English Language Teaching; Vol. 5, No. 8; 2012
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Problematizing Rating Scales in EFL Academic Writing Assessment: Voices from Iranian Context

Batoul Ghanbari[1], Hossein Barati[1] & Ahmad Moinzadeh[1]

[1] English Department, Faculty of Foreign Languages, University of Isfahan, Iran

Correspondence: Batoul Ghanbari, English Department, Faculty of Foreign Languages, University of Isfahan, Hezar Jerib Street, Isfahan, Iran. Tel: 98-917-775-3178. E-mail: btghanbari@gmail.com

Received: April 17, 2012 Accepted: May 10, 2012 Online Published: July 2, 2012

doi:10.5539/elt.v5n8p76 URL: http://dx.doi.org/10.5539/elt.v5n8p76

Abstract

Along with a more humanitarian movement in language testing, accountability to contextual variables in the design and development of any assessment enterprise is emphasized. However, when it comes to writing assessment, it is found that multiplicity of rating scales developed to fit diverse contexts is mainly headed by well-known native testing agencies. In fact, it seems that EFL/ESL assessment contexts are receptively influenced by the symbolic authority of native assessment circles. Hence, investigating the actualities of rating practice in EFL/ESL contexts would provide a realistic view of the way assessment is conceptualized and practiced. To investigate the issue, present study launched a wide-scale survey in the Iranian EFL writing assessment context. Results of a questionnaire and subsequent interviews with Iranian EFL composition raters revealed that rating scale in its common sense does not exist. In fact, raters relied on their own internalized criteria developed through their long years of practice. Therefore, native speaker legitimacy in the design and development of scales for the EFL context is challenged and the local agency in the design and development of rating scales is emphasized.

Keywords: performance assessment, EFL writing assessment, rating scale, construct validity, EFL rater

1. Introduction

Writing is a fundamental aspect of academic literacy and communicative competence in the current educated world (Behizadeh & Engelhard, 2011). The bulk of studies on writing assessment shows that articulating sound assessment of writing achievement is significant. As a result, it affects the quality of writing instruction (Hamp-Lyons, 1991, 2001; Messick, 1996). Hence, a rigorous investigation of writing assessment procedures is necessary in assuring whether it fulfills its function in the right way.

Among the many factors involved in performance assessment (McNamara, 1996), rating scale as spells out the criteria against which judgment of quality are achieved has an important role. In addition, as the scale embodies tacitly or explicitly the theoretical basis of the test, the design and development of the scale is of paramount importance in the validity of the assessment (McNamara, 1996).

The original function of rating scale as stating the purposes of the assessment in a specific assessment context addressing particular group of test-takers has been an important motivation for the presently diverse existence of rating scales in writing assessment **in especial**. A large number of scales with each having their specific language in defining the descriptors, levels and rankings all confirm that the above- mentioned elements of context and purpose exert considerable weights in the development of rating scales. Apparently, construct validity is relegated with the notions of purpose and context. Therefore, within current conceptualization of construct validity in writing assessment (Nemati & Ahmadi Shirazi, 2009), any assessment criterion that is not developed according to the specifications of the context and concomitantly not addressing the particular purposes of the assessment is not appropriate and useful for the assessment of writing in that context.

However, despite a booming in the number of rating scales in writing assessment that aim to meet the intricacies of the specific context of their development, there exists a receptive and unquestioning mode regarding the use of rating scales in many EFL contexts. In fact, rating scales are either put aside or in case of any application, some existing, internationally-known ones are drawn upon (Barkaoui, 2007; McNamara, 1996). Ignoring the particularities of the context and the specific objectives of assessment threaten the construct validity of the assessment in the first place. In other words, judgment of performance based on the above rating scales does not provide a true picture of the test-takers' abilities.

Acknowledging the above chaotic situation in the use of rating scales, an exploration of current practice in EFL writing assessment context seems warranted. An investigation of the context would show to what extent rating scales are in use and in case it is, whether it addresses the particularities of the EFL context and consequently the goals of assessment. For this aim, present study was launched to examine the Iranian context as a particular EFL writing assessment context to probe the functionality of the rating scales through analyzing the hands-on attitudes and experiences of raters. Findings would ground the on-going validity arguments on rating scales in this particular EFL context and in this way a more realistic picture of the functionality of the rating scales is shown in the context.

2. Theoretical Background

2.1 Rating Scale in Writing Assessment

A historical overview of the measurement theories involved in writing assessment in the 20th century reveals two dominant traditions, i.e. test-score tradition and the scaling tradition (Behizadeh & Engelhard, 2011). The test-score tradition that originates from the seminal work of Spearman (1904) is 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 of Generalizability Theory (Brennan, 1997; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), factor analysis and structural equation models (Joreskog, 2007).

Scaling theory as the other influential measurement tradition originally rooted in psychophysics in 19th century. The focus of scaling tradition is to provide some variable maps that configure the location of both items and individuals onto a latent variable scale that represents a construct. Inspired by E. L. Thorndike's epigram that 'Whatever exists at all exists in some amount' (Clifford, 1984), scaling tradition has continued through presenting different kinds of rating scales which have its theoretical basis in the strong statistical item-response theory.

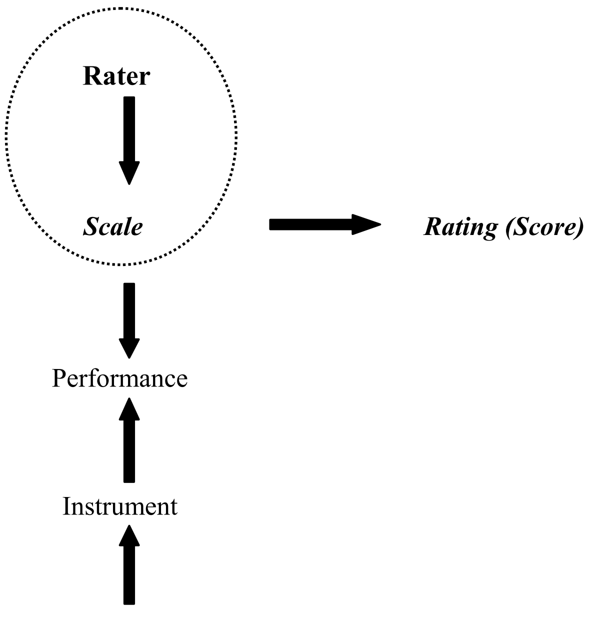

Figure 1. Factors in performance assessment: (adapted from McNamara, 1996)

The general appeal to rating scales was for the assessment of behavior or performance where judgments of quality against some rating scale are desired. In a much-quoted figure, McNamara (1996) schematically showed different components involved in a typical performance assessment (Figure 1). As shown, the interactive component of rater-rating scale as a distinct characteristic of performance assessment underscores the fact that rating scale and rater have to be rigorously studied. In other words, the qualities of the rating scale along with the characteristics of the rater significantly affect the ratings that are given regardless of the quality of the performance.

Furthermore, the conceptualization of rating scale as part of the test construct has opened a new horizon on examining different aspects of rating scales functioning in performance assessment. Writing assessment as a kind of performance assessment is affected by the quality of the rating scale used. To a large extent, the common thread of arguments on rating scales converges on the important issue of construct validity. Identifying the importance of rating scale in the quality of assessment, Weigle (2002, p. 109) summarizes McNamara (1996) on the centrality of the rating scale to the valid measurement of the writing construct:

The scale that is used in assessing performance tasks such as writing tests represents, implicitly or explicitly, the theoretical basis upon which the test is founded; that is, it embodies the test(or scale) developer's notion of what skills or abilities are being measured by the test. For this reason, the development of a scale (or set of scales) and the descriptors for each scale level are of critical importance for the validity of the assessment.

2.2 Validity Considerations

History of measurement theory in the recent hundred years shows that reliability or precision of measurement has consistently been the pursued concern of assessment specialists and validation issue in rating criteria has been dealt with mostly in passing, if at all (McNamara, 1996). However, as the criteria of assessment often make implicit reference to a psychological construct or constructs that turn out to be the goal of measurement, deciding on the criteria against which performance quality is judged considerably affects the construct validity of the test.

For example, McNamara (1996) after counting the dearth of research on construct validity in rating scales proceeds to question the assumptions behind the rating scales in order to see how they make sense in the contexts of their use. In his study, it was shown that idealization of native speaker performance is frequent in such scales which in turn have implications for the validity of the tests they are used to report, and for the fairness of gate-keeping decisions made on the basis of their use (McNamara, 1996, p. 183).

Another motivation to give more weight to validity in writing assessment is the general appeal to social and cultural nature of writing. On a par with the more ethical approaches to language testing (Hamp-Lyons, 2001), context, purpose and audience are considerably involved in designing and developing assessment procedures. Quoting Weigle (2002), in case social and cultural nature of writing is accepted, "the implication for the testing of writing is that writing ability cannot be validly abstracted from the contexts in which writing takes place"(p. 22). In other words, the very context of assessment minimally influences the kind of criteria used for the assessment.

Scoring Validity
Criteria/rating scale
Rater characteristics
Rating process
Rating conditions
Rater training
Post-exam adjustment
Grading and awarding

Figure 2. Aspects of scoring validity for writing (adapted from Weir 2005, p.47)

Similarly, Weir and Shaw (2007) in their socio-cognitive framework which views language testing and validation within a contemporary evidence-based paradigm, consider scoring validity along with cognitive validity and context validity as important elements to provide theoretical, logical and empirical evidence to support validity claims and arguments about the quality and usefulness of writing tests. In this conceptualization of validity, the first scoring validity parameter is that of the criteria and type of rating scale to be used (Figure 2). In fact, the choice of appropriate rating criteria and the consistent application of them by trained raters are considered significant in the valid assessment of ESL/EFL language performance (Alderson, Clapham and Wall 1995, Bachman and Palmer 1996, McNamara 1996). More specifically, appropriacy of rating scales has been attended by some scholars (Hamp-Lyons 1997; Norton 1997; Shohamy 1993) who believed that ethical accountability in language testing and assessment necessitates to rigorously consider the assumptions implicit in the rating rubrics we use, the decisions that we make on the basis of such rubrics, and the consequences these decisions might have for the life chances of test-takers.

Following the arguments on the appropriacy of rating scales, Moskal and Leyden (2000) dealt with the validity of rating rubrics in both apriori stages of the design and development of the scale and also the posteriori stage of checking the aspects of validity in the developed rubrics. In the apriori stage, after stating the purposes and objectives of the assessment, they state that scoring criteria should be developed in a way to match those pre-stated objectives. If some of the objectives are not represented in the rating rubrics or some of the rating rubrics are not related to the objectives, then appropriateness of the assessment and consequently rating rubrics is loosely held (Moskal & Leyden, 2000). A posteriori examination of validity may also be conducted after a preliminary development of rating rubrics. At this stage, different types of validity evidence may be examined concerning the developed rating scale. These concerted efforts are aimed to safeguard the validity of assessment and make the results of assessment be meaningful in the context.

Overall, the current turn to validity considerations in rating scales confirm the significance of rating procedures in establishing sound assessment. In the words of McNamara (1996), having the central role in the construct validity of much performance testing, rating scales deserve to be the subject of a much greater level of research.

3. The Present Study

Due to high importance of rating scales in writing assessment, it is expected that they have been intensively researched and discussed. In language assessment circles, however, it is not so. In the words of McNamara (1996, p. 182):

We are frequently simply presented with rating scales as products for consumption and are told little of their provenance and of their rationale. In particular, we too frequently lack any account of empirical evidence for their validity.

Moreover, the guiding principles of context and purpose demand a careful analysis of the rating scales in every particular assessment context. Inspired by the above arguments, present study was launched to widely explore the current rating practice in the particular EFL writing assessment context of Iran. Attempt was made to enquire whether rating scale as conceived in its theoretical configuration exists and also to reveal those hidden perceptions of raters on different aspects of rating scales. In order to achieve the above goals, the following research questions were posed in this study:

1. Do Iranian EFL raters use any explicit rating scale in their writing assessment?

2. How do Iranian EFL raters perceive context in the Iranian EFL writing assessment?

3. How do Iranian EFL raters perceive International native-developed rating scales in the Iranian EFL writing assessment?

4. Method

4.1 Participants

Overall 40 raters participated in the study. 10 raters took part in the pilot phase and in the main phase of the study 30 raters were present. Participants were either PhD holders as TEFL (Teaching English as a Foreign Language) teachers or were doctoral students in TEFL. They also varied in terms of age, gender and TEFL teaching background. Few of these people had passed any rater training courses, but all had a minimum 5 years of teaching and assessing writing experience. In the pilot phase, raters were selected based on their expertise and availability at the time of research. In the main phase of the study, however, a body of 30 raters was randomly selected from 10 major state universities in Iran. Table 1 describes the typical profile of two participants in this study.

Table 1. Typical profile of two raters participating in the study

	PhD[a]	PhD Student[b]
Role at the time of Research	TEFL teacher	TEFL student
EFL teaching experience (Years)	20	8
EFL rating experience (Years)	14	6
Received training in rating composition	No	No

n[a] = 19 n[b]= 11

4.2 Research Design

Due to the presence of both qualitative and quantitative modes of research, the design of the present study was a mixed-method one. In the first pilot phase of the study, the researcher conducted interviews with 10 experienced raters and in the main phase an exploratory questionnaire was developed and sent to 30 experienced EFL raters in the country. A further interview session with some of the raters was held and they were asked to elaborate on their responses to the questionnaire.

Table 2. The structure of the rating questionnaire

Subscale	No. of items	Description
1. Existence/Application of rating scale	10	Investigating whether there exists any rating scale in use
2. Context in writing assessment	6	Investigating the effect of contextual factors in EFL writing assessment
3. International rating scales	6	Investigating the appropriacy of Internationally-known rating scales in EFL writing assessment
4. Others	18	Investigating other aspects of EFL writing assessment

4.3 Context of the Study

Iranian EFL undergraduate writing assessment was the desired context of the present study. For this purpose, English departments in 10 major state universities were selected for inclusion in the study. As two courses of paragraph writing and essay writing were compulsory for undergraduates, it assured the researcher that writing assessment and particularly the issue of rating was of concern in the context.

4.4 Instruments

The instruments used in this study were two rounds of interviews and a researcher-made questionnaire. The first semi-structured interview scheme was designed and developed to be used in the pilot phase of the study. Its flexibility allowed the researcher to elicit the raters' views on different issues concerned with rating. Later, researcher developed a 40-item questionnaire aimed to tackle different aspects of rating. In writing the questionnaire items, researcher used the ideas of the expert EFL raters in the pilot interview sessions. After developing the first draft of the questionnaire, the researcher asked for the assistance of two expert raters to comment over the structure of the questionnaire and more importantly if they would allocate the items under the same subscales as the researcher. After resolving some ambiguities over a few items, two were found to have similar classification of items in the related subscale as that of the researcher. The inter-rater reliability estimate obtained was .95 which indicated a high degree of agreement among the raters and the researcher in classifying the items into similar subscales. In addition, after administering the questionnaire, the reliability for the total questionnaire estimated through Cronbach Alpha was .71 which due to its multi-scale nature such a moderate internal consistency measure is justified.

In a follow-up interview, the researcher asked some of the respondents to elaborate on their responses. The objective of the interview at this stage was to induce any further comments over some unexpected results that researcher could not interpret on the basis of questionnaire data.

4.5 Procedure

4.5.1 Data Collection

At the very beginning of the study and in order to obtain a general picture of the rating situation, the researcher conducted a pilot study with 10 experienced and well-known composition raters in the country. This body of raters who came from 5 prestigious state universities in Iran was contacted and interviewed by the researcher. The interview items followed a main theme that asked the raters to elaborate on the how of their ratings. Responses provided by the raters provided the researcher with some novel perspectives on rating and the rating scale in the EFL composition rating context of Iran. The interview data that was audio-recorded was later transcribed by the researcher for a qualitative analysis of the data. Through an in-depth content analysis of the interview data, some general patterns were derived and through further readings of the interview protocols they got more polished as distinct emerging patterns.

Following the pilot phase of the study, researcher developed a questionnaire on the rating practice. It aimed to explore different aspects of rating (Appendix 1). It's worth mentioning that present study was part of a larger project aimed to investigate and improve the rating practice in Iran. Therefore, the questionnaire included some subscales each including a number of items. In the main phase of the study, researcher randomly selected 3 raters in each of 10 major state universities in the country. The raters were asked to fill out the questionnaire. Following administering the questionnaire, some of the raters were asked for a retrospective interview to elaborate on the process of responding to the questionnaire and clarify some of their unexpected responses. The addition of an interview component right after the questionnaire is justified on the grounds that in case the researcher faces some unexpected results, he or she cannot interpret those on the basis of the questionnaire data. So, in order to corroborate the final findings the quantitative findings of the questionnaire is supplemented by the interview results (Dornyei, 2007). Therefore, after responding to the questionnaire, 7 of the raters agreed to participate in a follow-up interview (Appendix 2). The interview session was audio-recorded and lasted for 30-45 minutes. The questions posed to the raters centered on the first and third subscale of the questionnaire. Researcher asked the raters to elaborate on the items that were controversial. On the whole, 30 raters participated in the first stage of data collection- i.e. questionnaire and 7 of them for a subsequent interview. Clearly, the number was appropriate enough to embark on a detailed and in-depth analysis of the rating situation in the country.

4.5.2 Data Analysis

Both types of qualitative and quantitative data analyses were conducted in the present study. In order to analyze the interviews, the qualitative method of content analysis was utilized. The procedure followed by the researcher was first to carefully transcribe the audio-data by the researcher. Later, upon multiple readings of the interview transcripts, the emerging patterns were segmented and coded for analysis. To analyze the questionnaire, using SPSS-release 19- the required descriptive statistics was obtained.

5. Results

5.1 Quantitative Analysis: Analysis of the Questionnaire

Before addressing the research questions, it seems helpful to provide a sketch of the related subscales along with the associated items in the questionnaire.

As table 2 above shows, multiple scales were included in the structure of the questionnaire. In order to answer the first research question, items in the first sub-scale were analyzed. Descriptive statistics related to the performance of raters in the first subscale is presented in Table 3 below.

Table 3. Descriptive statistics for the first subscale in the questionnaire

Items	**Negative**		**Positive**		**No Idea**
	N	Percentages %	N	Percentages %	Percentages %
11	12	40	14	46.7	13.3
13	18	60	11	36.7	3.3
14	13	43.3	14	46.7	10
15	17	56.7	10	33.4	10
16	3	10	21	70	20
18	7	23.3	21	70	6.7
21	4	13.3	20	66.6	20
22	2	6.7	25	83.3	10
23	12	40	15	50	10
40	7	23.3	20	66.7	10
Mean	9.5	**31.66**	17.1	**57.01**	**11.33**

The individuals' responses to the items in the questionnaire were measured on a five-point Likert scale. However, for a more detailed analysis, responses in the two ends of the scale were collapsed to form one general category. Therefore, strongly agree and agree formed the category of positive and strongly disagree and disagree were merged to form the category of negative. The category of no idea was also considered in the analysis. Regarding the wording of the items in this subscale that presented a single idea in the form of a positive statement in the questionnaire (Appendix 1), the results clearly confirmed that presently there is no explicit rating scale in use in Iran. Analyzing the pattern of responses showed that although raters relied on their experience when scoring (items 13 & 40) and also their impressionistic and subjective scoring (item 15), they unanimously acknowledged that rating scales fulfilled an important role in writing assessment and it considerably enhanced the psychometric qualities of the assessment (items 21 & 22). Results of the interviews conducted after the questionnaire verified the pattern of responses obtained.

Similarly, in order to answer the second and third research questions, researcher analyzed the related items in the desired subscale. Displayed in Tables 4 & 5 below are the required descriptive statistics for the second and third research questions respectively.

Table 4. Descriptive statistics for the second subscale in the questionnaire

Items	**Negative**		**Positive**		**No Idea**
	N	Percentages %	N	Percentages %	Percentages %
5	7	23	22	73.4	3.3
6	8	26.7	13	43.4	30
7	2	6.7	23	76.7	16.7
8	13	43.4	15	50	6.7
26	5	16.7	25	83.4	-
35	3	10	22	73.3	16.7
Mean	6.33	**21.08**	20	**66.7**	**12.23**

Table 5. Descriptive statistics for the third subscale in the questionnaire

Items	**Negative**		**Positive**		**No Idea**
	N	Percentages %	N	Percentages %	Percentages %
19	14	46.7	11	36.7	16.7
20	15	50	7	23.3	26.7
25	15	50	3	10	40
30	2	6.7	18	60	33.3
32	7	23.3	7	23.3	53.3
33	2	6.7	23	76.7	16.7
Mean	9.16	**30.56**	11.5	**38.33**	**31.11**

The second research question explored Iranian EFL raters' perceptions over the role of context in their writing assessment. As table 4 shows, a substantial percentage of raters believed that contextual factors influenced their practice. As demonstrated in the above table, compared to agreed ones the percentage of dubious or disagreed responses was considerably lower.

Regarding the last research question, results showed that raters were receptive toward Internationally-known rating scales in the Iranian EFL context. But, as Table 5 shows there was a considerable doubt over the appropriacy of rating scales in the context. The sum of mean percentages in the two groups-negative and no idea-reached 61.67 which was robust enough to challenge the views of the raters in the positive camp. Figure 3 below presents the results of the questionnaire in a clear way.

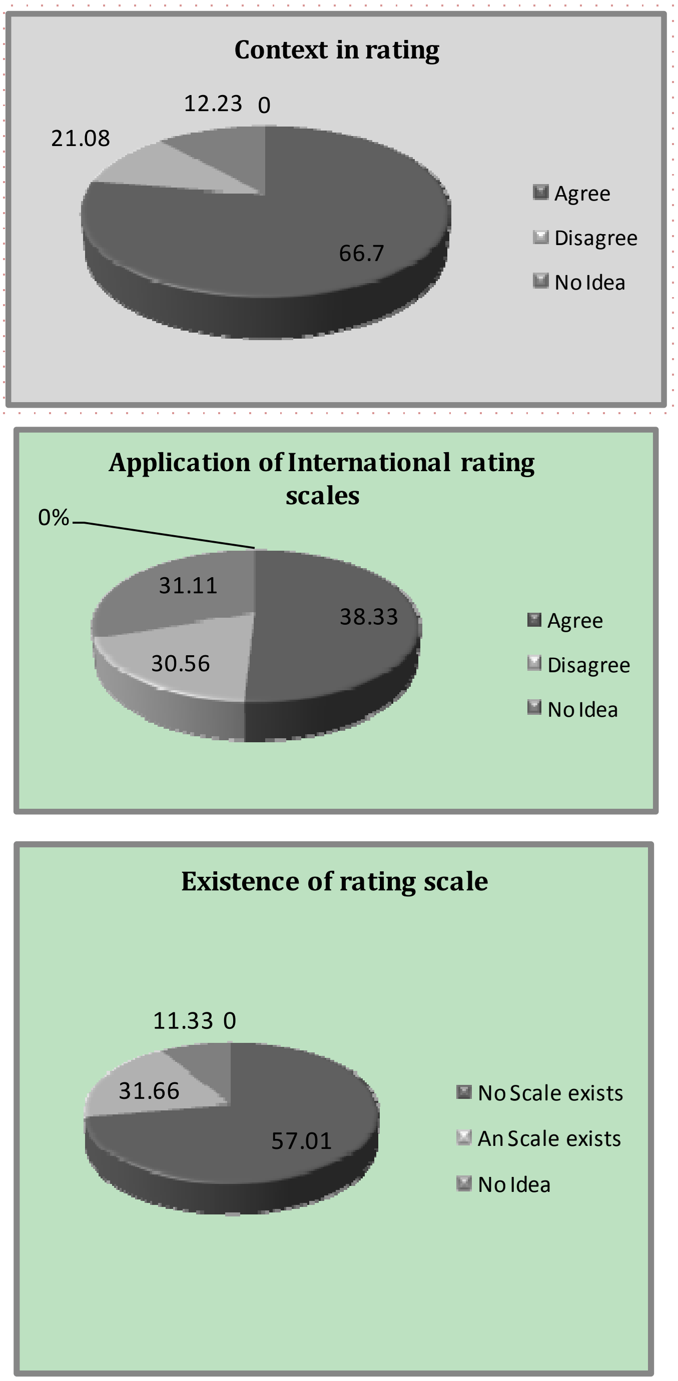

Figure 3. Results of the questionnaire on the existence/lack of rating scales, effect of context in rating and the appropriacy of International rating scales in the EFL writing assessment context of Iran

5.2 Qualitative Analysis: Analysis of the Interviews

As mentioned before, the purpose of interviews at this stage was to clear the ambiguous and unexpected results obtained in the quantitative phase of the study. Therefore, interviewees were mostly asked to elaborate on their pattern of responses in the first and third subscales. An analysis of the interview transcripts led to the emergence of two major themes. The two major patterns are presented below:

- *Existence of rating criteria but no explicit and objective rating scale in place*

The most frequently recurring theme in the interviews was that raters never identified themselves as having no rating criteria. Rather, they considered their own way of rating quite useful and the one which resulted out of their long experience in rating. They reasoned that their unique context of teaching and assessing writing had experientially directed them to their criteria of rating. As an example, rater 2 who had his analytic scoring approach believed that:

Rater 2: "look! I feel quite at ease with this rating approach. Within my past years of teaching and assessing writing, it has been quite helpful!"

Application of raters' own criteria in writing assessment was considered so obvious and they believed that upon years of experience they have internalized the criteria and apply them to the rating task quite automatically. Rater 3, when asked to verbalize his criteria of rating found it a difficult task:

Rater 3: "You know, [uhmm], I can't mention them explicitly. They are in my mind and when rating I brief them in my general impression of the composition."

None of the interviewees spoke of an explicit rating scale being in use in their rating; rather they developed a kind of *ad hoc* rating scale based on their own criteria when involved in the rating task.

- *Ambivalent attitude to native criteria in International rating scales*

A brief look at the results of the subscale on the appropriacy of International rating scales shows that there was a mixed attitude on the scales in the Iranian EFL rating context. On the one hand, they did not reject the widely-known scales such as Jacobs, et al. (1989) and considered such scales as authority in this regard. However, on the other hand, they rarely applied a rating scale directly in their practice. In an interesting reflection over the use of International rating scales, Rater 1 said so:

Rater 1: "I know that there are many rating scales, but I have not studied them…Actually, in my idea, there is no need for that! I think although they have been developed by natives, they are armchair scales. They have developed out of theoretical considerations and hence they do not tackle the particular context of writing assessment here in Iran."

Raters were quite ambivalent towards the direct application of native rating scales. The concerns expressed were not the ones usually expressed as hurdles in the application of an explicit rating scale; rather it covered issues such as context appropriacy, EFL learners' command of English, developmental issues in rating scales…Some of the problems that hindered the application of native scales was expressed by raters 3 and 5 as below:

Rater 3: "Well… the difference here lies in general educational goals and perceived needs of the learners. I think a writing course in an EFL undergraduate program cannot be compared with the one in a native context. I think categories of rating are the same but weighting of them and the way descriptors are defined differ. In fact have to differ! Descriptors must be modified to suit the context."

Rater 5: "Personally I believe that a scale developed for a foreign country is not necessarily applicable in our context. The least grounds of difference concern the proficiency level of learners here and there! For example, in a native rating scale the category of style might be included as an important rating criterion, but here undergraduate writers do not reach the level of style in their writing, so what's the use of having a category such as style in an EFL writing assessment context such as Iran?!"

A utilitarian view to International rating scales was also reported by some raters. While explaining his own rating procedure, Rater 6 commented that he strategically approached the native rating scales:

Rater 6: "I have not directly adopted my criteria of assessment from these scales; rather I consider them as good models of comparison. I look at the layout, descriptors for each category, scoring, and weighting of rating components and try to improve my rating criteria. Of course in my own way!"

While the comments expressed by the raters shed a positive light on the International rating scales, but they had serious reservations about the validity of the scales in the EFL writing assessment context of Iran. The vague attitude towards the legitimacy of native rating scales is taken up in more depth in the next section.

Above, quantitative and qualitative results for the three research questions were presented. The following section discusses the obtained results.

6. Discussion

Since Messick's (1989) introduction of the validity as an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores, the issue has found an unprecedented significance. Moreover, aligned with the socio-cultural and critical approaches to language testing, context and purpose of assessment are considered as two important yardsticks that considerably affect the construct validity of any rating scale. Different scholars (Knoch, 2009; Hamp-Lyons, 2007; McNamara, 1996; Moskal & Leyden, 2000, Nimehchisalem, 2011; Norton 2003; Shaw&Weir,2007; Weigle, 2002) have emphasized that rating scales should be developed to fit their particular context of usage. Therefore, validation studies are required to ascertain whether there is congruence between the realities of the context, objectives of assessment and the particular scale in mind. As part of a larger project that was aimed to develop a local rating scale in the Iranian EFL context, a survey was carried out to investigate the attitudes of Iranian expert raters along three lines of inquiry. The obtained results showed that raters conducted their impressionistic and subjective ratings based on their own criteria developed and internalized over the path of their rating practice. Their elaborations in the subsequent interviews showed that they firmly believed in their rating despite the fact that there was no use of any explicit rating scale in their rating. In sum, current impressionistic rating situation in the practice of Iranian EFL writing assessment can be argued on two main grounds:

At a general macro-level, the current state of rating practice can be the by-product of general educational policies in the EFL context. In the words of Fraizer (2003) writing assessment can be no longer something we pretend to 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 and Williamson, 1997). Writing and its assessment is not attended to seriously in the EFL curriculum. The practical rating context seriously lags behind the current arguments on validation of rating scales. In fact, pragmatic concerns including ever-increasing student population, stagnant funding, lack of rater training courses, and more seriously lack of an ordered validation program have caused raters to feel safe with their scoring and as a result, a vague rating situation combining elements of criterion-and norm-referenced approaches (Barkaoui, 2007) prevails.

At a local and micro-level, there is the fact that particularities of the context considerably influence the whole of assessment and raters' practice is no exception. When EFL raters are provided with rating scales that have been intuitively developed by some native scholars who are detached from the realities of EFL context of practice and their developed scales are even criticized in their own context (Fulcher,2011; Knoch, 2009), it's not odd to see them ignored in the real context of practice. Drawing on the framework of Bachman& Palmer (1996) and in order to evaluate the usefulness of rating scales, Weigle (2002) presents six criteria of reliability, construct validity, authenticity, interactiveness, practicality and impact. In this framework, Weigle (2002, p. 49) maintains that construct validity depends crucially on the definition of the ability of interest for a particular testing context. Hence, 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 context fail to function correctly in other contexts. The ambivalence over the practicality of Internationally-known rating scales in the Iranian EFL writing assessment context is a case in this regard.

As a result, the current practice of empirically developing rating scales has been triggered by strong contextual motivations. Dubious assumptions behind the rating scales (McNamara, 1996; Norton, 2003), unreal and impoverished picture of the context (Fulcher, 2011), low profile of psychometric qualities (Weigle, 2002; Moskal & Leyden, 2000) and ethical accountability movement in language testing (Hamp-Lyons 2001; Norton 1997; Shohamy 1993) have all prepared the ground for an empirical investigation of rating scales.

Majority of the raters in this study believed that native scales have to be appropriated in the context before application. In their ideas, unmediated application of native rating scales would surface a hidden conflict between the assumptions behind these scales on the one hand and the realities of the local context on the other hand. As an example, Haswell (2005, p. 2) questions the ESL Composition Profile (Jacobs, et al., 1981) as a commonly-used rating scale in diverse EFL/ESL contexts and contends that how this known scale is a tool "that is no different from dozens of similar guides by which raters have decided, and continue to decide, the academic fate of thousands upon thousands of second language students". In a similar vein, McNamara (1996) strongly questions the validity of rating scales and by tracking the origin of scale tradition in the FSI test in the 1950s shows how successive rating scales developed over the last four decades have been heavily influenced by the assumptions, and even the wording of the original work, and rare empirical validation has been done.

7. Conclusion

Despite current scholarship that are concerned with improving the psychometric characteristics of the rating scales (Barkaoui, 2007; Ecks, 2008; Lim, 2011; Johnson & Lim, 2009 and Schaefer, 2008), present study set out to localize the issue of rating scales in the EFL context of Iran. Findings vividly indicated a vacancy for an objective measure in EFL writing assessment. The simple alternative of international rating scales was also found to be impractical in the context. In fact, numerous rating scales which due to an ambiguity and inaccuracy in the definition of writing construct draw on the context and purpose to justify construct validity did not match the particular context of the study simply on the grounds they were developed (i.e. contextual appropriacy). It was argued that these scales are not neutral instruments; rather, there are serious value-laden assumptions woven in their structure. Originating in the 1950s at the heyday of psychometric-structuralist period in which a view of second language proficiency and its relation to first language proficiency gave the native speaker an important defining role as a kind of benchmark is a commonly shared view by many known rating scales expressed in their band level descriptions where the native level of proficiency is emphasized.

Concluding the article, two major recommendations are put forward. First, Due to the growing importance of written literacy as one of the educational achievements of every language learning enterprise, a major reconsideration in the place and importance of writing courses by the higher educational bodies is mandatory. As a first step in this regard, a revision in the assessment procedure and professional development of composition teachers through rater training courses would improve the current chaotic state. Second, problems counted in the present study encourage the development of a local rating instrument in the context. This proposal which is strongly supported in the literature would resolve many of the controversies on using rating scales in the context. Through addressing the particularities of the context and equipped with psychometric rigor, a local rating scale is a good option in safeguarding the soundness of writing assessment.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107. http://dx.doi.org/10.1016/j.asw.2007.07.001

Behizadeh, N., & Engelhard, G. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing writing, 16*(3), 189-211. http://dx.doi.org/10.1016/j.asw.2011.03.001

Brennan, R. L. (1997). A perspective on the history of generalizability. *Educational Measurement: Issues and Practice*, *16*(10), 14–20.

Clifford, G. J. (1984). *Edward L. Thorndike: The sane positivist*. Middletown, CT: Wesleyan University Press.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability of scores and profiles. In Behizadeh, N., & Engelhard, G. (2011), Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing*, *16*(3), 189-211.

Dornyei, Z. (2007). *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford: Oxford University Press.

Eckes, T. (2008). Rater types in writing performance assessments: a classification approach to rater variability. *Language Testing*, *25*(2), 155-185. http://dx.doi.org/10.1177/0265532207086780

Fraizer, D. (2003). The Politics of High-Stakes writing assessment in Massachusetts Why inventing a better assessment model is not enough. *Journal of writing assessment, 1*(2), 105-121.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, *13*(2), 208–238. http://dx.doi.org/10.1177/026553229601300205

Fulcher, G. (2003). *Testing second language speaking*. London: Pearson Longman.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performanc decision trees. *Language Testing, 28*(1), 5-29. http://dx.doi.org/10.1177/0265532209359514

Hamp-Lyons, L. (2001). Ethics, fairness(es) and developments in language testing. *Studies in* Language Testing, *9,* 30–34. Cambridge: Cambridge University Press.

Hamp-lyons, L. (1997). Ethics in language testing. In Claham, C., & D. Corson (Eds.) (1997), *Language Testing and Assessment: Volume 7, The Encyclopedia of Language and Education*. Dordrecht: Klewer academic publishers.

Hamp-Lyons, L. (1996). Ethical test preparation practice: The case of the TOEFL. Paper presented at the 18th Annual Language Testing Research Colloquium, Tampere, Finland.

Hamp-Lyons, L. (1991). Reconstructing academic writing proficiency. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 127–154). Norwood NJ: Ablex.

Haswell, R. H. (2005). Researching Teacher Evaluation of Second Language Writing via Prototype theory. In Paul Matsuda and Tony Silva (Eds.), *Second Language Writing Research: Perspectives on the Process of Knowledge Construction* (Erlbaum, 2005), 105-120.

Huot, B., & Williamson, M. (1997). Rethinking portfolios for evaluating writing: Issues of assessment and power. In K. B. Yancey, & I. Weiser (Eds.), *Situating portfolios: Four perspectives* (pp. 43-56). Logan: Utah University Press.

Jacobs, H. L., Zinkgraf, S. A., Wormouth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: a practical approach*. Rowely, MA: Newbury House.

Johnson, J. S., & Lim, G. S. (2009). The influence of rater language background on writing performance assessment. *Language Testing*, *26*(4), 485-505. http://dx.doi.org/10.1177/0265532209340186

Joreskog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck, & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions.* Mahwah, NJ: Erlbaum.

Knoch, U. (2007). Little coherence, considerable strain for reader: A comparison between two rating scales for the assessment of coherence. *Assessing Writing*, *12*(2), 108-128. http://dx.doi.org/10.1016/j.asw.2007.07.002

Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275-304. http://dx.doi.org/10.1177/0265532208101008

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*(2), 81-96. http://dx.doi.org/10.1016/j.asw.2011.02.003

Lim, G. (2009). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing Journal, 28*(4), 543-560. http://dx.doi.org/10.1177/0265532211406422

Mc Namara, T. (1996). *Measuring second language performance*. Harlow: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York, NY: Macmillan/American Council on Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. http://dx.doi.org/10.1177/026553229601300302.

Moskal, M. B., & Leydens, A. J. (2000). Scoring rubric development: validity and reliability. *Practical Assessment, Research & Evaluation*, *7*(10). Retrieved October 16, 2011, from, http://PAREonline.net/getvn.asp?v=7&n=10.

Nemati, M., & Ahmadi Shirazi, M. (2009). Writing Assessment Perspective: Too Simplistic or too Sophisticated? *TELL*, *3*(9), 27-63.

Nimehchisalem, V., & Mukundan, J. (2011). Determining the Evaluative Criteria of an Argumentative Writing Scale. *English Language Teaching*, *4*(1), 58-69.

Norton, B. (2003). 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. 69-73). Alexandria, VA: TESOL Publications.

Norton, B. (1997). Language, identity and ownership of English. *TESOL Quarterly*, *31*(3), 409–429. http://dx.doi.org/10.2307/3587831

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. *Language Testing*, *25*(4), 465-493. http://dx.doi.org/10.1177/0265532208094273

Shaw, D, S., & Weir, J. C. (2007). *Examining writing: Research and practice in assessing second language writing.* University of Cambridge ESOL Examinations. Cambridge University Press.

Shohamy, E. (1993). The exercise of power and control in the rhetorics of testing. In Hutta, A., Sajavaara, K., &

S. Takala (Eds), *Language Testing: new openings*. University of Jyvaskyla, Finland.

Spearman, C. (1904). “General intelligence,” objectively determined and measured. *American Journal of Psychology*, *15*, 201-293. http://dx.doi.org/10.2307/1412107

Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511732997.

Appendix A. Rating Questionnaire

	Questionnaire Items	**1** **Strongly Disagree**	**2** **Disagree**	**3** **No Idea**	**4** **Agree**	**5** **Strongly Agree**
	EFL Writing Courses					
1	The main purpose of writing courses is to develop students’ academic writing ability.					
2	I teach those genres of writing (e.g. expository) which are most related to academic writing.					
3	For undergraduates, teaching the basics of academic writing is the focus.					
4	I attend to the purpose, genre and audience in any writing course.					
5	In BA writing courses, low proficiency of the students causes teachers to ignore teaching some genres.					
6	EFL writing differs greatly from ESL writing.					
7	Context (native vs. non-native) affects the level of writing proficiency expected.					
8	Achieving native-like proficiency in writing is my goal in writing courses.					
	EFL writing scoring	**Strongly Disagree**	**Disagree**	**No Idea**	**Agree**	**Strongly Agree**
9	When scoring a writing text, I attend to different components of the text such as language, content, structure handwriting, style… one at a time.					
10	Scoring different aspects of a text in a separate way (analytic scoring) gives me more confidence when reporting results.					
11	I have my own way of analytic scoring (i.e. I do not use any existing analytic rating scale).					
12	For me, analytic scoring is frustrating and takes a lot of time.					
13	I look at the text and based on my own experience in rating, I give a total score.					
14	I look at the text and give a single general score based on a rating scale (holistic scoring).					
15	I think giving a score based on my impressions is quite trustable.					
16	I use a combination of holistic and analytic scoring in assessing writing.					
17	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).					
18	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**
19	International rating scales such as Jacobs, et al. (1981) directly guide my rating.					
20	All components of International rating scales are equally relevant to my rating.					

21	Rating scale plays a significant role in any assessment of writing.					
22	An explicit rating scale improves validity and reliability of my assessment.					
23	I have my own scoring criteria such as word choice, structure, spelling… but I don't consider them as a kind of analytic scoring.					
24	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).					
25	I think International scales are inappropriate as there are striking differences between the rating criteria in the international scales and those of mine.					
26	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).					
27	The function of writing assessment in the present Iranian undergraduate courses is diagnostic.					
28	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**
29	I assess students' compositions to provide them with a profile of their weaknesses and strengths in writing.					
30	It occurs that I have the same rating components as those of International scales but with different levels and descriptors that I define.					
31	Students are informed about my rating criteria early in the course.					
32	I think present international rating scales are quite suitable for scoring.					
33	In case of adapting an international rating scale, I redefine the level descriptors to adjust to my specific group of students.					
34	A local (e.g. Iranian) rating scale for writing assessment is needed to assure the validity of the scores.					
35	Particularities of any context would affect the rating components of any rating scale.					
36	As a rater, I am quite aware of different rating scales.					
37	Lack of a common rating scale would lead to bias, inconsistency and leniency/severity among the raters.					
38	Diversity of raters' criteria in writing assessment makes the construct of writing ability rather vague and elusive.					
39	The existence of a common rating scale would lead to a more fair writing assessment in ELT departments.					
40	My rating experience justifies the scores I assign out of my general impressions of the text.					

Appendix B. Interview scheme

1. Did you face any controversial item/s in the questionnaire? Elaborate on them please.
2. Do you use any special rating scale when rating a composition?
3. In brief, can you describe your typical rating?
4. Have you ever used any International rating scale such as Jacobs, et al. (1981) in your rating?
5. In case of using an International rating scale, how do you react to the level descriptors? Adapt? Adopt? Or just as a guide?
6. Is an International rating scale appropriate in Iranian EFL context?
7. Can you name some advantages and some disadvantages of International rating scales?