Tabaran Institute of Higher Education
Iranian Journal of Language Testing
Vol. ٤, No. ٢, October ٢٠١٤

ISSN ٢٢٥١-٧٣٢٤

Iranian EFL Writing Assessment: The Agency of Rater or Rating Scale?

Nasim Ghanbari[١], *Hossein Barati*[٢]

Abstract

The present study explores the practice of Iranian raters in the EFL writing assessment context of the country. For this aim, early in the study a questionnaire developed by the researcher was administered to thirty experienced raters from ten major state universities in Iran. Later, five of the raters were chosen to participate in a follow-up think aloud session aimed to further explore the rating process. Findings of the questionnaire casted doubt on the existence of an explicit rating scale in the context. The following think aloud protocols further revealed that despite the apparent superiority of the raters in the rating process, neither the raters nor the rating scale-as two central components of the performance assessment- had a real agency in the Iranian EFL rating context. Lack of a common rating scale caused the raters to draw on ad hoc rating criteria and raters' idiosyncratic practices resulted by a missing rater training component created a context in which the construct of writing ability is greatly underrepresented. Along with locating the sources of the issue in both the rating scale and the raters, this study emphasizes that Iranian raters are in urgent need of training and constant monitoring programs to acquire rating expertise over time. As a requirement for the training programs, development of an explicit rating scale is strongly encouraged in the context.

Keywords: *EFL writing assessment, Rating scale, Rater, Validation, Think aloud protocols*

١. Introduction

Along with the influencing factors of task, and test-taker characteristics in performance assessment, the interactive component of rater and rating scale has received a considerable attention in different models of performance assessment (Kenyon, ١٩٩٢; Fulcher, ٢٠٠٣; McNamara, ١٩٩٦; Skehan, ١٩٩٨; Weir, ٢٠٠٥). In all these models, the rating obtained is in the first place the direct result of an interaction between the rater and the rating scale. In other words, the raters who judge the performance using some benchmark rating criteria constitute an important part of any assessment of performance. Although the factors involved in performance assessment have further refined in later elaborations of the performance models (e.g. Fulcher, ٢٠٠٣), the original importance attached with the rater and rating scale is still maintained. Rater and rating scales are no longer single entities that are combined together in

[١] *English Language and Literature Department, Faculty of Humanities, Persian Gulf University, Bushehr, Iran. Email: btghanbari@gmail.com ; btghanbari@pgu.ac.ir*

[٢] *English Language and Literature Department, Faculty of Foreign Languages, University of Isfahan, Isfahan, Iran. Email: h.barati@gmail.com*

ISSN ٢٢٥١-٧٣٢٤

a straightforward manner producing a measure of performance called rating. Rather, each has a multi-layered structure that configures the performance assessment in a myriad of ways.

In this regard, rater variability and the construct of rating scale that no longer sustains the conception of a neutral ruler in the rating context combine to create an unpredictable rater-rating scale interaction that might lead to the underestimation or overestimation of either's role in the performance assessment of writing. Following this line of inquiry, a huge body of research studies has made efforts to improve the functionality of either rater (Eckes, ٢٠٠٨; Johnson & Lim, ٢٠٠٩; Lumely, ٢٠٠٢; Schaefer, ٢٠٠٨; Shaw, ٢٠٠٢) or rating scale (Barkaoui, ٢٠٠٧; Knoch, ٢٠٠٧; North, ١٩٩٥; Upshur & Turner, ١٩٩٥) in rating situations. The attempt here has been to control the raters' undesirable factors of bias, variable rating experience and diverse degrees of professional training on the one hand. In addition, development of rating rubrics that are empirical, context-based, explicit, detailed and in line with the developmental route of second language acquisition is encouraged on the other hand.

Nevertheless, it happens that the nature of the interaction between the rater and the rating scale is ambiguous in some rating contexts. For example, in impressionistic rating contexts where there is not a serious concern for a rigorous assessment of writing, rating is looked upon as an individual rater's task (Barkaoui, ٢٠٠٧). As a result, in the absence of assessment quality criteria of reliability and validity, rating is viewed as the personal task of the rater and the way he/she conceptualizes the construct determines the final outcome of the assessment. It is clear that in these contexts the Pandora's Box of rating incorporates a mysterious combination in which the functionality of each rating component is unknown.

EFL writing assessment as a non-native rating context has been criticized for the prevalence of an impressionistic rating mood (Barkaoui, ٢٠١٠) in which raters rely on their intuitions (Brown, ١٩٩٥) and a vague combination of norm-referenced and criterion-referenced rating elements reign the context. In an attempt to explore the prevailing rating situation in EFL contexts, the present study investigates the particular Iranian EFL writing assessment context. Focusing on two components of rater and rating scale, the deficiencies associated with each are counted. This study particularly investigates the function of the rater and the rating scale in the rating process.

For this purpose, early in the study, the raters were surveyed for how they perceived the task of rating with regard to the rating scale. These preliminary findings which demonstrated the way raters conceived the role of the rating scale were followed by think aloud sessions to get access to a realistic profile of the rating process. Therefore, the principal aim of the present study was to provide answers to the following two research questions:

- Does rating scale have any place in the Iranian EFL writing assessment context?
- How do the rater and rating scale function in the Iranian EFL writing assessment?

٢. Literature Review

٢.١. Writing Assessment

In an attempt to provide a more valid picture of the construct of writing ability, there has been a major shift in language testing towards the development and use of performance tests within the past decades. Within this new mode of assessment student writing is assessed by raters using some kind of rating scale which makes it different from the traditional fixed-response assessment. This type of assessment provides the advantage of directly measuring candidates' productive language skills.

ISSN ٢٢٥١-٧٣٢٤

In this regard, different models have specified the factors that influence performance (e.g. Fulcher, ٢٠٠٣; Kenyon, ١٩٩٢; McNamara, ١٩٩٦; Skehan, ١٩٩٨, Weir, ٢٠٠٥). Among different models that have dealt with performance, two models of Fulcher (٢٠٠٣) and Weir (٢٠٠٥) are briefly mentioned. In review of each model, the place of rater and rating scale is considered. In the context of oral performance assessment, Fulcher (٢٠٠٣) provides a detailed model explaining how different factors might affect performance outcomes. As Figure ١ shows along with task and test-taker components, rater and rating scale influence the performance.

In this model, rater characteristics and rater training affect the rater performance. Moreover, scoring philosophy and construct definition of the rating scale impact the outcome of the rating process. As the model shows, the interactive component of rater-rating scale plays an important role in the performance assessment. In fact, when the two factors are in place and function properly, the validity of the assessment outcomes would be enhanced.

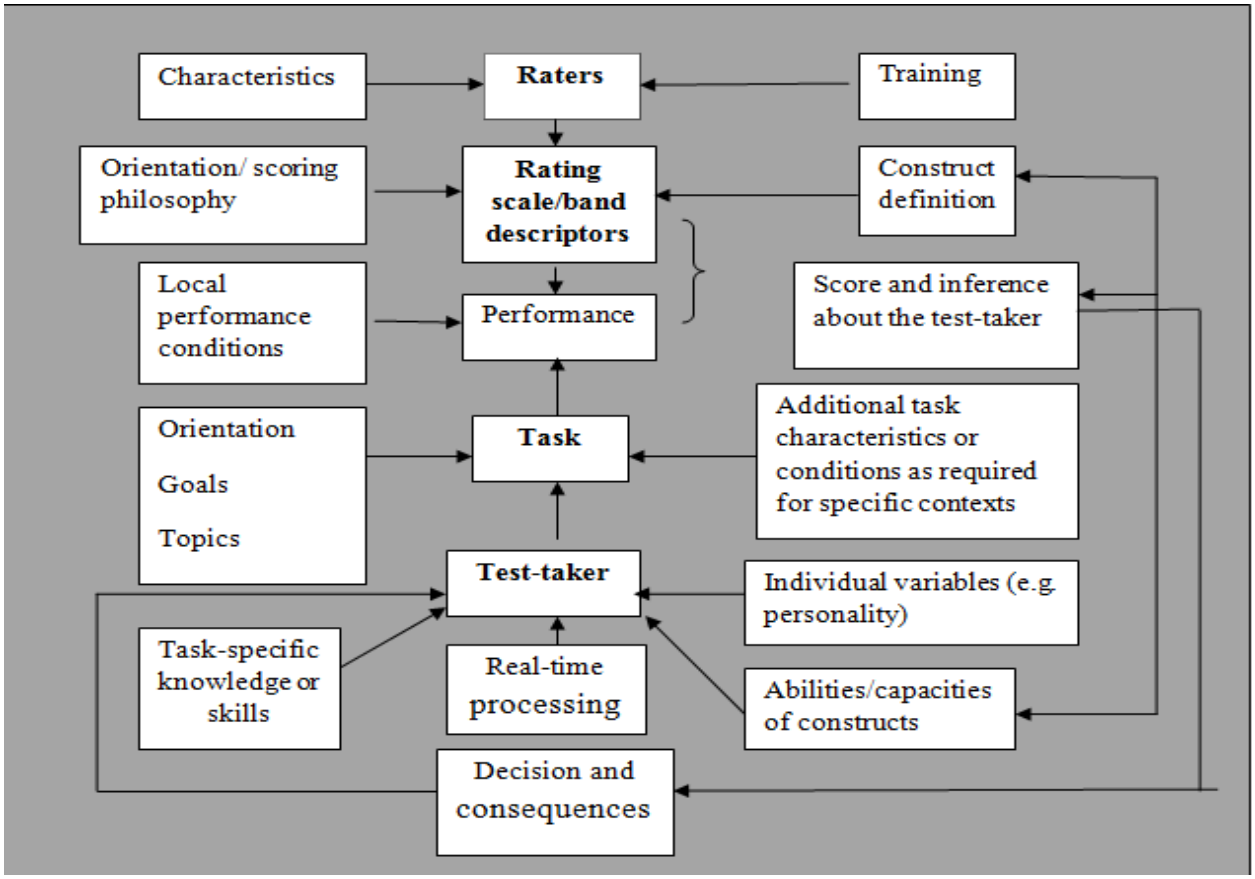
Later in ٢٠٠٥, Weir proposed his socio-cognitive model for the validation of performance assessment (Figure ٢). In his opinion, there are three important validity components at the heart of any language performance assessment. They are cognitive validity, context validity and scoring validity. As Figure ٢ shows, achieving scoring validity is highly important since if we cannot rely on the ratings done, it has unfortunate consequences for the validity of the developed tasks (Shaw & Weir, ٢٠٠٧).

The first scoring validity parameter is that of the criteria and type of rating scale. In fact, the choice of appropriate rating criteria and the consistent application of them by trained raters are regarded as key factors in the valid assessment of second language performance (Alderson, Clapham, & Wall ١٩٩٥, Bachman & Palmer ١٩٩٦, McNamara ١٩٩٦). In addition to rating scale, rater-related factors also influence the validity of the performance assessment. Therefore, rating scale and rater determine the scoring validity and in general they affect the whole of the assessment enterprise.

As the above two models showed, rater and rating scale immensely affect the assessment of performance. Hence, due to their importance, they are further discussed below.

Figure ١. Expanded model of oral test performance (Fulcher, ٢٠٠٣)

ISSN ٢٢٥١-٧٣٢٤



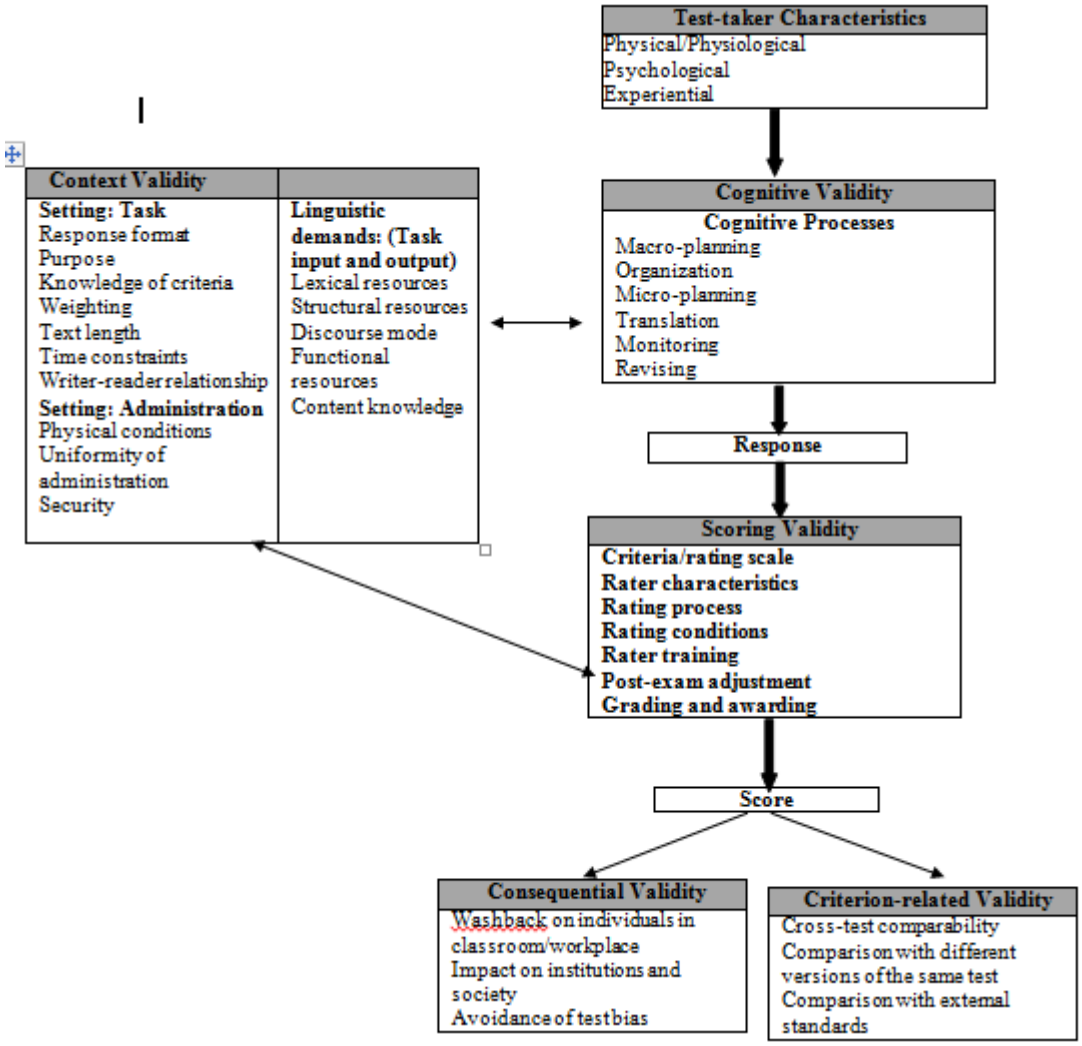
٢.٢. Rater

In the complex world of language assessment, the presence of raters distinguishes performance assessment from traditional assessment. In contrast to traditional assessment where scores are elicited solely from the interaction between test-taker and the task, in performance assessment variables inherent to the rater significantly affect the outcome (McNamara, ١٩٩٦). Therefore, along with the increasing prevalence of performance assessment of writing in both large-scale and classroom assessment, considerable attention has been paid to the raters and what they do when involved in the rating task.

This line of inquiry has advanced in two directions. The first deals with the issue of rater variability. In fact, rater variability is manifested in a variety of ways (Weigle, ٢٠٠٢). A number of studies have shown the ways in which raters can vary.

Figure ٢. Framework for conceptualizing writing performance (Weir, ٢٠٠٥)

ISSN ٢٢٥١-٧٣٢٤



Some of these rater effects are severity effects, halo effect, central consistency effect and bias effects. Severity effect implies that raters may differ in terms of their overall severity compared to the other raters. Halo effect occurs when the raters cannot distinguish between different conceptually distinct traits and they instead rate based on their general impression. The third rater variable is central tendency in which raters avoid the extreme ratings and have a tendency for the midpoint of the scale (Landy & Farr, ١٩٨٣). Finally, biased raters tend to rate unusually harshly or leniently on one aspect of the rating situation (Schaefer, ٢٠٠٨). Rater background including rater's native language, experience, and training is another source of variability among the raters which have been investigated by a good number of scholars (Barkaoui, ٢٠١٠; Eckes, ٢٠٠٨; Lim, ٢٠٠٩, Lumley, ٢٠٠٢).

The next line of research looks inside the rating process. It is motivated on the grounds that understanding rating process reveals how different rater-related factors and rating context variables interact to shape the final rating outcome (Lumley, ٢٠٠٢, ٢٠٠٥). Investigations into the rating define rating as an indeterminate process (Lumley, ٢٠٠٢) in which raters had different interpretations of the rating task when they evaluated the same texts (Deremer, ١٩٩٨). These studies which mostly use introspective measures such as think aloud protocols have shown that rating is an unpredictable task which put human rater in interaction with a multitude of rating factors in the assessment context (Huot, ١٩٩٣; Pula & Huot, ١٩٩٣, Vaughan, ١٩٩١).

ISSN ٢٢٥١-٧٣٢٤

The above-mentioned studies underscore the importance of raters in performance assessment. In fact, in the words of Lumely (٢٠٠٢), rater lies at the center of the process. By improving the inherent subjectivity associated with raters the validity of the assessment considerably improves and the raters' stance in assessment is improved as well.

٢.٣. Rating scale

A historical overview of the measurement theories involved in writing assessment in the ٢٠th century reveals two dominant traditions, i.e. test-score tradition and the scaling tradition (Behizadeh & Engelhard, ٢٠١١). The test-score tradition that originates from the seminal work of Spearman (١٩٠٤) was primarily concerned with measurement error and the decomposition of an observed score into two components of true score and some error components. This tradition which continued under the name of classical test theory led to the emergence of some more powerful and sophisticated theories such as generalizability theory (Brennan, ١٩٩٧; Cronbach, Gleser, Nanda, & Rajaratnam, ١٩٧٢), factor analysis and structural equation models (Joreskog, ٢٠٠٧).

Scaling theory as the other influential measurement tradition originally rooted in the psychophysics in ١٩th century. The focus of scaling tradition was to provide some variable maps that configure the location of both items and individuals onto a latent variable scale that represented a construct. Inspired by E.L.Thorndike's epigram that "Whatever exists at all exists in some amount" (Clifford, ١٩٨٤), scaling tradition has continued through presenting different kinds of rating scales which have its theoretical basis in the item-response theory.

In writing assessment, rating scale as the operational definition of the construct provides a series of constructed levels that students' performance is judged against. As Figure ١ shows rating scales are not mere tools of assessment but they are also realizations of theoretical constructs which carry a particular rating philosophy. Furthermore, a primary purpose of the rating scale is to constrain different interpretations of the raters to help them to articulate and justify their rating decisions to come to reliable, orderly and categorized ratings (Lumley, ٢٠٠٢).

Many scholars have studied different aspects of the rating scale. For example, some scholars have attended to the development and validation of scales (Knoch, ٢٠٠٧; Hawkey & Barker, ٢٠٠٤; Nakatsuhara, ٢٠٠٧; Tyndall & Kenyon, ١٩٩٦), different types of rating scales and their influence on rating (Barkaoui, ٢٠٠٧, ٢٠١٠; East, ٢٠٠٩), rater's elaboration of the rating scale categories (Deremer, ١٩٩٨; Lumley, ٢٠٠٢, ٢٠٠٥). All these studies aimed to improve the function of rating scales which in turn affected the outcome of assessment.

In sum, the above brief review illustrates the important components of rater and rating scale in writing assessment. Although it would be immature to consider a linear and direct effect for each on the ratings assigned, the two components interact with each other and the rating context in complex and multi-faceted ways to come to a final judgment. Nevertheless, this complex process presupposes the presence of both the rater and the rating scale in the assessment process. It was the search for the existence and functionality of these two rating components that motivated the present study. The following section explains the methodology pursued in the present study.

٣. Methodology

٣.١. Context of the study

The context of the present study was the Iranian EFL undergraduate academic writing assessment. Here, writing courses are treated as usual part of the undergraduate program.

ISSN ٢٢٥١-٧٣٢٤

Two courses of paragraph writing and essay writing form the core practice in this regard. The focus of paragraph writing is to enable learners to learn the essentials of composing simple paragraphs in English. This course, moreover, is considered as the requisite for the essay writing course that prepares learners to write essays in English at a more advanced level. Upon completing the courses, learners are expected to have developed sufficient mastery for their further academic coursework. For each of the courses, teachers are responsible for the rating task as well.

٣.٢. Participants

A group of thirty raters who held PhDs in TEFL were randomly selected from ten major state universities in Iran to participate in the study. They varied in terms of age, gender and TEFL teaching background. Regarding the gender distribution, ٢٤ (٨٠ %) of the participants were male and the remaining ٦ (٢٠ %) were females. Few of these people (٣ out of ٣٠ raters) had attended any rater training courses, but all had a minimum five years of experience in teaching and assessing writing. Table ١ below provides a brief summary of the profile of the participants in this study.

Table ١. Descriptive statistics of the raters in the study

	N	*Minimum*	*Maximum*	*Mean*	*SD*
Teaching experience	٣٠	٥	٣٠	١٦.٦٧	٥.٨٦٨
Rating experience	٣٠	٥	٢٣	١٠.٤٧	٥.٥٨٨
Age	٣٠	٣٠	٥٣	٤٤.٤٣	٦.٦٧٨
Valid N (listwise)	٣٠				

Along with the raters, a group of hundred students were also randomly selected to write essays in this study. Table ٢ displays the background characteristics of the student writers.

Table ٢. Background information of the student writers in the study

	Gender	*N*	*Major*	*N*
	Male	٢٨	Literature	٦٧
	Female	٧٢	Translation	٣٣

٣.٣. Instruments

٣.٣.١. Researcher-made questionnaire

The first draft of the questionnaire included sixty items which asked the raters to respond to a five-point Likert-like scale ranging from strongly disagree to strongly agree. Since this study was part of a larger study that aimed to investigate different aspects of rating in the Iranian EFL writing assessment context, therefore, in addition to a group of items that directly

ISSN ٢٢٥١-٧٣٢٤

addressed the issue of rating, items which explored other related issues were also inserted in the questionnaire.

Upon the development of the early draft of the questionnaire and in order to improve the quality of the instrument, it was decided to do a pilot study. For this aim, five raters, with similar characteristics as those of the main study, were asked to take part in piloting session for the questionnaire. After the completion of the questionnaire, the raters were asked to comment over any problems they identified with the instrument. In so doing, they were encouraged to express any further ideas that might have been ignored by the researcher, check for the relevance of the items to the major themes of the questionnaire and underline any wording or conceptual problems existing in the items. The piloting phase of the questionnaire resulted in comments that suggested the questionnaire was too long for the busy university raters to respond; also, the wording of some items was ambiguous and some of them did not specifically address rating as the main concern of the questionnaire. As a result, a group of ٢٠ items were omitted from the questionnaire. Some of the items were omitted for wording or conceptual ambiguity. Further, some others which the raters did not agree on their related subscales were considered as misleading and consequently were omitted.

Following piloting the early draft of the questionnaire, the final form of the questionnaire which included forty items was prepared. As mentioned, the questionnaire included different subscales (Table ٣). For the particular purpose of the current study, the sixteen items in the subscale existence/ application of rating scales was used (see Appendix).

Table ٣.The structure of the questionnaire developed in the study

Subscale	*No. of items*	*Description*
١. Existence/application of rating scale	١٦	Investigates whether there exists any rating scale in use
٢. Context in writing assessment	٦	Investigates the effect of contextual factors in EFL writing assessment
٣. International rating scales	٦	Investigates the appropriacy of International rating scales in EFL writing assessment
٤. Others	١٢	Investigates other aspects of EFL writing assessment

ISSN ٢٢٥١-٧٣٢٤

As for the reliability of the questionnaire, a group of forty raters were asked to complete the questionnaire. The measure of Cronbach Alpha was estimated for the whole questionnaire as well as the desired sub-scale of the questionnaire. The measure of Alpha obtained for the whole questionnaire was ٠.٦٠ while it turned out to be ٠.٦٨ for the desired subscale in the questionnaire. The reason for the low reliability index of the whole questionnaire can be sought in the very structure of the questionnaire. The questionnaire acted as an exploratory device in this study. In other words, different subscales in the questionnaire included items which were not conceptually close to each other. Consequently, the items did not show high internal consistency with those in other subscales. And since Alpha is sensitive to internal consistency of the items, it considerably lowered the estimates of reliability for the whole questionnaire. Needless to say, items in the desired subscale were more consistent with each other which caused the higher-though low- reliability index in this subscale.

With regard to the validity of the questionnaire, six raters as experts were asked to find the correspondence between the items in the questionnaire and the underlying constructs. The inter-coder correlation estimated through a particular Kappa procedure (Fleiss, ١٩٧١) turned out to be ٠.٨١ indicating a high degree of agreement among the raters in classifying the items into their related subscales. The general consensus among the raters over the underlying constructs of the questionnaire was indicative of the validity of the questionnaire.

٣.٣.٢. Writing texts

Under real-exam condition one hundred writing texts were collected from the participants. Student writers were required to write on the paper sheets prepared by the researcher and to provide information such as gender, major and place of study. Due to more familiarity of the students with the argumentative genre in academic writing tasks, student writers were assigned an argumentative writing topic.

٣.٤. Data collection procedure

Since this study was part of a larger research project which aimed to develop and validate a local rating scale in the Iranian context, different number of texts was required in each phase of the study. So, to meet this requirement, early in the study, a body of one hundred undergraduate students, randomly selected from four undergraduate TEFL classes at the English Department, University of Isfahan was asked to write a well-organized argumentative essay within a ٤٥-minitue time frame. The collection of texts was conducted in early fall of ٢٠١١.

Next, to administer the questionnaire the researcher randomly selected three raters from each of the ten major Iranian state universities which were randomly considered in the study (e.g. University of Tehran, Shiraz University, Ferdowsi University, Tarbiat Modares University, etc.). Then, the questionnaire was sent to the raters who were in different universities across the country. The purpose of the questionnaire was to elicit the raters' attitudes towards different aspects of rating and rating scale in the country. The raters could either complete the paper version of the questionnaire or an electronic version via email which they could fill in on the computer and send back to the researcher. The collection of the completed questionnaires lasted about three months.

Following the administration of the questionnaires and in order to closely investigate the rating process, a group of five raters, who had responded to the questionnaire, were asked to take part in a think-aloud session. Therefore, two months after completing the questionnaires, these five raters agreed to verbalize their thoughts while each rating a sample

ISSN ٢٢٥١-٧٣٢٤

of three writing texts. Intentionally, the time interval of two months was considered in order to prevent the probable memory effect which could interfere with their true performance in the think-aloud sessions. Fifteen writing texts were randomly selected from a pool of hundred texts and were given an ID number from one to fifteen. The texts were then put in groups of three and they were randomly assigned to the five raters.

At the beginning of the think aloud session, the researcher briefly explained the procedure of the think aloud to the raters. The raters as experienced academics in TEFL were quite aware of the procedure and hence it facilitated the administration of the think-aloud session. According to Cohen (١٩٩٤) to ascertain the reliability and accuracy of the recoded information, the raters were allowed to use whichever language they felt easy in their think-aloud process, whether English or Persian. The administration of the questionnaires and the collection of the subsequent think aloud protocols (TAPs) were conducted in late fall ٢٠١١ and early winter ٢٠١٢ (Figure ٣).

Figure ٣. Summary of data collection procedures

Phase I

١.Development of the questionnaire

٢. Piloting the questionnaire

٣. Administration of the questionnaire

Phase II

٤. Orientation session for think-aloud procedure

٥. Rating ١٥ essays while thinking aloud

٣.٥. Data analysis

The analysis of the questionnaire was concerned with the desired subscale which investigated the existence and/or application of rating scale in writing assessment in the country (Table ٣). The sixteen items constituting this subscale were carefully analyzed following the categories they were grouped under. Later, the researcher provided frequency counts and percentages for the items constituting each of the categories.

TAPs were carefully transcribed by the researcher. To start the coding process, initially the researcher considered the patterns emerged from the analysis of the questionnaire as a framework in mind. Additionally, some other categories emerged during the coding process. Upon multiple rounds of analyses, categories developed in this way were put together to form the main themes of the TAPs. In order to establish the reliability of the coding conducted by the researcher, a colleague of the researcher who was a rater-participant

ISSN ٢٢٥١-٧٣٢٤

in the study coded a portion of five verbal protocols. A substantial inter-coder agreement was found between the researcher and the rater-colleague (Kappa=٠.٨٧, p <٠.٠٠١).

٤. Results

In this part, obtained results are presented based on the research questions posed earlier in the study.

٤.١. Research question I: Does rating scale have any place in the Iranian EFL writing assessment context?

Through analyzing the sixteen items which explored the place of the rating scale among the Iranian EFL raters, the researcher could cluster the items around three major categories (Table ٤).The first group of items looked into the process of rating and investigated the respondents' opinions about rating in the Iranian context. The second group of responses located the place of rating scale in the experiential practice of the raters, and finally the last category presented the raters' views regarding the place and function of rating scales in writing assessment in Iran. It's worth mentioning that presentation of the raters' responses in Tables ٤ to ٧ -though measured on a five-point Likert scale- is in a way that the two ends of the scale are merged with their neighboring points. The justification is that this study was the first phase of a larger research project and the degrees of agreement or disagreement did not affect the interpretation of the results obtained in this phase, so the two ends of the questionnaire were merged with their neighboring categories when reporting the results.

Table ٤. Analysis of the questionnaire items in the desired subscale

Category ١: Rating process	Items: ١١, ١٤, ١٥, ١٧, ١٩, ٢٣, ٢٤, ٣٦
Category ٢: Views on rating scales	Items: ٢١, ٢٢, ٢٨, ٣٧, ٣٩
Category ٣: Experience in rating	Items: ١٣, ١٨, ٤٠

As Table ٥ below shows, the raters who contributed to this study doubted the existence or use of a common rating scale in their rating. While a substantial number of raters disagreed with an impressionistic approach to rating (٥٦.٦٦٪, Item ١٥) and strongly believed that all raters had some criteria in their scoring (٨٠ %, Item ١٧), they held differing attitudes about rating scale in their own practice. For example, when asked whether they solely relied on their own scoring or used any known scale at all, the pattern of the raters' responses showed that although they had their own analytic scoring (Item ١١) which included the explicit criteria of word choice, structure, spelling, etc. (Item ٢٣) they were hesitant to reject the efficiency of the existing rating scales (Item ٢٤). In other words, their responses to items ١٩ and ٣٦ revealed that they were aware of the existing scales but they followed their own approach to rating which was not directly determined by the existing rating scales. It seems that the prestige attached with the rating scales in writing assessment created a foggy atmosphere in which although expert Iranian raters followed their own rating criteria, they did not like to strongly reject the efficiency of these scales. The substantial weight associated with the 'no idea' category in Item ٢٤ is a testimony in this regard.

Moreover, there was a paradox in the raters' responses. On the one hand, the validity of the impressionistic rating in which no explicit scale is considered was strongly questioned (Item ١٥). On the other hand, in both their holistic and analytic scoring (Items ١١ & ١٤), they contended that they were not concerned with using extant rating scales and believed instead

ISSN ٢٢٥١-٧٣٢٤

that their own rating criteria fulfilled their purpose quite well. It appeared that the raters viewed impressionistic scoring different from the use of their idiosyncratic rating criteria. In other words, in raters' idea, drawing on their own criteria was a kind of using scale in their rating, while they stated that they did not attend to the existing rating scales.

Table ٥. Analysis of items ١١, ١٤, ١٥, ١٧, ١٩, ٢٣, ٢٤ and ٣٦ in the questionnaire

Items	*Agree (%)*	*Disagree (%)*	*No Idea (%)*
Item١١: I have my own way of analytic scoring (i.e. I do not use any existing analytic rating scale).	٤٦.٦٦	٤٠	١٣.٣٣
Item ١٤: I look at the text and give a single general score based on a rating scale (holistic scoring).	٤٦.٦٦	٤٣.٣٣	١٠
Item ١٥: I think giving a score based on my impressions is quite trustable	٣٣.٣٣	٥٦.٦٦	١٠
Item ٢٣: I have my own scoring criteria such as word choice, structure, spelling	٥٠	٤٠	١٠
… but I don't consider them as a kind of analytic scoring.			
Item ٢٤: When it comes to my actual rating, I find existing rating scales less effective	٣٣.٣٣	٣٦.٦٦	٣٠
(i.e. there are inconsistencies between my criteria and those of the scales).			
Item ١٧: I think all raters have some criteria for their scoring though	٨٠	١٠	١٠
they may not be in the familiar format of present scales (analytic or holistic).			
Item ١٩: International rating scales such as Jacobs, et al. (١٩٨١) directly guide my rating.	٣٦.٦٦	٤٦.٦٦	١٦.٦٦
Item ٣٦: As a rater, I am quite aware of different rating scales.	٦٣.٣٣	٣٠	٦.٦٦

Five items (٢١, ٢٢, ٢٨, ٣٧ and ٣٩) comprised the second category. These items investigated the rater's general views on rating scales. As Table ٦ shows, majority of the raters (about ٦٧ %) believed that rating scales had an important function in writing assessment (Item ٢١). They strongly believed that the application of a rating scale considerably would improve the psychometric qualities of such an assessment (Item ٢٢, ٢٨ and ٣٧). As further evidence to item ٢٢, the raters' responses to items ٢٨ and ٣٧ showed that they were concerned with issues such as raters' consistency and bias. In addition, there was a considerable consensus among the raters that the existence of a common rating scale would enhance the fairness of writing assessment (Item ٣٩).

Table ٦. Analysis of items ٢١, ٢٢, ٢٨, ٣٧ and ٣٩ in the questionnaire

Items	*Agree (%)*	*Disagree (%)*	*No Idea (%)*
Item ٢١: Rating scale plays a significant role in any assessment of writing.	٦٦.٦٦	٢٠	١٦.٦٦

ISSN ٢٢٥١-٧٣٢٤

Item ٢٢: An explicit rating scale improves validity and reliability of my assessment.
٨٣.٣٣ ٦.٦٦ ١٠
Item ٢٨: Rating a composition is quite an individual act
٦.٦٦ ٧٦.٦٦ ١٦.٦٦
(e.g. no concern for inter-rater agreement).
Item ٣٧: Lack of a common rating scale would lead to bias, inconsistency and
٨٣.٣٣ ٣.٣٣ ١٣.٣٣
leniency/severity among the raters.
Item ٣٩: The existence of a common rating scale would lead to a fair writing assessment.
٨٦.٦٦ ٦.٦٦ ٦.٦٦

Finally, the last category revealed the importance of experience in shaping the particular rating approach of the raters. In other words, the raters relied on their scoring criteria since they had a strong experiential basis. To the extent that this body of experiential knowledge even justified and assured them of their impressionistic scoring (Items ١٣ & ٤٠). It may seem that raters' tendency to impressionistic scoring be at odds with their devotion to a set of criteria for their scoring indicated above. Probably, high agreement on item ١٨ specifies that through professional involvement in rating, the raters had developed a set of criteria that guided their rating. Table ٧ below presents the results of analysis for the three items of ١٣, ١٨ and ٤٠.

Table ٧. Analysis of items ١٣, ١٨ and ٤٠ in the questionnaire

Items	*Agree (%)*	*Disagree (%)*	*No Idea (%)*
Item ١٣: I look at the text and based on my own experience in rating, I give a total score.	٦٠	٣٦.٦٦	٣.٣٣
Item ١٨: Upon experience, I have learned to keep all the rating criteria in my mind and score based on them.	٧٠	٢٣.٣٣	٦.٦٦
Item ٤٠: My rating experience justifies the scores I assign out of my general impressions of the text.	٦٦.٦٦	٢٣.٣٣	١٠

In sum, analysis of the raters' responses in the three emerged categories showed that rating scale in its conventional sense does not have any place in the EFL writing assessment context of Iran. While the raters confirmed the importance of the rating scale in writing assessment, they claimed that they rarely adopted any existing scale in their practice. Instead, they relied on their own rating criteria which were not in the form of a scale and enjoyed a strong experiential basis.

٤.٢. Research question II: How do the rater and rating scale function in the Iranian EFL writing assessment context?

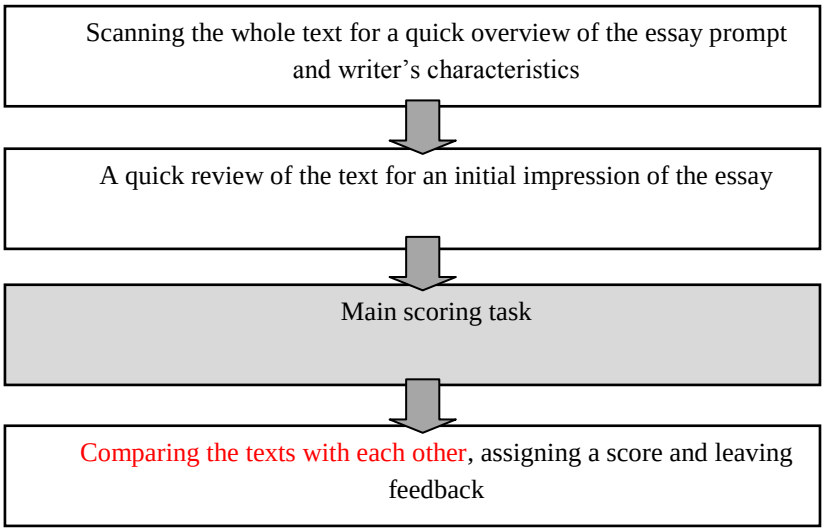
Here the researcher made attempts to have a closer look at the rating task. The aim here was to see how rating scale or criteria-as raters called it- and the rater functioned in the rating task. Results of analyzing fifteen TAPs showed that there were four distinct stages in the

ISSN ٢٢٥١-٧٣٢٤

rating process that almost all the five raters went through (Figure ٤). A brief look at the stages (e.g. I, II and IV) shows that it was the raters who had a more prominent role in the rating process. In fact, in the absence of an explicit rating scale, the raters were the final judges to articulate and justify a rating that could best meet the requirements of reliable and orderly ratings in their view.

Despite the general similarities of the raters in stages, I, II and III, they displayed lots of variations when it came to the main scoring task. In fact, lack of a rating scale as a tool to channel and constrain the raters' different reflections on the texts had created considerable variations and inconsistencies among them. This chaotic practice was well-revealed when the raters commenced on the main scoring stage. They showed diverse reactions when dealing with the texts. On the whole, following the analysis of TAPs at this stage (i.e. main scoring stage) two major themes emerged which are presented below.

Figure ٤. Stages of the rating process



٤.٢.١. No evidence for any existing rating scale in use

Despite the raters' vague attitude to the use and application of rating scales (see ٤.١), their actual rating performance in the think aloud sessions showed that they did not draw on any established rating scale. By the term rating scale, majority of the raters had their own set of scoring criteria in mind. In fact, none of the raters even referred to any well-known rating scale, rather they referred to their own criteria by saying as "my criteria are…." For example, at the end of his rating, Rater ٣ articulated his final judgment as:

> Rater ٣: Well! If I wanna rate this text on organization, punctuation and grammar uhmm based on my alphabetical scoring system A, B, C, I will assign C to her.

In his second rating, Rater ١ elaborated on his rating criteria. It was clear that he did not use or even refer to any known scale in his rating:

> Rater ١: So again! Let's start from the beginning. Again I have some issues with the paragraphs, the way they have been divided, no space…no

ISSN ٢٢٥١-٧٣٢٤

indentation, uhhh, again I would say grammatically uhhh he has some flaws to call flaws but very bad handwriting ,content-wise that's I would say ٨٥ percent OK , formally speaking that's ٧٥ percent OK! so if I get, if this student gets I would say ١٥ and ١٨, the average would be ١٥ plus ١٨ divided by ٢… doesn't again have an eye catcher or an interesting topic, interesting sentence is needed in the beginning, this piece of writing lacks it, uhhh…again content wise, conclusion is again not clearly defined so, upon further contemplation this student deserves ١٥ plus ١٧ divided by two this will be some of them write carelessly uhhh punctuation is important, handwriting is important at least to me and for me…

٤.٢.٢. Raters' scoring criteria were justified to them

As mentioned, each of the raters was assigned three essays for their rating. The consistent application of the rating criteria when scoring each of the texts showed that upon experience, the raters had developed their own rating style. In other words, the way they approached the erroneous instances in the texts and the way they reasoned about them and finally justified their rating confirmed that the raters had a robust set of criteria for their scoring. For example, during their think aloud sessions it happened that the raters stopped and reflected on their criteria. Rater ٥, for example, tried to clarify his rating criteria when deciding on a particular ill-formed instance of the text:

Rater ٥: So I wanna rate let's say holistically and there are many criteria. First of all, there are many irrelevant ideas. Alright! Because we have the topic and the ideas and content should be related to the topic but there are many irrelevant sentences and ideas not focusing on the main topic…they do not contribute…let's say to the elaboration of the topic and communicatively proper ideas…so number one the content and the irrelevance of the content and many of the sentences…the second criteria concentrates on the organization of the…let's say the paragraph, alright?! So the paragraph should be based on reasons, here the topic asks why did you choose English as your academic major…so the logical relationship should be based on a couple of reasons concerning let's say the person choosing English as the academic major but I did not see the essay follows such line of reasoning throughout the paragraph…alright!...

As further evidence, the raters identified their rating approach with particular terms. For instance, Rater ٢ stated that he was form-oriented in his rating. To him, form concerned the grammatical aspects of the texts. His rating showed that he was extremely sensitive to the grammatical soundness of the language.

Rater ٢: not only I wanted…OK …not only did I want[rater underlines the correct form] makes more sense, not only did I want to…on a new world and a new culture but… but contrast…uhmm but also maybe better here… to enter innermost professional components of English and… uhhm maybe to for parallelism to work… for example translating tests of different topics and subjects or being an English teacher, being maybe…makes sense[rater puts a

ISSN ٢٢٥١-٧٣٢٤

tick under being] Ok!... more attention to form than to meaning so…being is OK.

In sum, findings of TAPs showed that the raters had a prominent role in the rating process. In fact, in the absence of a standard rating scale, the raters relied on their own rating criteria to manage the rating process. However, it created a chaotic rating practice. Different raters weighted different aspects of the essays. Even when they were concerned with the similar aspects (e.g. grammar, content, etc.) they variably weighted them in their general evaluation of the texts. On the whole, what could be inferred from TAPs was the apparent dominance of the raters in the rating process. It was the raters' subjectivities which unfolded in their rating criteria and they finally determined the scoring process.

٥. Discussion

The common thread among many studies that have investigated different aspects of rater and rating scale in writing assessment has been to improve the functionality of the two to yield the most reliable and valid assessments of the performance (Barkaoui, ٢٠٠٧, ٢٠١٠; East, ٢٠٠٩; Eckes, ٢٠٠٨; Gamaroff, ٢٠٠٠; Johnson & Lim, ٢٠٠٩). However, despite the paramount importance of these two factors in performance assessment of writing, findings of the present study showed that Iranian EFL writing assessment context seriously lags behind the current practices in the field. Results of the questionnaire survey with thirty experienced raters in the country showed that the raters relied on their own rating criteria and the existing rating scales had no place in guiding their rating decisions.

As further evidence, results of TAPs revealed that it was the raters that defined the rating route. Although this finding is in line with many studies that consider the rater at the center of the rating (Cumming, Kantor, & Powers, ٢٠٠١; Erdosy, ٢٠٠٤; Lumley, ٢٠٠٥), a closer look at the process of rating shows that the raters in this study did not fulfill their expected role in the rating process. According to previous research (e.g., Cumming, Kantor, & Powers, ٢٠٠٢; Lumley, ٢٠٠٥; Milanovic, Saville, & Shuhong, ١٩٩٦) decision-making behaviors of the raters and the aspects of writing they attend to while rating define the rating process. The present study showed that although Iranian raters followed some general steps in their rating (Figure ٤) when it came to the main scoring stage, they showed great variations in the aspects of the texts they attended to and the weights they assigned to them. To explicate the issue, two interrelated issues of rater expertise and the place of rating scale are further explained.

As Lumley (٢٠٠٥) rigorously laid out the process of performance writing assessment enterprise, raters occupy a middle position between two ends. On one end are students writing samples which show disordered complexity (Lim, ٢٠١١) and on the other hand are the institutional requirement of providing standardized and reliable test scores as a basis for making decisions. Such a consequential role of the raters creates an unprecedented need to provide rating competence among the raters. In fact, raters have to provide appropriate ratings in a consistent way. It is on these grounds that test programs administer rater training which usually incorporates familiarization activities, practice rating and feedback and discussion (Lane & Stone, ٢٠٠٦). Regardless of the quality of these rater training programs, numerous research suggest that rater training is effectual even at least in a short interval after the training sessions (Congdon & McQueen, ٢٠٠٠; Engelhard, ١٩٩٤; Shohamy, Gordon & Kraemer, ١٩٩٢; Weigle, ١٩٩٨). Majority of the raters in this study stated that they had not

ISSN ٢٢٥١-٧٣٢٤

experienced any rater training courses. Some even commented that their training was a self-made effort and it was their long years of teaching and rating writing experience that had shaped their rating philosophy.

In fact, training and constant monitoring of the raters as important instruments for checking the quality of the ratings are quite missing in the Iranian context. In this situation, individual raters' experiences which are shaped by many background factors over time determine the way they approach the rating task. Consequently, different conceptions of what a good piece of writing is prevail in the context. In this situation, ratings resulted from the raters' experiences which as mentioned are quite personal and the probable similarities with other raters are mere chance. In the words of Lim (٢٠١١), rater experience and rater expertise are two distinct concepts. While experience considers the length of time a rater has been involved with rating or the amount of rating he has done, expertise refers to the time when raters scoring performance is consistently good. In the complex task of writing assessment, it is ideally expected that as a result of a rigorous training program, raters' experiential knowledge lead to develop expertise among them. Otherwise, raters translate their own idiosyncratic experiences into their ratings and as a result large amount of variations undermine the scoring task. Needless to say, this chaotic practice has unfortunate consequences for the validity of the scoring and the ensuing decision-making tasks.

Closely related to rater training is the notion of rating scale in writing assessment. Findings of the present study showed that rating scale in its conventional sense (Davies, Brown, Elder, Hill, Lumley, & McNamara, ١٩٩٩) does not exist in the Iranian EFL writing assessment context. While the raters advocated the existence of scales in theory, their practical ratings reflected in TAPs showed that they had their own rating criteria and they rarely applied any of existing rating scales in their rating. This foggy atmosphere can be argued on two main grounds. First comes the fact that particularities of the context considerably influence the whole of assessment and raters' practice is no exception. When EFL raters encounter with rating scales that have been intuitively developed by some native scholars who are quite detached from the realities of EFL context of practice and their developed scales are even criticized in their own context (Fulcher, et al., ٢٠١١; Knoch, ٢٠٠٩; McNamara, ١٩٩٦), it's not odd to see them ignored in the real context of practice in a remote EFL context such as Iran.

Moreover, aligned with the socio-cultural and critical approaches to language testing, the particular context and purpose of assessment are considered as two important yardsticks that considerably affect the construct validity of any rating instrument. Different scholars (Knoch, ٢٠٠٩; Hamp-Lyons, ٢٠٠٧; McNamara, ١٩٩٦; Moskal & Leyden, ٢٠٠٠, Nimehchisalem & Mukundan, ٢٠١١; Norton ٢٠٠٣; Shaw & Weir, ٢٠٠٧; Weigle, ٢٠٠٢) have also emphasized that rating scales should be developed to fit their particular context of use. Therefore, validation studies are required to ascertain whether there is harmony between the realities of the context, objectives of assessment and the particular scale in mind.

Weigle (٢٠٠٢, p. ٤٩) maintains that "construct validity depends crucially on the definition of the ability of interest for a particular testing context." Hence, the definition of ability and subsequent criteria of assessment are minimally determined by the context and purpose of assessment. In other words, scales that come in to existence based on the theoretical and intuitive mentalities in some native contexts fail to function appropriately in other contexts. The consideration of these contextual issues may have implications for the marginal role of extant rating scales in some non-native assessment contexts such as Iran.

The second argument relates the current Iranian EFL writing assessment context to the wider issue of educational policies. Undoubtedly policy-makers as important stakeholders greatly influence the whole assessment enterprise. Writing assessment is no exception as well. As Fraizer (٢٠٠٣) put it writing assessment can be no longer something we pretend to

ISSN ٢٢٥١-٧٣٢٤

do on our own within the realm of academic institutions; rather, power is an element of the assessment process that cannot be ignored, even though we in academic institutions have often sidestepped questions of who controls writing assessment and curriculum in the schools (Huot & Williamson, ١٩٩٧). The pattern of responses resulted from the questionnaire and the idiosyncratic performance of the raters in TAPs can be considered as indications that there is no clear assessment policy with regard to the writing assessment in the Iranian EFL context. In particular, lack of an ordered validation program has created an incoherent practice in which either of writing and its assessment pursues different goals and consequently objectives of the course are lost in the individualistic and impressionistic practice of the raters. Apparently, these two factors which integrate micro and macro perspectives in dealing with the realities of the rating context can greatly account for the missing place of rating scales in their common sense.

٦. Conclusion

Realities of the rating practice in this particular Iranian EFL writing assessment showed that neither of the rater nor the rating scale functioned appropriately. As a remedy, a number of suggestions can be made. With regard to the raters, there is a pressing need for rater training courses in the country. The over-emphasis on formal and mechanical features of writing among the raters showed that Iranian EFL raters had a superficial conception of the writing ability. Moreover, in the absence of rating scales as guiding tools, the raters had an unpredictable scoring route and consequently this great variability significantly affected the reliability and validity of their ratings. Through a concerted training program, the raters would be guided and monitored through their rating. Over time, they would develop the required expertise for their assessment tasks.

Regarding the rating scale, development of a local scale that addresses the particularities of the Iranian EFL assessment is encouraged. The existence of such an objective measurement instrument safe-guards the practice from any inconsistencies in the assessment. In addition, it remarkably facilitates the rater training programs.

The current study has some implications as well. The first implication of the present study is for the EFL writing assessment theory. Present study showed that localizing writing assessment should be preceded with a pathological investigation of the context. In other words, localizing models, rating procedures, etc. in the EFL context does not occur at once; rather, it requires some theoretical foundations to be built among the community of writing instructors. In this way, local attempts would be built on sound theoretical foundations and mere experiential knowledge of the raters does not equate with the local interpretations of the concepts.

Furthermore, present study contributes to EFL performance assessment models and models of rater decision-making processes. Since most of the performance assessment models (McNamara, ١٩٩٦; Skehan, ١٩٩٨, & Fulcher, ٢٠٠٣) were conceived in the ESL performance assessment context and when needed were adapted to EFL contexts, this study demonstrated some idiosyncrasies of EFL context.

Regarding the limitations, despite the country-wide scope of the present study which included ١٠ major state universities, the study was conducted with a limited number of raters, i.e. ٣٠ raters participated in the study overall. Also, due to providing rich data on the mental processes of the participants, TAP was a major data collection instrument in the study. In addition, some of the raters were not well-oriented with the procedure. Although the researcher provided them with a training session prior to data collection, it might be that the raters' mental processes were not completely represented in their verbalized thoughts. Having

ISSN ٢٢٥١-٧٣٢٤

some retrospective measures following think-aloud sessions would supplement the results obtained from the verbal protocols and reduce the veridicality threat (Barkaoui, ٢٠١١).

Acknowledgements

The authors are grateful to the thirty Iranian raters who participated in this study. We also thank the anonymous IJLT reviewer for his/her valuable comments on the earlier drafts of this paper.

References

Alderson, J. C, Clapham, C. & Wall, D. (١٩٩٥). *Language test construction and evaluation*. Cambridge: Cambridge University Press

Bachman, L. F. & Palmer, A.S. (١٩٩٦). *Language testing in practice*. Oxford, UK: Oxford University Press.

Barkaoui, K. (٢٠٠٧). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, ١٢ (٢), ٨٦-١٠٧.

Barkaoui, K. (٢٠١٠). Do ESL essay raters' evaluation criteria change with experience? A mixed-methods, cross-sectional study. *TESOL Quarterly*, ٤٤(١), ٣١-٥٧.

Barkaoui, K. (٢٠١١). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing Journal*, ٢٨ (١), ٥١-٧٥.

Behizadeh, N. & Engelhard, G. (٢٠١١). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing writing*, ١٦(٣), ١٨٩-٢١١.

Brennan, R. L. (١٩٩٧). A perspective on the history of generalizability. *Educational Measurement: Issues and Practice*, ١٦ (١٠), ١٤–٢٠.

Brown, A. (١٩٩٥). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, ١٢, ١-١٥.

Clifford, G. J. (١٩٨٤). Edward L. Thorndike: The sane positivist. In Behizadeh, N. & Engelhard, G. (٢٠١١). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing writing*, ١٦(٣), ١٨٩-٢١١.

Cohen, A. D. (١٩٩٤). Verbal reports on learning strategies. *TESOL Quarterly*, ٢٨(٤), ٦٧٨-٦٨٤.

Congdon, P. J. & Mc Queen, J. (٢٠٠٠). The stability of rater severity in large-scale assessment programs. *Journal of Educational Measurement*, ٣٧(٢), ١٦٣-١٧٨.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (١٩٧٢). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. In Behizadeh, N. & Engelhard, G. (٢٠١١). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, ١٦(٣), ١٨٩-٢١١.

Cumming, A., Kantor, R., & Powers, E. D. (٢٠٠١). *Scoring TOEFL essays and TOEFL ٢٠٠٠ protocol tasks: An investigation into raters' decision making and development of a*

ISSN ٢٢٥١-٧٣٢٤

preliminary analytic framework. (TOEFL Monograph Series, Report No. ٢٢.). Princeton, NJ: Educational Testing service.

Cumming, A., Kantor, R. & Powers, D. E. (٢٠٠٢). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, ٨٦ (١), ٦٧–٩٦.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (١٩٩٩). *Dictionary of language testing*. Cambridge: Cambridge University Press.

DeRemer, M. (١٩٩٨). Writing assessment: Raters' elaboration of the rating task. *Assessing Writing, ٥*, ٧–٢٩.

Eckes, T. (٢٠٠٨). Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing*, ٢٥(٢), ١٥٥-١٨٥.

East, M. (٢٠٠٩). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. *Assessing Writing*, ١٤(٢), ٨٨-١١٥.

Engelhard, G. (١٩٩٤). Examining rater errors in the assessment of written compositions with a many-faceted Rasch model. *Journal of Educational Measurement*, ٣١(٢), ٩٣-١١٢.

Erdosy, M. U. (٢٠٠٤). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions*. (TOEFL Research Report RR-٠٣-١٧). Princeton, NJ: Educational Testing Service.

Fleiss, J. L. (١٩٧١). Measuring nominal scale agreement among many raters. *Psychological Bulletin*.٧٦ (٥), ٣٧٨–٣٨٢.

Fraizer, D. (٢٠٠٣). The Politics of High-Stakes writing assessment in Massachusetts: Why inventing a better assessment model is not enough. *Journal of writing assessment,* ١(٢), ١٠٥-١٢١.

Fulcher, G. (٢٠٠٣). *Testing second language speaking*. London: Pearson Longman.

Fulcher, G., Davidson, F. & Kemp, J. (٢٠١١). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing*, ٢٨(١), ٥-٢٩.

Gamaroff, R. (٢٠٠٠). Rater reliability in language assessment: The bug of all bears. *System*, ٢٨ (١), ٣١–٥٣.

Hamp-Lyons, L. (٢٠٠٧). Worrying about rating. *Assessing Writing*, ١٢(١), ١-٩.

Hawkey, R. & Barker, F. (٢٠٠٤). Developing a common scale for the assessment of writing. *Assessing Writing*, ٩ (٢), ١٢٢-١٥٩.

Huot, B. A. (١٩٩٣). The influence of holistic scoring procedures on reading and rating student essays. In M.M. Williamson & B. A. Hout (Eds.), *Validating holistic scoring for writing assessment: Theoretical and Empirical foundations* (pp. ٢٠٦-٢٣٦). Cresskill, NJ: Hampton.

Johnson, L. & Lim, G. (٢٠٠٩). The influence of rater language background on writing performance assessment. *Language Testing*, ٢٦(٤), ٤٨٥-٥٠٥.

Joreskog, K. G. (٢٠٠٧). Factor analysis and its extensions. In: R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at ١٠٠: Historical developments and future directions*. Mahwah, NJ: Erlbaum.

Kenyon, D. (١٩٩٢). An investigation of the validity of the demands of tasks on performance-based tasks of oral proficiency. In Knoch, U. (٢٠٠٧a). *Diagnostic writing assessment: The development and validation of a rating scale*. PhD dissertation, University of Auckland. Retrieved from: http://researchspace.auckland.ac.nz.

Knoch, U. (٢٠٠٧). *Diagnostic writing assessment: The development and validation of a rating scale*. PhD dissertation, University of Auckland. Retrieved from: http://researchspace.auckland.ac.nz.

Knoch, U. (٢٠٠٩), Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, ٢٦ (٢), ٢٧٥-٣٠٤.

ISSN ٢٢٥١-٧٣٢٤

Landy, F. J., & Farr, J. L. (١٩٨٣). The measurement of work performance: Methods, theory, and application. San Diego, CA: Academic Press.

Lane, S., & Stone, C.A. (٢٠٠٦). Performance assessment. In R.L. Brennan (Ed.), Educational measurement (٤th ed., pp. ٣٨٧-٤٣١). Westport, CT: American Council on Education.

Lim, G. (٢٠٠٩). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing Journal,* ٢٨(٤), ٥٤٣-٥٦٠.

Lim, G.S. (٢٠١١). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, ٢٨(٤), ٥٤٣-٥٦٠.

Lumely, T. (٢٠٠٢). Assessment criteria in a large scale writing test: what do they really mean to the raters? *Language Testing*, ١٩ (٣), ٢٤٦-٢٧٦.

Lumley, T. (٢٠٠٥). *Assessing second language writing: The rater's perspective*. Frankfurt: Peter Lang.

Mc Namara, T. (١٩٩٦). *Measuring second language performance*. Harlow: Longman.

Milanovic, M., Saville, N. and Shuhong, S. (١٩٩٦): A study of the decision-making behavior of composition markers. In Milanovic, M. and Saville, N., (Eds.), *Performance testing, cognition and assessment. Selected Papers from the ١٥th Language Testing Research Colloquium, Cambridge and Arnhem* (pp. ٩٢-١١٤). Cambridge: Cambridge University Press. Retrieved from http://books.google.com/books.

Moskal, M. B. & Leydens, A. J. (٢٠٠٠). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, ٧(١٠). Retrieved from http://PAREonline.net/getvn.asp?v=٧&n=١٠.

Nakatsuhara, F. (٢٠٠٧). Developing a rating scale to assess English speaking skills of Japanese upper-secondary students. *Essex Graduate Student Papers in Language & Linguistics*, ٩, ٨٣-١٠٣.

Nimehchisalem, V., & Mukundan, J. (٢٠١١). Determining the evaluative criteria of an argumentative writing scale. *English Language Teaching*, ٤(١), ٥٨-٦٩.

North, B. (١٩٩٥). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System*, ٢٣ (٤), ٤٤٥–٦٥.

Norton, B. (٢٠٠٣). Bonny Norton responds: On critical theory and classroom practice. In J. Sharkey & K. Johnson (Eds.), *The TESOL Quarterly dialogues: Rethinking issues of language, culture, and power* (pp. ٦٩-٧٣). Alexandria, VA: TESOL Publications.

Pula, J. J., & Huot, B.A. (١٩٩٣). A model of background influences on holistic raters. In Williamson, M.M. and Huot, B.A., editors, *Validating holistic scoring for writing assessment: theoretical and empirical foundations*. Cresskill, NJ: Hampton Press, ٢٣٧±٦٥.

Schaefer, E. (٢٠٠٨). Rater bias patterns in an EFL writing assessment. *Language Testing*, ٢٥(٤), ٤٦٥-٤٩٣.

Shaw, D, S. & Weir, J.C. (٢٠٠٧). *Examining writing: Research and practice in assessing second language writing.* University of Cambridge ESOL Examinations. Cambridge: Cambridge University Press.

Shaw, S. D. (٢٠٠٢). The effect of training and standardization on rater judgment and inter-rater reliability. In Shaw, S. & Falvey,P. (٢٠٠٨). The IELTS Writing Assessment revision project: towards a revised rating scale. *Research Reports,* ١, ١-٢٩٥. Cambridge: Cambridge ESOL.

ISSN ٢٢٥١-٧٣٢٤

Shohamy, E., Gordon, C. M., & Kraemer, R. (١٩٩٢). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, ٧٦(١), ٢٧-٣٣.

Skehan, P. (١٩٩٨). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Spearman (١٩٠٤) Spearman, C. (١٩٠٤). "General intelligence," objectively determined and measured. In Behizadeh, N. & Engelhard, G. (٢٠١١). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing writing,* ١٦(٣), ١٨٩-٢١١.

Tyndall, B., & Kenyon, D. M. (١٩٩٦). Validation of a new holistic rating scale using Rasch multi-faceted analysis. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. ٣٩-٥٧). Clevedon, UK: Multilingual Matters.

Upshur, J. A. & Turner, C. E. (١٩٩٥). Constructing rating scales for second language tests. *ELT Journal*, ٤٩(١), ٣–١٢.

Vaughan, C. (١٩٩١). Holistic assessment: What goes in the raters' mind? In L. Hamp-Lyons (Ed.), *Assessing second language in academic contexts* (pp. ١١١-١٢٦). Norwood, NJ: Albex.

Weigle, S.C. (١٩٩٨). Using facets to model rater training effects. *Language Testing*, ١٥(٢), ٢٦٣-٢٨٧.

Weigle, S. C. (٢٠٠٢). *Assessing writing*. Cambridge, UK: Cambridge University Press.

Weir, C, J. (٢٠٠٥). *Language testing and validation*. Great Britain: Palgrave.

ISSN ٢٢٥١-٧٣٢٤

Appendix
Rating Questionnaire

	Questionnaire Items	**١ Strongly Disagree**	**٢ Disagree**	**٣ No Idea**	**٤ Agree**	**٥ Strongly Agree**
	EFL Writing Courses					
١	The main purpose of writing courses is to develop students' academic writing ability.					
٢	I teach those genres of writing (e.g. expository) which are most related to academic writing.					
٣	For undergraduates, teaching the basics of academic writing is the focus.					
٤	I attend to the purpose, genre and audience in any writing course.					
٥	In BA writing courses, low proficiency of the students causes teachers to ignore teaching some genres.					
٦	EFL writing differs greatly from ESL writing.					
٧	Context (native vs. non-native) affects the level of writing proficiency expected.					
٨	Achieving native-like proficiency in writing is my goal in writing courses.					
	EFL writing scoring	**Strongly Disagree**	**Disagree**	**No Idea**	**Agree**	**Strongly Agree**
٩	When scoring a writing text, I attend to different components of the text such as language, content, structure handwriting, style… one at a time.					
١٠	Scoring different aspects of a text in a separate way (analytic scoring) gives me more confidence when reporting results.					
١١	I have my own way of analytic scoring (i.e. I do not use any existing analytic rating scale).					
١٢	For me, analytic scoring is frustrating and takes a lot of time.					
١٣	I look at the text and based on my own experience in rating, I give a total score.					

ISSN ٢٢٥١-٧٣٢٤

١٤	I look at the text and give a single general score based on a rating scale (holistic scoring).					
١٥	I think giving a score based on my impressions is quite trustable.					
١٦	I use a combination of holistic and analytic scoring in assessing writing.					
١٧	I think all raters have some criteria for their scoring though they may not be in the familiar format of present scales (analytic or holistic).					
١٨	Upon experience, I have learned to keep all the rating criteria in my mind and score based on them.					
	EFL writing scoring	**Strongly Disagree**	**Disagree**	**No Idea**	**Agree**	**Strongly Agree**
١٩	International rating scales such as Jacobs, et al. (١٩٨١) directly guide my rating.					
٢٠	All components of International rating scales are equally relevant to my rating.					
٢١	Rating scale plays a significant role in any assessment of writing.					
٢٢	An explicit rating scale improves validity and reliability of my assessment.					
٢٣	I have my own scoring criteria such as word choice, structure, spelling… but I don't consider them as a kind of analytic scoring.					
٢٤	When it comes to my actual rating, I find existing rating scales less effective (i.e. there are inconsistencies between my criteria and those of the scales).					
٢٥	I think International scales are inappropriate as there are striking differences between the rating criteria in the international scales and those of mine.					
٢٦	The students' command of English affects the selection of rating criteria both quantitatively and qualitatively (e.g. low proficiency might lead me to omit or adjust some of the rating components).					
٢٧	The function of writing assessment in					

ISSN ٢٢٥١-٧٣٢٤

	the present Iranian undergraduate courses is diagnostic.					
٢٨	Rating a composition is quite an individual act (e.g. no concern for inter-rater agreement).					
	EFL writing scoring	**Strongly Disagree**	**Disagree**	**No Idea**	**Agree**	**Strongly Agree**
٢٩	I assess students' compositions to provide them with a profile of their weaknesses and strengths in writing.					
٣٠	It occurs that I have the same rating components as those of International scales but with different levels and descriptors that I define.					
٣١	Students are informed about my rating criteria early in the course.					
٣٢	I think present international rating scales are quite suitable for scoring.					
٣٣	In case of adapting an international rating scale, I redefine the level descriptors to adjust to my specific group of students.					
٣٤	A local (e.g. Iranian) rating scale for writing assessment is needed to assure the validity of the scores.					
٣٥	Particularities of any context would affect the rating components of any rating scale.					
٣٦	As a rater, I am quite aware of different rating scales.					
٣٧	Lack of a common rating scale would lead to bias, inconsistency and leniency/severity among the raters.					
٣٨	Diversity of raters' criteria in writing assessment makes the construct of writing ability rather vague and elusive.					
٣٩	The existence of a common rating scale would lead to a more fair writing assessment.					
٤٠	My rating experience justifies the scores I assign out of my general impressions of the text.					