Rating Scales Revisited: EFL Writing Assessment Context of Iran under Scrutiny

BATOUL GHANBARI
English Department, Faculty of Foreign Languages, University of Isfahan, Iran

HOSSEIN BARATI
English Department, Faculty of Foreign Languages, University of Isfahan, Iran

AHMAD MOINZADEH
English Department, Faculty of Foreign Languages, University of Isfahan, Iran

Bio Data:

Batoul Ghanbari is a PhD candidate in TEFL at University of Isfahan, Iran. Her areas of interest include writing assessment, second language testing and sociolinguistics.

Dr. Hossein Barati is an assistant professor in TEFL at English Department, University of Isfahan, Iran. His areas of research include language testing & assessment and research methodology in second/foreign language acquisition.

Dr. Ahmad Moinzadeh is an assistant professor in TEFL at English Department, University of Isfahan, Iran. His major areas of interest center on linguistics and SLA research.

Abstract

Through addressing particular ideologies regarding language, meaning, level of proficiency and target writer and reader, rating criteria define and control the *what* and *how* of the assessment process. But a point which has been neglected or intentionally concealed due to concerns of practicality and the legitimacy of the native authority in setting assessment guidelines in EFL writing assessment contexts is the appropriateness of the scale**.** To raise attention to the current vague rating situation and consequently remedy the state, present study followed two lines of argument. First, drawing on the socio-cognitive framework of Weir (2005) for validating writing assessment, it is discussed that the important characteristic of scoring validity necessitates an appropriate choice of rating rubrics. Second, through posing a critical argument, deficiencies of the present practice of adopting rating scales are revealed and consequently it is discussed how assessment circles in native countries by setting rating standards control and dominate the whole process of writing assessment. To add more flesh to the argument, the ESL Composition Profile of Jacobs, et al. (1981) for its popularity in the Iranian EFL academic writing assessment is analyzed. A preliminary examination of the Profile shows that quite different underlying assumptions are involved. The study ends with a call to add a more local taste to the rating scales. To a large extent, developing a local rating scale that gives agency to the intricacies of Iranian EFL context in designing and developing the scoring criteria in writing assessment would be promising.

Keywords: writing assessment, academic writing, rating scale, validity, construct validity, ESL Composition Profile (Jacobs, et al., 1981).

Introduction

Within the past few decades, writing assessment has been a constant concern to the extent that any new publications on written composition have some references to the issues related on evaluating writing. Due to the ascending importance of writing among all sections of the present modern society that values written communication as an index of educational growth, pronouncing judgment on a piece of writing text has found a significant place (Gere, 1980).

However, assessing writing faces challenges on two major frontiers: on the one hand, program-level decisions regarding placement in different levels of a course or admission purposes necessitates a rigorous assessment plan, and on the other hand Pandora's Box of performance assessment reveals itself in the writing (Mc Namara, 1996) as there are still vague grounds in the articulation of a sound and explicit basis in scoring writing (Gere, 1980).The ability to make sound decisions about the writing ability of individual writers is the de facto function expected from the scoring procedures involved. Therefore, any malfunctioning in the writing assessment might pop up this basic but critical question in mind: do scoring procedures work correctly to accomplish their expected purpose in providing a sound appraisal of writers' writing ability?

Inspired by the above line of inquiry, the present study proceeds to give a second thought to the procedures of writing assessment. In this line, the venerable tradition of using rating scales in writing assessment is investigated. Upon contextualizing the concept of rating scale in its theoretical background and analyzing the value-laden nature of the scales involved, the writer proceeds to underscore the appropriateness of rating scales in safeguarding the validity of assessment outcomes provided through the scales. The discussion is touched at a deeper level when a critical appraisal of the assumptions behind rating scales is called for. The critical argument supports the claim for selecting appropriate rating scales in the writing assessment. Moreover, along with theoretical arguments and in order to give a more realistic taste to the points expressed, ESL Composition Profile (Jacobs, et al., 1981) as a commonly-used rating framework in ESL/EFL[1] writing assessment context of the country is analyzed to reveal any inconsistency between the assumptions of the original scale developers and the realities of the EFL writing assessment contexts such as Iran.

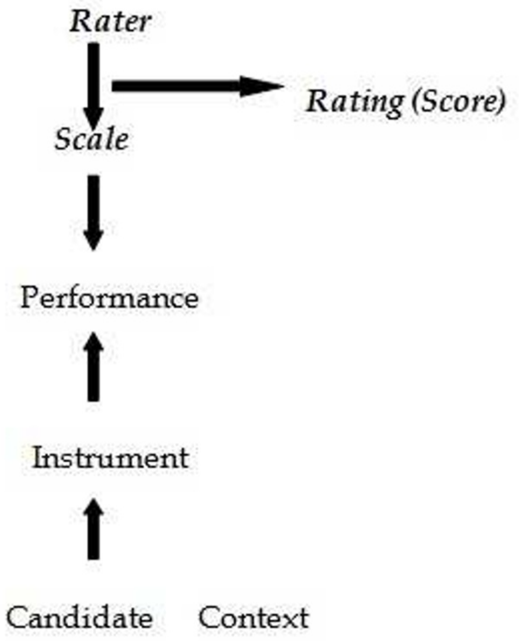
[1] English as a Second Language/ English as a Foreign Language

Overall, the study subscribes to the view of developing a rating scale which minimally addresses the particularities of the EFL context of Iranian writing assessment. This proposal which is strongly supported in the literature (McNamara, 1996; Norton, 2000; Ostovar& Hajmalek, 2010) is promising in both amending the important psychometric feature of construct validity of writing assessment and also in shifting the locus of control to the people who have the highest stake in any assessment enterprise, i.e. local stake-holders.

Rating Scales in Performance Assessment

In a much-quoted figure, McNamara (1996) schematically represents different factors that affect the final score given to a test-taker in a typical performance assessment context (Figure 1). While scoring in a traditional fixed-response assessment involves an interaction between just candidate and the instrument, in performance assessment there is some additional component which involves a rater or judge to assess a sample of performance through a scale or other kind of scoring schedule (Weigle, 2002).

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



This new interactive component between rater and rating scale-which mediates the scoring of the performance- has opened a new horizon of investigation for assessment specialists. In the words of McNamara (1996), we should seek information on the scale and the rater with the same rigor we did for the instrument and subject in the traditional assessment.

Inextricably, the conceptualization of rating scale as the de facto test construct (Norton, 2003) has created an unprecedented position for it in many discussions on performance assessment. For instance, Weigle (2002) in her discussion of validity of writing assessment considers gathering information on rating scales as an important evidential basis of any ongoing validation enterprise. Needless to say, the crucial role of rating scale in any assessment of performance deserves a more rigorous analysis. However, as McNamara (1996, p.182) desperately claims:

> 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.

Undoubtedly, rating scales and their attendant effects in performance assessment should be studied with a more scientific rigor. As the concern of the present study, the next part delineates the place of rating scales in writing assessment.

Rating Scale in Writing Assessment

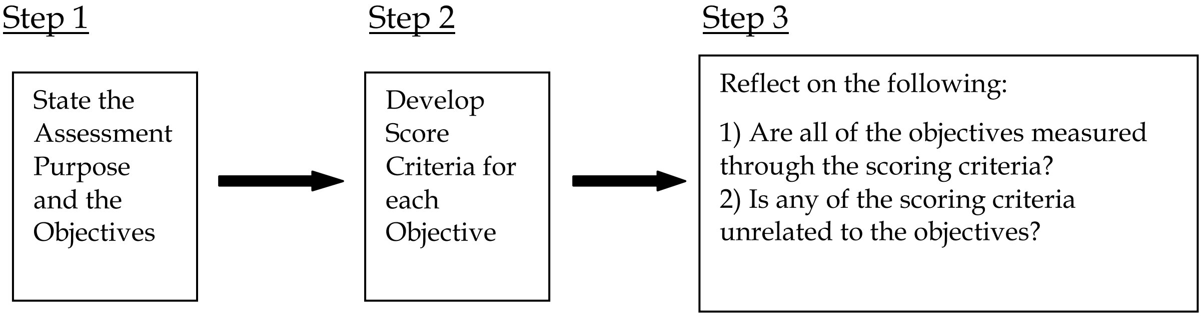
On a par with debates on rating scales in general performance assessment, the quality of writing assessment is to a great extent dependent upon the criteria used when assessing a piece of writing. The major argument on rating scales converges on the issue of construct validity in writing. The construct definition of writing as a fleeting and complex task is considered as an important step in any investigation of the validity. In fact, the issue of construct validity in writing must be tackled in at least three ways: First, the task must elicit the type of writing that we want to test. Second, the scoring criteria must take in to account those components of writing that are included in the definition of the construct and third, the raters must actually adhere to those criteria when scoring writing samples (Weigle, 2002).

When it comes to the second point, the issue of scoring presents itself in a challenging way. The main problem with writing assessment is that objective scoring as implemented in multiple-choice tests cannot be used. Therefore, various raters might assess the same text differently and hence come to divergent results (i.e. scores). As an example, the impressionistic and individualistic approach to essay writing in the EFL context of Iran have faced several problems (Nemati, 2007). First, since the scoring criteria are unique to each individual rater, large discrepancies in essay scores often occur. The large amount of rater variability then undermines the validity of the writing test as a measure of students' EFL writing abilities. Also, lack of explicit and detailed rating scales make it difficult to resolve the score discrepancies. Second, teachers, particularly those who are novice in the local context make it difficult to infer and implement these assessment criteria consistently (Barkaoui, 2007). Even experienced teachers might show great variability in weighting different aspects of a written text (ibid). As a result, the attempt to improve subjective scoring in assessing writing has been a point of investigation among many field practitioners (Barkaoui, 2007; Cumming, 1990; Farzanehnejad, 1992; Knoch, 2009; Hamp-Lyons, 1996; Nemati, 2007; Zomorodian, 1998). One way out of the problem of fluctuations in the raters' ideas in relation to the same written text has been suggested as using a scoring framework (Bachman, 1990; Mc Namara, 1996). Scoring frameworks or better say, rating scales, have thus been the focus of the writing assessors so that many different scales have been introduced; each holding particular assumptions about *how* of the assessment procedure. Hence, rating scales fulfill central roles in assessing writing to the extent that the validity of the results relies on the rating procedure adopted (Weir, 1990).

A similar argument was posed by Moskal and Leydens (2000). In an attempt to illustrate the process of developing a rating rubric, they highlighted the significance of rating criteria as the instantiation of assessment objectives in the process of assessment (Figure 2). As shown below, rating scales mediate the relation between the theoretical dimensions of assessment, i.e. objectives and purposes and the final scoring emerged.

In counting the importance of rating scales, the writers proceed to say that if some objectives are not represented in the scoring scale or if some of the scoring criteria are not related to the objectives, then, the appropriateness of the whole assessment and the rubric is dubious. Undoubtedly, rating scales and the way they are conceptualized, designed and developed affect the outcome of the assessment.

Figure 2. Evaluating the appropriateness of scoring categories to a stated purpose (adapted from Moskal& Leydens, 2000)


Step 1
Step 2
Step 3
State the Assessment Purpose and the Objectives
Develop Score Criteria for each Objective
Reflect on the following:
1) Are all of the objectives measured through the scoring criteria?
2) Is any of the scoring criteria unrelated to the objectives?

In the same vein, 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.*

Moreover, another motivation to focus on rating scales is a renewed interest in diagnostic assessment of writing (Knotch, 2011). With rating scale as its integral component, a diagnostic mode of assessment has encouraged writing assessment specialists to embark on a more explicit and detailed development of rating scale to get access to a more detailed and objective profile of the writers strengths and weakness in writing.

As can be inferred from the above, there is a consensus over the importance of rating scales in writing assessment. However, rating scales are not just neutral psychometric instruments used to provide merely assessment results. They bear particular assumptions about the test, test-taker, test-user, reader, writer, etc. (Weigle, 2002). As a matter of fact, the process of developing a rating scale is deeply involved with many considerations which are determined and controlled by the assumptions of the scale constructors. To get a more nuanced understanding of the considerations involved in any scale development, the five practical steps that

Weigle (2002) introduces are explained. These considerations which need to be weighted carefully for a rating scale to be valid (Knotch, 2011), are briefly described in the next section.

Steps in Rating Scale Development

Weigle (2002, pp. 122-125) presents some steps that should be taken in to account in the process of scale development. As will be discussed below, the steps indicate that rating scales are founded on the assumptions and concerns of their constructors in their specific context and to serve their particular goals. In sum, they are developed to target test-takers with particular features who aim to accomplish specific goals in specific contexts.

- **What type of rating scale should be used?** Decision about the type of scale is the concern of scale developer at this stage. The common types of analytic, holistic or primary trait methods are options that scale developer should select owing to the general applicability and the concerns of the assessment.

- **Who is going to use the rating scale or what is the purpose of the scale?** The context and the purpose of the test necessitate the appropriateness of the format of the scale, the theoretical orientation of the description and the formulation of the definitions (Knotch, 2011). In the words of Alderson (1991), the purpose of the test affects the formulations of the descriptors.

- **What aspects of writing are most important and how will they be divided up?** The scale developer needs to decide over the rating criteria to use a basis for assessment. Therefore, the criteria used are reflections of the scale developer's concerns.
- **What will the descriptors look like and how many scoring levels will be used?**

The range of performances that can be expected and also what the test results will be used for, will determine the format of the descriptors. To decide on how band levels should be distinguished from each other as well as the types of descriptors to be used will be decided by the scale developers.

- **How will scores be reported?** The very use of the test scores will determine the manner the scores will be reported. Moreover, it affects the decisions over whether different categories on the scale should be weighted.

The above five explicit procedures in rating scale development vividly shows the influence of several important parties in constructing rating scale and consequently the validity of the assessment. Therefore, it provides rating scales with an important role in the writing assessment.

This theoretical significance of rating scales in the writing assessment context should be investigated having in mind the chaotic situation in the construction of rating scales. Many scholars (Brindley, 1998; Fulcher, 2003; Knotch, 2011; McNamara, 1996; Turner, 2000; Upshur and Turner, 1995) have pointed out the dominance of an atheoretical view in rating scale development. For example, Fulcher (2003) points out that many rating scales are developed based on intuition, which means that a group of teachers or language testers develop the scale, possibly by adapting an existing one. The outcome is that the issue of an explicit rating scale is taken for granted in different contexts of writing assessment. Different raters might either rely on their own impressionistic judgment of writing tasks or in case of using any scale; it might be a mere adoption of some existing scales in the literature.

Regarding the chaotic rating situation in writing, many scholars have attempted to improve the subjective assessment of writing (Fulcher, 1996; Fulcher et al., 2011; Knotch, 2007; Maftoon & Akef, 2010; North & Schneider, 1998; Upshur & Turner, 2002). These lines of queries aim to create more objectivity in the writing assessment. However, the appropriatness of rating scales with regard to the context of their use has been given a cursory attention.

Although writing assessment has faced some basic challenges over the use of rating scales, the move to a validated rating scale that acts appropriately in its context should not be hidden behind the concerns of practicality or simply the illusion of a universal writing ability encouraged in many theoretically-developed rating scales. In order to problematize the rating situation, we draw upon two perspectives and reveal the deficiencies in the following two sections.

Weir’s Socio-cognitive Framework

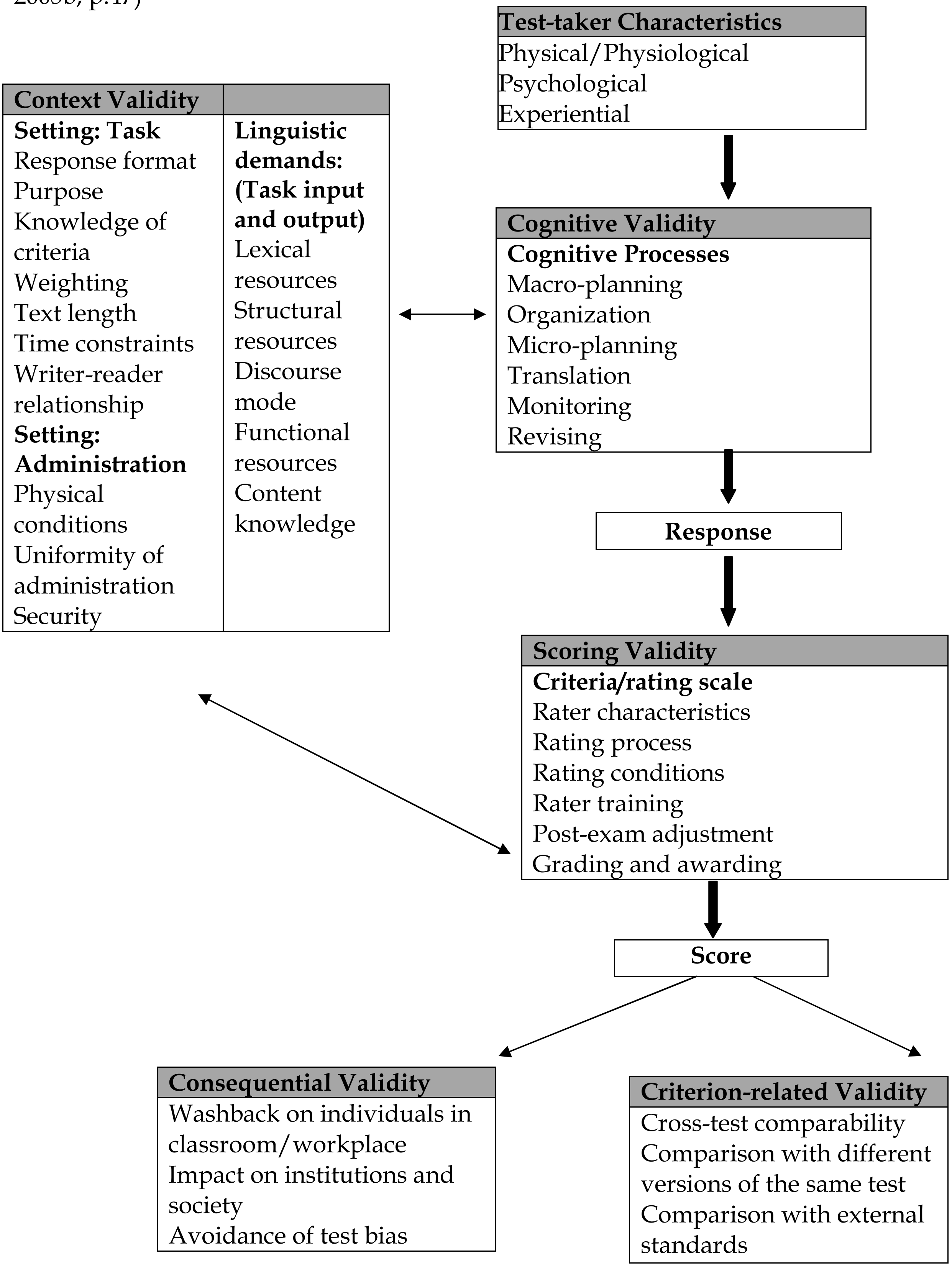
Weir (2005) presented an explicit framework for validation in language testing (Figure 3). He showed that at the heart of any language testing activity, there is a tripartite relationship between three crucial components:

the test-takers’ cognitive abilities
the context in which the task is performed, and
the scoring process

These three internal dimensions of any language test referred to as cognitive validity, context validity and scoring validity respectively constitute an innovative conceptualization of construct validity which bears sound theoretical and direct practical relevance for language testers. The symbiotic relationship between the contextual parameters laid out in the task and the cognitive processing involved in task performance reminds us that language use –and also language assessment-is both a socially situated and a cognitively processed phenomenon. The socio-cognitive framework as a unified approach to establishing the overall validity of a test is intended to depict how the various validity components (the different types of validity evidence) fit together both temporally and conceptually.

As can be inferred from the pictorial representation above, achieving scoring validity is highly important since if we cannot depend on the ratings done, it matters little that the tasks we develop are potentially valid in terms of both cognitive and contextual parameters (Shaw& Weir, 2007).

Figure 3. Framework for conceptualizing writing performance (adapted from Weir 2005b, p.47)



The first scoring validity parameter is that of the criteria and type of rating scale (Figure 4). In fact, the choice of appropriate rating criteria and the consistent application of rating scales by trained examiners are regarded as key factors in the valid assessment of second language performance (Alderson, Clapham and Wall 1995, Bachman and Palmer 1996, McNamara 1996).

Figure 4. Aspects of scoring validity (adapted from Weir, 2005)

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

Appropriateness of the rating criteria refers to a sequence of activities in which the test developer first establishes appropriate criteria based on the purpose of the assessment and the construct being measured and then determines levels of performance in relation to these criteria. Faulty criteria or scale along with other components of scoring validity shown in Figure 4 can all lead to a reduction in scoring validity and consequently to the risk of construct irrelevant variance.

Purpose of the assessment which is defined and delineated in every particular context and the construct being assessed are two yardsticks which determine the appropriateness of any rating scale (Shaw& Weir, 2007). Overall, the socio-cognitive framework of Weir (2005) as an explicit descriptive metaphor clearly demonstrates the value of an appropriate rating scale. Thus, any writing assessment context that ignores this aspect of scoring validity would lead to the scores that are not valid indicators of ability. The model of Weir (2005) proposes to take into account all the factors that have a bearing over the triangular relationship among cognitive, context and scoring validity. In this respect, all subcategories in each validity type can be investigated to see how it works in relation to the other. Therefore, as a monitoring checklist, this model specifies the place of appropriate rating scales in the writing assessment.

A Critical Appraisal of Rating Scale

The rating scale tradition originated in the FSI[2] test in the 1950s and 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 little empirical validation of them has been attempted (McNamara, 1996). For example, a series of research studies looking at the performance of native speakers on a reading test where performance was reported in FSI-type band reveals that the assumptions lying behind the descriptors in the scale cannot be sustained by empirical evidence. More research effort must go into the validation of such rating scales, which are

[2] Foreign Service Institute

central to the construct validity of the instruments with which they are associated (ibid).

For instance, McNamara (1996) in a review of studies that investigated the performance of native speakers on an EAP test looked critically at the assumptions of rating scales. He showed that idealization of native speaker performance is frequent in such scales which had 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. In addition, studies by some scholars (Bachman, 1990; Hughes, 1980; Oller and Conrad, 1971) showed that the idealization of native speaker cannot be sustained on the grounds that performance on the test involves factors other than straight second language proficiency, and these factors are included in the assessment, then we may expect there to be an overlap in the performance of native and non-native speakers; and the performance of native speakers will be highly variable.

Therefore, it goes without saying that rating scales are not just neutral practical artifacts; rather, they are used to make decisions about the life chances of individuals. Hence, they have some certain assumptions about language, meaning, writers and readers. Recent debates on ethics and fairness in language testing also demand us to critically analyze rating scales (Hamp-Lyons 1997; Norton 2000; Shohamy 1993) to uncover the hidden ideologies involved. However, a critical analysis of rating scales has been given passing attention (McNamara, 1996) mostly due to a common trust in so-called psychometrically sound instruments designed and developed by the most eligible authorities in the field, i.e. native speaker scale developers. The dominance of native authorities has encouraged a prescriptive tradition of passive acceptance by many “scale consumers” who openly accept the dominance (Odell & Cooper, 1980).

Referring to our arguments on the place and structure of any rating scale, it is evident that the issue of native speaker idealization and legitimacy in rating scales is confined to vague results in achievement; rather it shows a monopoly of native orthodoxy in the way assessment proceeds in nonnative countries. The impressionistic writing assessment which recognizes no need for the inclusion of any rating scale on the one hand and a strong reliance on the rating scales without adapting the levels or descriptors to the specific context are two seemingly opposing forces that end to the same result: the maintenance of a native dominance in writing assessment context.

In line with the above argument and in order to critically analyze a writing assessment procedure, the ESL Composition Profile (Jacobs, et al., 1981) as a well-known rating scale would be examined to reveal any implicit assumptions and/or hidden values involved in the categories of the scale.

ESL Composition Profile

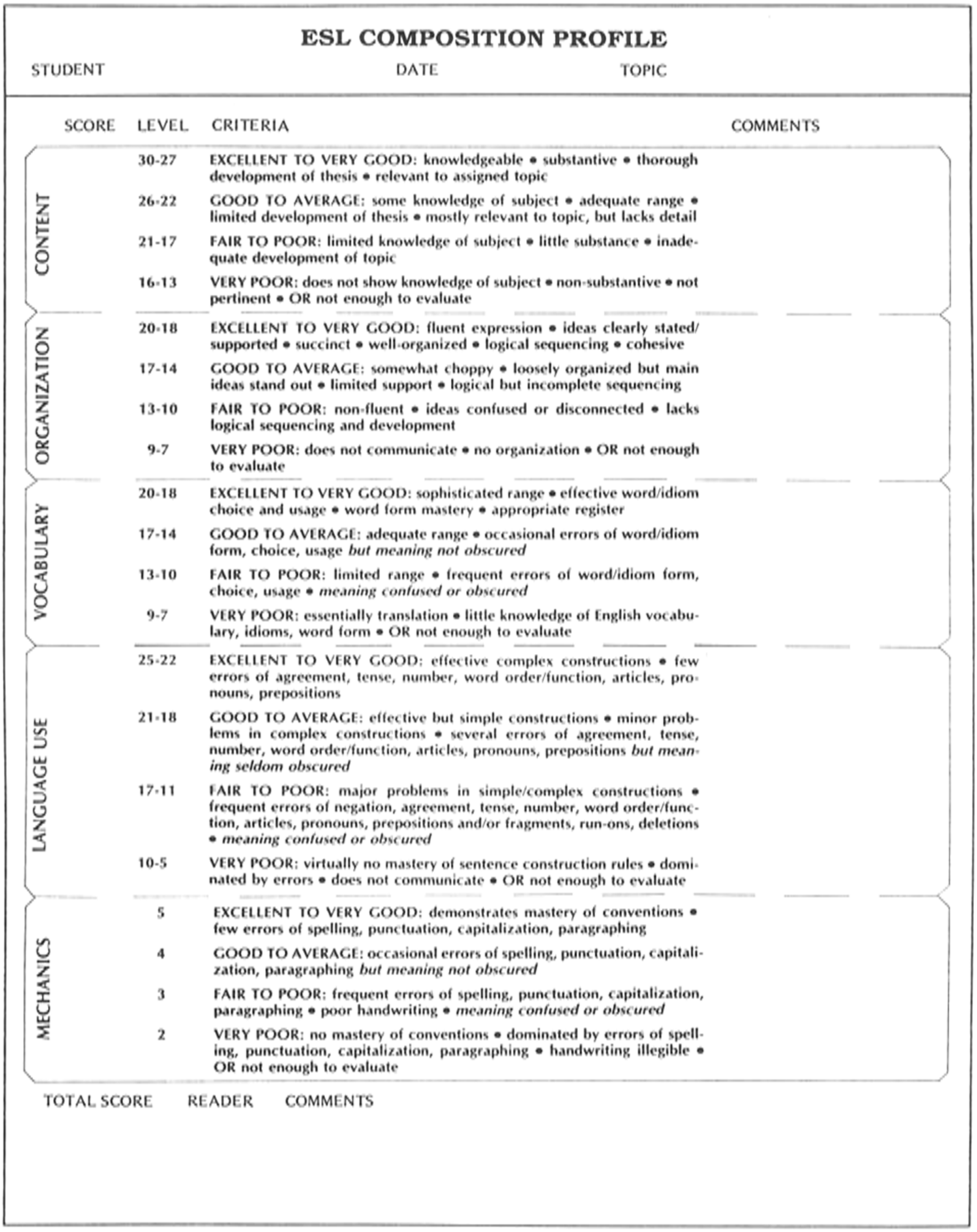
The very basic assumption that one can identify with the analytic scale of ESL Composition Profile (Figure 5) is that one can and must identify distinct qualities which one looks for in “good" writing. Haswell (2005, p.2), calls the ESL Composition Profile by Jacobs, Zingraf, Wormuth, Hartfiel and Hughey (1981) as a tool that “it is no different than dozens of similar guides by which raters have decided, and continue to decide, the academic fate of thousands upon thousands of second language students”. He proceeds to count the three main features of the Profile:

1. A limited number of basic criteria or main traits (e.g., content, organization, vocabulary, language use, and mechanics).

2. A fitting of each trait into a proficiency scale, the levels of which are also small in number and usually homologous or corresponding (e.g., 1, 2, 3, or 4 for each trait).

3. breakdown of each trait into sub-traits, which are also small in number and homologous or corresponding (See Figure 5 below).

Figure 5. ESL Composition Profile (Jacobs, et al., 1981)



ESL COMPOSITION PROFILE

STUDENT DATE TOPIC

	SCORE	LEVEL	CRITERIA	COMMENTS
CONTENT		30-27	EXCELLENT TO VERY GOOD: knowledgeable • substantive • thorough development of thesis • relevant to assigned topic	
		26-22	GOOD TO AVERAGE: some knowledge of subject • adequate range • limited development of thesis • mostly relevant to topic, but lacks detail	
		21-17	FAIR TO POOR: limited knowledge of subject • little substance • inadequate development of topic	
		16-13	VERY POOR: does not show knowledge of subject • non-substantive • not pertinent • OR not enough to evaluate	
ORGANIZATION		20-18	EXCELLENT TO VERY GOOD: fluent expression • ideas clearly stated/ supported • succinct • well-organized • logical sequencing • cohesive	
		17-14	GOOD TO AVERAGE: somewhat choppy • loosely organized but main ideas stand out • limited support • logical but incomplete sequencing	
		13-10	FAIR TO POOR: non-fluent • ideas confused or disconnected • lacks logical sequencing and development	
		9-7	VERY POOR: does not communicate • no organization • OR not enough to evaluate	
VOCABULARY		20-18	EXCELLENT TO VERY GOOD: sophisticated range • effective word/idiom choice and usage • word form mastery • appropriate register	
		17-14	GOOD TO AVERAGE: adequate range • occasional errors of word/idiom form, choice, usage *but meaning not obscured*	
		13-10	FAIR TO POOR: limited range • frequent errors of word/idiom form, choice, usage • *meaning confused or obscured*	
		9-7	VERY POOR: essentially translation • little knowledge of English vocabulary, idioms, word form • OR not enough to evaluate	
LANGUAGE USE		25-22	EXCELLENT TO VERY GOOD: effective complex constructions • few errors of agreement, tense, number, word order/function, articles, pronouns, prepositions	
		21-18	GOOD TO AVERAGE: effective but simple constructions • minor problems in complex constructions • several errors of agreement, tense, number, word order/function, articles, pronouns, prepositions *but meaning seldom obscured*	
		17-11	FAIR TO POOR: major problems in simple/complex constructions • frequent errors of negation, agreement, tense, number, word order/function, articles, pronouns, prepositions and/or fragments, run-ons, deletions • *meaning confused or obscured*	
		10-5	VERY POOR: virtually no mastery of sentence construction rules • dominated by errors • does not communicate • OR not enough to evaluate	
MECHANICS		5	EXCELLENT TO VERY GOOD: demonstrates mastery of conventions • few errors of spelling, punctuation, capitalization, paragraphing	
		4	GOOD TO AVERAGE: occasional errors of spelling, punctuation, capitalization, paragraphing *but meaning not obscured*	
		3	FAIR TO POOR: frequent errors of spelling, punctuation, capitalization, paragraphing • poor handwriting • *meaning confused or obscured*	
		2	VERY POOR: no mastery of conventions • dominated by errors of spelling, punctuation, capitalization, paragraphing • handwriting illegible • OR not enough to evaluate	

TOTAL SCORE READER COMMENTS

Upon investigating the Profile, some "invisible" assumptions become known. These points are discussed below.

- **Troubling and mysterious origin**

ESL Composition Profile has been a popular L2 essay rating method since its introduction in 1981. This wide application gives this legitimacy that the scale has been developed by L2 researchers and based on the compositions of L2 writers. The validity of scale is accepted through a wide lore among raters and composition researchers alike. However, Haswell (2005) puts the question in this way that why the form is in its existing form. In other words, why content, organization, vocabulary, language use, and mechanics and not creativity, logic, suspense, tradition, shock-appeal, humor, cleverness—and the second list could go on.

When the history of the Profile is queried, it is surprising to find that one of the widely used rating scales is based on such shaky and invalid grounds. The origin of these five traits of the Profile refers to 1958 when the scale was derived from the grades and marginal comments written on student homework. The graders and commenters included a few non-ESL teachers who had no TESOL experience. The writers were first-year students at Cornell, Middlebury College, and the University of Pennsylvania, probably none of them second language students. Later, three researchers at Educational Testing Service (ETS) factored the commentary, passed the factoring on to a colleague of theirs at ETS, Paul Angelis, who passed it on to the authors of the "ESL Composition Profile" (Jacobs et al., 1981). In the meantime, one of the original five factors, flavor, got dropped, and another, wording, got divided into vocabulary and language use, but no new factors were added (Haswell, 2005).

It is worth noting that how a long-established rating scale such as ESL Composition profile turns out to be armless when facing a question that just re-orients the scale from another perspective. As revealed, the non-ESL origin of the Profile alone might threaten it as an inappropriate scale in judging the writing performance of the ESL learners.

- **Homology of the sub-traits**

Imagine content as a trait on the scale. It is divided to four sub-traits of knowledge of the topic, substance, development of the topic, and relevance which are associated with levels. The hidden feature of homology prevents for example a writer who has a "limited knowledge of the topic" to relate to the topic in a relevant way. Although significant, this feature of homology has been little mentioned by composition researchers (Haswell, 2005).

- **Holistic categorization**

For a long time in its history, ESL Composition Profile has been named under the wide category of analytic scales. By weighting content as the most important component and mechanics the lowest, it befits writers who show uneven accomplishment in different writing scales. This characteristic has been lauded as an advantage of analytic scoring compared to holistic for different purposes of research, placement, rater training and program validation, those needed to defend commercial testing or research studies. But a closer look reveals that the Profile is no different from holistic scoring. It asks the rater to conduct holistic scoring five times. In fact, both of the scoring (holistic and profile scoring) apply a similar categorization frame.

- **Uneven weighting**

Tedick (2002) contends that the weighting scheme of any scale depends on factors like the task, purpose and learners' level. Making decision on the weighting of each sub-construct is immensely important to the extent that some scales such as TEEP[3] (Weir, 1983) have followed an equal-weight scheme. In the ESL Composition Profile (Jacobs, et al., 1981), different weights are assigned to each subscale. Content has the highest weight (30% of the total score). Moderate weights are given to language use, organization and vocabulary (25%, 20% and 20% of the total mark, respectively), while mechanic receives the lowest (only 5% of the total mark).

Considering the troubling origin of the scale mentioned above, this kind of differential weighting is seriously under question. The Profile is of low validity due to its initial non-ESL motives, therefore; it goes without saying that composition of ESL writers cannot be rated according to some weighting out of statistical procedures. Moreover, even imagining an ESL base for the Profile, how it can justify its pattern of weighting in an EFL context such as Iran where different objectives and purposes are involved.

As an example, mechanics is rated as the lowest on the Profile, while attention to this aspect of EFL composition is considerable especially at primary levels of writing proficiency. In the same vein, content might not be a primary concern of EFL composition raters as they seek for a text come out of well-organized bases which involve considerations of language use and organization in the terms of ESL Composition Profile.

As went above, a preliminary critical scrutiny of ESL Composition Profile (Jacobs, et al., 1981) revealed that the scale is not merely a neutral rating instrument; rather, several ideological assumptions dominate and direct the scale.

Conclusion: Towards a Local Rating Scale

As a rebel to the mainstream tradition in language testing, the present study set out to re-analyze rating scales in writing assessment with regard to their original function, i.e. providing a sound picture of the writers' ability. In this regard, the discussion followed two strands of arguments. First, socio-cognitive framework of Weir (2005) reminded the importance of an appropriate selection of rating criteria and henceforth questioned the validity of any scale developed and devised on unknown grounds. Next, the critical argument provoked some serious concerns over the validity of rating scales. The discussion showed that the multi-layered ideological structure of rating scales caused reservations for their undisputed application. The argument revealed that many rating scales in use were known to ultimately derive from the FSI scale, developed originally 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. In the remainder of the study, ESL Composition Profile (Jacobs, et al., 1981) was investigated to reveal the seamy side of the scale. A critical look at the scale identified several value-laden assumptions built into the structure of the scale. The kind of assumptions that had serious consequences for a fair assessment that is envisioned as the ultimate goal in educational assessment contexts.

When it comes to the particular writing assessment context of Iran, the problem of a vague rating situation where writing assessment is considered as a

[3] Tests of English for Educational Purposes

quite individualistic and impressionistic task and there exists no serious concern over the use of an explicit rating scale complicates the process of scoring writing. Even in case of using any rating scale, the problems counted in the present study widely exist. Therefore, any recommendation for infusing objectivity in rating practices in the Iranian writing assessment context should be mediated.

In conclusion, this article reminds that in addition to the long-term obligation of continually examining and testing the evaluation procedures and the assumptions that underlie them, a local rating scale as it takes into account the particularities of each assessment context would lead to more valid outcomes. Such a proposal for the development of a context-based rating scale is justified and supported by both validity argument of Weir (2005) and critical discussions in the field. In this way, a local definition of rating criteria by the local raters would depict a promising perspective for this important aspect of writing assessment.

References

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

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing*, *12*(2), 86-107.

Brindley,G.(1998). Describing language development? Rating scales and SLA. In L. Bachman, & A.Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 112-140).Cambridge: Cambridge University Press.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7*(1), 31-51.

Farzanehnejad, A. R. (1992). *A new objective measure for calculating EFL writing tasks* (Unpublished M.A. thesis). University of Tehran,Tehran, Iran.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

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

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238.

Gere, R. A. (1980). Written composition: Toward a theory of evaluation. *College English, 42*(1), 44-58.

Hamp-lyons, L. (1997) Ethics in language testing. In C. Claham, & D. Corson (Eds.), *Language testing and assessment: The encyclopedia of language and education* (Vol 7). 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 P. Matsuda, & T. Silva (Eds.), *Second language writing research: Perspectives on the process of knowledge construction* (pp. 105-120). Erlbaum.

Hughes, A. (1989). Testing for language teachers. In T. McNamara (Ed.), *Measuring second language performance*. Harlow: Longman.

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.

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,* 81-96.

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

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.

Maftoon, P., & Akef, K. (2010). Developing rating scale descriptors for assessing the stages of writing process: The constructs underlying students' writing performances. *Journal of Language and Translation, 1*(1), 1-17.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing*, *15*(2), 217–263.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*(1), 49–70.

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

Nemati, M. (2007). To be or not to be: A search for new objective criteria to evaluate EFL compositions. *Pazhuhesh-e Zabanha-ye Kahreji, 32,* 175-186.

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. (2000). Writing assessment: Language, meaning, and marking memoranda. In A. Kunnan (Ed.), *Fairness and validation in language assessment* (pp. 20-29). New York: Cambridge University Press.

Odell, L., & Cooper, C. (1980).Procedures for evaluating writing: Assumptions and needed research, *College English*, *42*(1), 35-43.

Oller, J. W., & Conard, C. A. (1971). The close technique and ESL proficiency. In T. Mc Namara (1996). *Measuring second language performance*. Harlow: Longman.

Ostovar, F., & Hajmalek, M. (2010). *Writing assessment: Rating rubrics as a principle of scoring validity.* Paper presented at the fifth conference on issues in English language teaching in Iran (IELTI-5),University of Tehran, Iran.

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

Shohamy, E. (1993). The exercise of power and control in the rhetorics of testing. In A. Hutta, K. Sajavaara, & S. Takala (Eds.), *Language testing: New openings*. University of Jyvaskyla, Finland.

Tedick, D. J. (2002). Proficiency-oriented language instruction and assessment: Standards, philosophies, and considerations for assessment. In Minnesota Articulation Project, D. J. Tedick (Ed.), *Proficiency-oriented language instruction and assessment: A curriculum handbook for teachers*. CARLA Working Paper Series. Minneapolis, MN: University of Minnesota, The Center for Advanced Research on Language Acquisition. Retrieved from http://www.carla.umn.edu/articulation/polia/pdf_files/standards.pdf.

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3–12.

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation*. Great Britain: Palgrave.

Weir, C. J. (1990). *Communicative language testing*. Englewood Cliffs, NJ: Prentice Hall.

Weir, C. J. (1983). *Identifying the language problems of overseas students in a tertiary education in the United Kingdom* (Unpublished doctoral dissertation). University of London, UK.

Zomorodian, M. (1998). *Iranian EFL teachers' and students' assessment of the student essays* (Unpublished M.A thesis). Iran University of Science and Technology,Tehran, Iran.