds of researchers as an important tool for reflection in EFL/ESL contexts; however; the role of teachers' CT has been largely neglected. On that ground, the present study tried to open a new window of opportunity for related teacher studies where and when the students' CT is a matter. Therefore, the author decided to study the effects of collaboration in writing assessment on teachers' CT skills. Also, the criteria applied by novice and experienced raters in a subjective writing assessment context were tried to be formed. This will help raters, especially novice raters, to have a better grasp of their own CT indices and rating criteria while assessing writing, compare themselves with more experienced raters regarding those issues, and let them know about their problems and the truth that what should be done, especially in problematic areas, in a marking process.

Teachers are the main members who have the responsibility of implementing assessment in all educational settings. Therefore, it is important to make sure that they are provided with the opportunity to obtain the necessary skills they need to conduct high quality assessments. It was found in the present study that collaboration in writing assessment improves teachers' CT skills, and this improvement had a direct relationship with their experience level. Accordingly, the more experience the better for teachers to act more appropriately while assessing a passage. Not only did the experienced raters perform better than their novice colleagues, they also did so with less effort. The main reason is believed to be that the cognitive skills become automatic with extensive practice.

Among the main concerns in writing assessment, the proper choice of rating criteria is of paramount importance to enable raters decide on the students' proficiency levels appropriately. For this purpose, teachers need to be equipped with specific tools. Accordingly, attending at formal degree courses, going to training workshops and highlighting various issues in the process of writing assessment can address some of their needs. However, the theoretical knowledge acquired by these methods also need to be supplemented by relevant practical experience in developing assessment skills. It is also practical that the model utilized by the experienced raters be used in training novice raters so that they may acquire some of the skills, if not all, which are necessary to improve their assessing performance. Collectively, it is expected that the model provided here regarding CT and assessment criteria could help raters- and also rater trainers- to develop a full understanding of the writing assessment mechanism. However, hope is held that this study adds to the debate and discussion about how best assessing skills can be developed. The findings also can be added to a growing body of literature on effective techniques through which CT can be fostered in an educational setting- including language teaching. In conclusion, the researcher believes that the raters, regardless of their experience level, are able to use the suggestions as a starting point for the development of their own CT and assessment criteria, and incorporate them into their own classes. And finally, expert instructors need to bear the responsibility to facilitate a continuous growth on the part of less experienced raters; otherwise, they would simply be left on their own to figure it out.

5. Conclusion

Nobody denies that EFL learners are dealing with complex cultural and contextual issues in the realm of learning a foreign language. In order to find ways to deal with the challenges they face, teachers need to have an ability to make the right judgments. They need to get familiarized with new techniques so that they feel the usefulness and significance of their abilities to teach. In this regard, it is hoped that the findings reported here might improve the field. There are still other questions related to beliefs and practices not fully answered here including: What is the influence of pairing novice and experienced, or experienced and expert raters on their CT and criteria? Participants at the present study did not have any time restriction during the process of data collection, which caused some ambiguity for the researcher at the data analysis. How are CT skills challenged by time restriction? What is the effect of group assessment instead of paired assessment? Though, still other questions could be raised.

Dialogue among scholars and practitioners about the beliefs and practices of teachers' CT in our profession is critical as we seek to improve our professional development programs and the quality of education services delivered to young children. A range of qualitative techniques such as interviews, retrospective recall or verbal protocol analysis provide empirically-based descriptions and evaluations of written products that help us understand rater behavior in more specific ways. These insights can be fed back directly into pedagogical measures such as rater training courses or rating scale and rubric development. Ultimately, though inconsistency in writing assessment cannot be fully eliminated, this investigation represents one more step towards minimizing deficiencies and proposing models to improve the quality of the raters' work.

6. References

Axelrod, B. N., Greve. K. W., & Goldman, R. S. (1994). Comparison of Four Wisconsin Card Sorting Test Scoring Guide with Novice Raters.*Journal of Psychological Assessment, 1*(2), 115-121.

Batanieh, F. R., & Zghoul, L. H., (2006). Jordanian TEFL graduate students' use of critical thinking skills (as measured by the cornell critical thinking test, level Z).*The International Journal of Bilingual Education and Bilingualism, 9*(1), 33-50.

Birjandi, P., & Bagherkazemi, M. (2010). The relationship between Iranian EFL teachers' critical thinking ability and their professional success. *English Language Teaching, 3*(2), 135-145.

Brown, M. N. & Keelley, S. M. (1994). *Asking the right questions.* Englewood Cliffs, NJ: Prentice Hall.

Burgess, M. L. (2009).Using web CT as a supplemental tool to enhance critical thinking and engagement among developmental reading students. *Journal of College Reading and Learning, 39*(2), 231-253.

Carroll, D. W. (2007). Patterns of student writing in a critical thinking course:

A quantitative analysis. *Assessing writing, 12,* 213-227*.*

Choy, S. C. &Oo, P. S. (2012). Reflective thinking and teaching practices: A precursor for incorporating critical thinking into the classroom? *International Journal of Instruction, 5*(1), 167–182.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Dantas–Whitney, M. (2002). Critical reflection in the second language classroom through audio taped journals. *System, 30(4),* 543–555.

Deremer, M. L. (1998). Writing assessment: raters' elaboration of the rating task. *Assessing writing, 5*(1), 7-29.

Dinkelman, T. (1999). An inquiry into the development of critical reflection in secondary student teachers. *Teaching and Teacher Education, 16(3)*, 195-222.

Facione, P. A. (2011).*Critical Thinking: What It Is and Why It Counts.* Millbrae, CA: Measured Reasons and The California Academic Press.

Facione, P. A. (2000).The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill*. Informal Logic, 20*(1), 61-84.

Farhady, H., Ja'farpur, A., & Birjandi, P. (1994). *Testing Language Skills: From Theory to Practice.* Tehran: SAMT Publications.

Ghafar Samar, R. & Ahmadi, M. (2012). *Collaborative writing assessment as a tool to enhance EFL teachers' critical thinking: Insights into think-aloud protocols*. Paper presented at the 9th Malaysia International Conference on ELT, Ipoh, Malaysia.

Ghahremani-Ghajar, S., & Mirhosseini, S. A. (2005). English class or speaking about everything class? Dialogue journal writing as a critical EFL literacy practice in an Iranian high school. *Language, Culture and Curriculum, 18*(3), 286-299.

Hattie, J. (2003).*Distinguishing Expert Teachers from Novice and Experienced Teachers: What is the research evidence?* Australian Council for Educational Research.University of Auckland.

King, K.O. (2010). *Using young adult literature and literary theory to teach middle school students: How to read through critical lenses.* Unpublished MS Thesis, Dominican University of California, San Rafael, CA.

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304.

Lieberman, A., & Miller, L. (2008).*Teachers in professional learning communities: Improving teaching and learning.* New York: Teachers College Press.

Lishchinsky, O. Sh. (2010). Teachers' critical incidents: Ethical dilemmas in teaching practice. *Teaching and Teacher Education, 27*, 648-656.

Lumley, T. (2005). Assessing second language writing: the rater's perspective*. Language Testing and Evaluation Series, 3*, 294-314.

Lynch, B. (2001). Rethinking assessment from a critical perspective. *Language Testing,18*(4) 351–372.

McNamara, T. F. (1996). *Measuring second language performance*. London: Longman.

Mei, W. S. (2007). Investigating Raters' Use of Analytic Descriptors in Assessing Writing. *Reflections on English Language Teaching, 9*(2), 69–104.

Paul, R.W. (1990). *Critical thinking: what every person need s to survive in a rapidly changing world.* Foundation for Critical Thinking, Santa Rosa, CA.

Philips, A. (2010). Teaching critical appraisal to students in the behavioral and Life sciences. *Psychology Teaching Review, 16*(2), 641-657.

Poehner, M. (2008). Dynamic assessment and the problem of validity in the classroom. *Calper working paper series,10*, 543-555.

Ramasamy, S. (2011). An analysis of informal reasoning fallacy and critical thinking dispositions among Malaysian undergraduates. *Language Testing Journal, 11*(2), 109-131.

Somervell, H. (1993). Issues in assessment, enterprise and higher education: The case for self-, peer and collaborative assessment. *Assessment and Evaluation in Higher Education, 18*(3), 221–233.

Soeherman, S. (2010).*The relationships of critical thinking skills, critical thinking dispositions, and college theological students in Indonesia.* Biola University, Indonesia.

Stapleton, P. (2001). Assessing critical thinking in the writing of Japanese university students: Insights about assumptions and content familiarity. *Writing Communication, 18*(4), 506–548.

Stout, C. J. (1993). The dialogue journal: A forum for critical consideration. *Studies in Art Education, 35*(1), 34-44.

Tsang, A. K. (2011). Online reflective group discussion – connecting first year undergraduate students with their third year peers. *Journal of the Scholarship of Teaching and Learning*, *11*(3), 58 – 74.

Twardy, R., C. (2005). Argument Maps Improve Critical Thinking. *Revised draft for publication in Teaching Philosophy*.

Wright, I. (2002). *Is that right? Critical thinking and the social world of the young learner*. Scarborough: Pippin Publishing Corporation.

Received: 14/01/2016
Accepted: 24/06/2016