Shiraz University

International Branch

Department of Foreign Languages and Linguistics

Judgments of IELTS Writing Task II by Non-native and Native English Speaking Teacher Raters: An Outlook on Inter-rating Variability

By

Mohammad Farri

Supervised

By

Dr A. Ahmadi

September 2013

Abstract

Judgments of IELTS Writing Task II by Non-native and Native English Speaking Teacher Raters :An Outlook on Inter-rating Variability

BY

Mohammad Farri

The purpose of this study was to launch a thorough investigation concerning the possibility of differing orientations to the writing proficiency construct by native and non-native English speaking teacher raters. It mainly revolved around the International English Language Testing System (IELTS) that is widely administered and employed as a measure of general proficiency in English. While in Iran the performance on this test is regularly assessed by trained non-native English speaking (NNES) raters, the question has arisen as to whether the standards they apply in judging writing performance are comparable to those of native English speakers (NES). To answer this question, the researcher compared the samples scored by 40 NES raters (20 males+ 20 females) with the same 40 samples that were rated by NNES raters, 20 males (10 experienced + 10 inexperienced) + 20 females (10 experienced + 10 inexperienced) on the basis of 4 rating criteria, namely Task Response (TR), Coherence and Cohesion (CC), Lexical Resources (LR) and Grammatical Range and Accuracy (GRA). The results demonstrated that with regard to LR and GRA male experienced NNES raters were significantly different from their counterparts in NES group. Female experienced NNES raters were also significantly different from female NES raters on the basis of LR, CC, LR and GRA. Male NES raters were also significantly different from experienced female NES raters with regard to LR criterion. The experienced NNES males overall were closer to NES raters and experienced female NNES raters had the biggest difference. In this study it was also discovered that concerning CC, GRA, and Holistic scores male and female NNES raters were significantly different based on the factor of experience while the same difference was not found to be significant on the basis of TR and LR. It was also revealed that gender did not play any significant role between male and female NNES raters with regard to TR, CC, LR, GRA, and holistic scores.

Key words: Writing task, IELTS, Native and Non-native raters and Inter-rater variability

چکیده

ارزیابی فعالیت نوشتاری دو آیلس توسط معلمان بومی و غیر بومی انگلیسی: بررسی تفاوتهای ارزیابی بین گروهی

به کوشش محمد فری

هدف اصلی این تحقیق بررسی کامل و دقیق وجود گرایشهای متفاوت در نحوه ارزیابی مفهوم مهارت نوشتاری توسط مصححین بومی و غیر بومی می باشد. تمرکز این تحقیق روی آزمون بین المللی آیلس که هدف آن ارزیابی مهارت عمومی داوطلبان می باشد است. این آزمون در حالی در کشور ایران برگزار می شود که ارزیابی عملکرد داوطلبین آنرا افرادی به عهده دارند که خود غیر بومی بوده و صرفا برای آشنایی با نحوه ارزیابی این آزمون آموزش دیده اند یا با استناد به تجربه شخصی این کار را انجام می دهند. حال این پرسش ممکن است همواره به ذهن افراد متخصص این فن خطور کند که معیارهای ارزیابی مورد استفاده مصححین غیر بومی چقدر با معیارهای ارزیابی مورد استفاده مصححین بومی زبان انگلیسی وجه مشترک دارند. برای پاسخ به این پرسش پژوهشگر این تحقیق نمونه های تصحیح شده توسط چهل نفر از مصححین بومی (بیست مرد و بیست زن) را با همان نمونه های تصحیح شده توسط چهل مصحح غیر بومی که بیست نفر آنها مرد (ده نفر با تجربه و ده نفر بی تجربه) و بیست نفر دیگر زن (ده نفر با تجربه و ده نفر بی تجربه) بودند مقایسه کرد. در این مقایسه معیارهای نمره دهی به ترتیب عبارت از توانایی در و پاسخ درست به پرسش، انسجام و یکپارچگی در مهارت نوشتاری، توانایی در استفاده از کلمات مناسب متن و گستره دقت دستوری بودند. بر اساس نتایج حاصله با توجه به دو معیار توانایی در استفاده از کلمات مناسب متن و گستره دقت دستوری مصححین مرد غیر بومی با همتای خود در گروه مصححین بومی تفاوت معنی داری را نشان دادند. مصححین غیر بومی زن با تجربه همچنین با همتای بومی خود بر اساس تمامی چهار مورد معیارهای نمره دهی ذکر شده تفاوتهای معنی داری را تجربه کردند. در یک مقایسه دیگر مصححین زن غیر بومی با تجربه بر اساس معیار نمره دهی توانایی در استفاده از کلمات مناسب متن با مصححین مرد بومی تفاوت معنی داری را ترسیم نمودند. در کل مصححین مرد با تجربه بومی به مصححین بومی نزدیکتر بوده و در همین مقایسه مصححین زن با تجربه غیر بومی بیشترین تفاوت معنی دار را داشتند. در این تحقیق همچنین مشخص شد که بر اساس معیار تجربه تفاوت معنی داری بین مصححین زن و مرد غیر بومی دررابطه با معیارهای انسجام و یکپارچگی در مهارت نوشتاری ، گستره دقت دستوری و نمره دهی کلی وجود داشت. در صورتیکه در رابطه با دو معیار توانایی در فهم و دادن پاسخ درست به پرسش و توانایی در استفاده از کلمات مناسب متن هیچ تفاوت معنی داری بین مصححین بر اساس عامل تجربه وجود نداشت. همچنین در پایان مشخص شد که عامل عامل جنسیت هیچ تفاوت معنی داری را بین مصححین زن و مرد غیر بومی در رابطه با چهار معیار نمره دهی فوق الذکر و همچنین نمره دهی کلی ایجاد نمی کند.

کلمات اصلي: **فعالیت نوشتاری, آیلس, معلمان بومی و غیر بومی و ارزیابی بین گروهی**

Table of contents

Lists of Figures

Content **Page**

List of Tables

Content **Page**

List of Abbreviations

CAE = Certificate in Advanced English

CPE = Certificate of Proficiency in English

EFL = English as a Foreign Language

ESL = English as a Second Language

ESP = English for Specific Purposes

FCE = First Certificate in English

IELTS = International English Language Testing System

NES = Native English Speaker

NNES = Non- Native English Speaker

TOEFL = Test of English as a Foreign Language

UCLES = University of Cambridge Local Examinations Syndicate

CHAPTER ONE

Introduction

1.1 Preliminaries

IELTS (International English Language Testing System) is generally considered as a proficiency test in which both receptive skills namely, Listening and Reading and productive skills, namely Speaking and Writing are measured. This test is administered in most countries around the world. Having had special aims in their minds, applicants for this test can take either or both academic and general modules.

General module is highly recommended for vocational purposes while the academic module is mainly required for those prospective applicants who want to peruse their education. Academic Writing proficiency in IELTS is thoroughly evaluated on the basis of two tasks. IELTS Task one is generally referred to as information transfer task since the testees are required to extract the explicit information from a visual, namely a table, graph, pie chart, or flow chart. The second part or task in the writing section of the IELTS exam highly requires the testees to present their own views concerning a specific topic or given statement. The time allocation for the first task is 20 minutes while for the second task is 40. In the final stage of the writing assessment more points will be assigned to the second task since the minimum number of words required is 250. The second task can be presented in two forms namely, account or argument. In task two the recommended criteria are Task Response (TR), Coherence and cohesion (CC), Lexical Response (LR) and Grammatical Range & Accuracy (GRA).

On the basis of testees' written performance their grades are presented from 0 to 9. In IELTS test, there is no pass or fail distinction and the applicants are highly recommended to decide on their academic or vocational requirements.

1.2. Rater variables

There exist some variables that exert their direct influence on assigning special scores to scripts. When raters are engaged in scoring script, they usually bring along with them the personal experiences and values that have been accumulated during their professional lives. Here the impact of providing raters with the related trainings cannot be totally ignored since the main philosophy behind their being administered, would be providing the raters with the related techniques to come up with a sort of harmony between both experienced and inexperienced raters.

On the basis of the available literature, there are three sources of rater variability in linguistic assessment especially writing. According to Mc Namara (1996) the first source of variability associated with raters can be the candidate him/herself and his or her related abilities to perform certain tasks. The second source is the nature of the task itself and the possibility thereof to provide the candidate with two alternatives and last but not least, the raters themselves can be the source of existing variability.

Mc Namara (1996) believes that raters regardless of being experienced or not may demonstrate different levels of discrepancies with their counterparts on the bases of different scenarios. Firstly they may differ from each other on the very concept of leniency and their amount of tendency shown towards it. Secondly raters might demonstrate a certain level of prejudice to certain tasks or even testees. Thirdly raters might also not be very consistent in their rating tendencies or behaviours and finally they may show the discrepancies based on their interpretations of the employed rating scales.

1.2.1 Verbal protocol

1.2.1.1 What is verbal protocol?

The involvement of test takers' cognitive process while doing a special task or coming up with an answer has always been the root of controversy among language teachers and evaluators. What goes on in the brain when it is stimulated to solve a problem and how it reaches the desirable results can open new horizons in the field of language learning and language acquisition and as a result the outcomes can applied to and utilised in the field of language assessment. One of the methods applied to discover the divergent aspects of cognitive process is the employment of verbal protocols. Technically speaking verbal protocols are defined as the valuable sources which contain invaluable recorded data with regard to test taker's verbalisation of the process that he or she employed to reach a solution to assigned problem.

1.2.1.2 Verbal protocol advantages

Wilson (1994) enumerates the following advantages for verbal protocols in writing assessment:

- Providing researcher with insights into the processes utilised by the test taker to solve a problem
- Providing the lay person with a set of information that is to a high extent quite easy to grasp
- Providing researchers with quickly gathered information
- Providing researchers with the basis concerning the underlying processes involved in language production and paving the way for further investigation

1.2.1.3 Verbal protocol disadvantages

Wilson (1994) stipulates the following advantages for verbal protocols in writing assessment:

- Verbalisations by themselves cannot be considered a good demonstration of the process itself
- Verbalisations may be biased
- Verbalisations cannot be considered to cover all the aspects of the employed processes

When it comes to rater variables, the available literature mainly focuses on two areas of writing assessment, namely L1 first language and L2 second language. L1 writing assessment is replete with different ideas with regard to different aspects concerning writing evaluation. In most of the available scenarios where the main concentration is on L1, researchers have employed different techniques specially the most common one, namely verbal protocol to get some data that would enable them to instigate the difference between experienced and inexperienced raters. Huot (1988) in his studies demonstrated that experienced and inexperienced raters have something in common when they evaluate and score the scripts in L1. He believes that the primary areas of attention for both of these writers are the gist and the content of the script. Huot (1988) also discovered that experienced raters in comparison with the inexperienced ones utilise the scoring techniques with more coherence. Of course the same findings were expertly presented by Cumming (1989) and Conner and Carrel (1993) with regard to their studies in L2.

1.2.2 Raters' attributes

Personal characteristics attributed to script raters have always been considered as a necessary launching pad to investigate the source of discrepancies among raters regardless of the very fact that the background had been L1 or 12. In this area there are lots of studies that have dealt with the different aspects of raters' attributes. For instance Keech and McNelly (1982)

made a comparison between three groups of raters, namely high school students, inexperienced teacher raters as well as professional teachers on the basis of their assigned holistic grades to scripts written in L1 environment.

On the basis of the data analysis they discovered that the first group, namely school students assigned the lowest grades to scripts with regard to their holistic impression while the highest mark was assigned to the professional teacher raters with the inexperienced teacher in between. With regard to the factor of leniency Sweedler Brown (1985) discovered that in comparison with the rater trainers the inexperience raters were more lenient in assigning scores to scripts written in L2.

1.2.3 Raters' cultural background

Raters 'cultural background is another factor of great importance when it comes to the raters' variables and the effects thereof in the realm of writing assessment. On the basis of present studies such as those conducted by Kobayashi and Rinnert (1999), Land and Whitley (1989) and Hinkle (1994) it was noticed that L1 raters demonstrate stronger tendency to show their affirmation to those writings that had utilised the rhetorical patterns that they were more familiar with.

1.2.4 Raters' training

Training to create harmony amongst raters has always been considered to be of significant importance especially in L2 writing assessment (Anderson, Clapham & Wall 1995, Bachman & Palmer 1996, McNamara 1996, Weir & wall 1988). It has been experimentally proved that training cannot always lead to the improvement of reliability in our assessment (Lumely & McNamara 1995, Weigle 1998).

In the area of writing assessment, there have always been factors such as the degree of the severity or leniency demonstrated by raters regardless of their cultural and linguistic background that formed the source of controversy. These two factors according to Black (1962) are directly influenced by the very nature of the task that the applicant has been exposed to and according to Lunz and O'Neil (1997) cannot be affected that much by training. Training per se can be very effective in providing the raters with an acceptable threshold to stick to , as well as in creating consistency while adapting different approaches to scoring (Lunz, Wright & Linacre 1990).

On the basis of studies performed by Weigle (1994, 1998) rater training can to some extent improve the reliability rate and help the raters curb their inclination to demonstrate harshness or clemency in their score assigning. Anderson et al. (1995) also believes that in the realm of second language teaching and writing assessment rater training plays a significant role in diminishing or alleviating rater variability.

In addition, by undergoing a focused training raters can be helped to apply scoring schemes with more consistency (Connor & Carrell, 1993; Weigle, 1994) while, the advantages of exposing raters to training may be transitional (Lumley & McNamara, 1995), and the contextual effects cannot be eliminated (Hughes, Keeling, & Tuck, 1983).

Based on studies conducted by Hamp-Lyons (1990) there are factors such as, the context in which training occurs, the type of training given, the extent to which both training and reading are monitored, and the kind of feedback that readers were given that can play an important and effective role in maintaining both the reliability and the validity of the scoring of essays.

1.2.5 Raters' expectations

When a rater is engaged in assigning scores to scripts he or she usually brings his or her special anticipation to the table, so if scripts do not meet those expectations automatically he or she will not assign the appropriate marks to them. Even for some researchers such as Stock and Robinson (1989) raters' expectations are as significant as textual quality in grade assignments.

1.3 Rating scales

In the area of both L1 and L2 writing assessment, there exists another area that tells the raters apart from each other, regardless the fact that the assessment is done in L1 or L2. The employment of certain rating skills and not the others has always been considered as the root of discrepancy among raters and language evaluators. Vaughan (1992) believed that raters would employ different approaches to assess a script holistically. He demonstrated that raters may utilise either grammar oriented approach rater or first impression dominates approach in their assessment.

According to Upshur and Turner (1995) the principal function of assessment criteria, is to compare the collected samples to certain, recommended descriptors. In assessing writing there are many factors that should be taken into account to forge relative harmony among raters. Developing a specific writing scale that does not keep raters away from achieving reliability is a very difficult task for both test developers and administrators.

In the available literature two main rating scales, namely holistic scales ,in which only a single score is given to a script and analytic scales, in which different aspects of writing such as task response (TR) ,cohesion and coherence (CC), grammatical range and accuracy (GRA) and lexical resources each is assigned a grade, are presented.

In a study conducted by Freedman (1979) when the professional writers were compared with the college writers based on the scores that they assigned, the discrepancy was laying in analytic rather than holistic employment of the scales. According to the same study professional writers could achieve a much higher grade when their writings were analytically rated while the college writers could constantly achieve the same grade regardless of either scale. In the literature there are some controversies with regard to the effectiveness of both analytic and holistic scales, for example Bauer (1981) believes that in comparison with that analytic scale, holistic scale is more cost effective but analytic scale is considered to be more trustworthy when it comes to writing assessment.

Based on the studies conducted by Bachman and Palmer (1996) and Mc Namera (1996) language assessment especially in L2 atmosphere cannot be achieved or considered valid unless appropriate rating scales and criteria are employed and observed by trained raters.

The recommended writing standards manifest themselves in three forms namely, band descriptors, marking schemes and assessment criteria. According to Wolf (1995) since a standard cannot be appropriately conveyed by the written criteria it is highly recommended that raters utilise exemplar scripts to base their judgements on. Anderson (1991) enumerates the problems that raters normally associated with using rating scales in their assessing testees' writing.

The first problem is the one that raters usually face when they try to include the testees in certain level based on the interpretation of the assessment criteria. The second area of difficulty arises from the ambiguity that exists in the borderlines between the beginning and the end of the band descriptors. The third problem is the tendency of the raters towards the descriptors themselves. Some of the raters usually find the recommended criteria