Journal of Applied Linguistics and Language Research
Volume 3, Issue 1, 2016, pp. 185-196
Available online at www.jallr.com
ISSN: 2376-760X

An Evaluation of the Construct Validity of Iranian National Test of English at High Schools

Masoud Zhoghi
Assistant professor, Department of ELT, Ahar Branch, Islamic Azad University, Ahar, Iran

Gholamreza Rostami *
PhD candidate, Department of ELT, Ahar Branch, Islamic Azad University, Ahar, Iran

Hamideh Gholami
M.A candidate, Department of ELT, Maragheh Branch, Islamic Azad University, Maragheh, Iran

Abstract
The main purpose of the current study was to investigate the construct validity of the national final tests of English for grade three senior high school students in Iran. For this purpose, a total number of forty students and forty EFL teachers in Iran were selected. A quantitative method was adopted to analyze how this test has changed over fourteen years time span based on item analysis and correlation coefficients. Results revealed that there was a significant difference between 2000 and 2014 versions of final national tests of English for grade three high schools students in Iran in terms of test items, item facility, item difficulty, and item discrimination. The second finding showed that the total score of the 2014 test correlated with every subtest. Similarly, different subtests of the 2000 version correlated with each other. In addition, EFL teachers believed that the final national test of English for grade three high school do not have construct validity.
Keywords: construct validity, correlation coefficient, item analysis

INTRODUCTION

Construct validity is required to use performance to infer the possession of certain psychological traits or qualities. These are all hypothetical qualities, called constructs, which are assumed to exist in order to account for behavior in many different specific situations. To describe a person as being highly intelligent, for example, is useful because the term suggests it a series of associated meaning which indicate what his behavior is likely to be under various conditions (Birjandi, 2010).

Before interpreting test scores in terms of these broad behavioral descriptions, however, language teachers and test developers must first establish the constructs

* **Correspondence:** Gholamreza Rostami, Email: rrostami185@gmail.com

which are presumed to be reflected in the test scores actually do account for differences in test performance. This process is called construct validation. In determining construct validity, the aim is to identify all factors which influence test performance and to determine the degree of influence of each. The process includes the following steps: 1) identifying the construct which might possibly account for test performance, 2) formulating testable hypotheses from the theory surrounding each construct, and 3) gathering data to test these hypotheses (Brown, 1997).

The English language is a required subject which is taught in Iranian high schools and as a result, English text books are naturally developed by Iranian text book developers in which reading skills and grammar are emphasized. The students are evaluated on these text books by teacher made tests. These students must take teacher- made tests twice before sitting for the national final test (Birjandi, 2010). The national final exams are administered centrally. To come up with a standardized test, a few competent and qualified teachers are invited to design the test in the Ministry of Education. In addition, there is an office inside the Ministry of Education responsible to plan, design, copy, and distribute the papers (Birjandi, 2010).

However, the problems which are frequently reported in relation to the final national test of English for third graders in Iran are (1) inadequacy of number of item numbers, (2) poor item wording, and most importantly lack of (3) piloting, item analysis, reliability, and validity studies. These are the concerns that are commonly addressed in the validation process. In order to make a comparison between two national final exams belonging to different school years, four characteristics of these tests i.e. item facility, item difficulty, item discrimination and construct validity were studied (Brown, 1997).

As one way of assessing the construct validity of a test is to correlate its different test components (Alderson & Clapham, 2000), the present study attempted to focus on correlation item analysis of the 2000 and 2014 versions of final national grade 3 in Iranian senior high schools in order to see how it has changed over time. The study, therefore, addresses the following research questions:

- **RQ 1:** How do the test item types, descriptive statistics, item facilities, item discriminations of the 2000 and 2014 versions of Iranian final national tests differ?
- **RQ 2:** How well do the 2000 and 2014 versions of Iranian final national tests correlate?
- **RQ 3:** What are the strengths and the short coming of the current version of this test?
    - **3a**: How well does the total score of the 2014 versions of Iranian final national tests correlate with every subtest?
    - **3b:** How well do the different subtests of the 2000 version of Iranian final national test correlate with each other?
- **RQ 4:** How do language teachers evaluate final national grade 3 based on construct validity?

REVIEW OF THE RELATED LITERATURE

Construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the markers intend it to measure. The goal is to determine the measuring of scores from the test, to assure that the scores mean what we expect them to mean (Bachman, 1990). According to Ebel and Frisble (1991), construct validation is the process of gathering evidence to support the contention that a given test indeed measures the psychological construct the makers intend it to measure. The goal is to determine the meaning of scores from the test, to assure that the scores mean what we expect them to mean.

Therefore, construct validity cannot be determined by a single type of evidence. Language teachers should make predictions which are in harmony with the theory underlying the construct and test them one by one. Where the data are in harmony with our predictions, they provide support for the validity of our interpretations of the scores as a measure of the particular construct. Where the data are contrary to our predictions, the language teachers revise the test interpretation, reformulate the theory underlying the construct, or improve the experimental designs used to obtain the data (Birjandi, 2010).

Since test scores cannot be interpreted as a measure of only one construct, the process of construct validation typically includes a study of the influence of several factors. We might, for example, ask to what extent the scores on our reasoning test are influenced by reading comprehension, proficiency skills, and speed (Birjandi, 2010) .

The key element in construct validity, then, is the experimental verification of the test interpretations we propose to make. This involves a wide variety of procedures and many different types of evidence. As evidence accumulates concerning the meaning of the test scores, our interpretations are enriched and we are able to make them with greater confidence (Brown, 1997).

In short, the construct validity of a test should be demonstrated by collecting evidence. For example, taking the unified definition of construct validity, we could demonstrate it using content analysis, correlation coefficients, factor analysis. Naturally, all of the above would be a tremendous amount of work, so a group of Iranian test developers is willing to put in to demonstrating the construct validity of their test .Competent test developers will stop when they feel they have provided a convincing set of validity arguments (Brown, 1996).

Empirical Studies

A number of researchers have been studied and analyzed different aspect of construct validity in different construct. In this part, some of these studied will be presented.

Bachman (1995) investigated the construct validity of examinations including FCE and TOFEL. He examined pattern of relationship in examinee's performance on the tests, both at the level of total test scores and where appropriate, at the item level .

Cronbach (1995) examined the construct validity of psychological tests. Personality tests and some tests of ability are interpreted in terms of attributes for which there is no adequate criterion. This paper indicates what sorts of evidence can substantiate such an interpretation, and how such evidence is to be interpreted.

Yujie (2007) examined the construct validity of an EFL test for Ph.D. candidates from a quantitative perspective with two versions of the English entrance examination for doctoral candidates at an institution in China as a case study. A quantitative method is adopted to analyze how this test has changed over a nineteen years times based on item analysis and correlation coefficients.

Pae, Greenberg, and Morris (2007) examined construct validity and measurement invariance of the Peabody picture vocabulary Test-III from A in the performance of struggling adult readers. Tavakoli (2011) carried out a study which tries to investigate skills assessed by the items in the tests; hence, the construct validity of the test, the study was conducted to find out the construct validity of reading paper of the first certificate in English (FCE) in Iranian EFL context.

Yarmohamadi and Sadighi (2014) investigated the construct validity of a nationwide large-scale English proficiency test, finding of the study indicated that the structure and expression section of TOLIMO early measures what is supposed to measure and the scores of this section can be interpreted as real indicators of examinee's ability level in structure and writing. The finding implies that the test demonstrated construct validity.

METHOD

Participants

A total of 40 Iranian high school students in grade three in Marzieh senior high school, Miandoab, West Azerbaijan, Iran in 2014 participated in this study. All the students were female and were pre-intermediate and intermediate level students, with the average age of 18. As well, forty EFL teachers in Miandoab, West Azerbaijan were selected in order to answer the questioners related to construct validity of final national test of version 2000 and 2014.

Instruments and Materials

Two research instruments were used in this study: paper tests and two questionnaires. The 2000 and 2014 versions tests were selected to study to measure how much this test has changed over 14 years. The compositions of two versions with the rubrics are provided in table 1 and 2.

Tables 1 and 2 outline the basic structure of both versions. Structurally, the 2014 test varies from the 2000 test in the increase of: a) the overall test time from 90 minutes to 120 minutes, b) the weight of the cloze test, c) the grammar test, d) the reading test, e) the pronunciation section, and e) the dictation section.

Table 1. A structural overview of the 2000 FNTEHSSI

Section	Sub – section	Item types	point	Item number
1. Dictation	A-B-C-D-E-F-G-H	Blank Filling	3	12
2. Vocabulary	A, B	Blank – Filling	6	12
3. Reading	A, B	MCQ & T.F	6	7
4. Conversation	A	Matching	3	6
5. Pronunciation	A	MCQ	1	2
6. Picture	A	Full Answer	1	1
7. Structure	A, B, C, D	MCQ Put in correct order Correct word	10	10
Total			30	50

Table 2. A Structural overview of the 2014 of the FNTEHSSL

1. Section	Sub – section	Item types	point	Item number
2. Dictation	A-B-C-D-E-F G.H.I.O	Blank filling	4	12
3. Vocabulary	A,B	Blank filling	6	12
4. Reading	A,B	MCQ Open - end T.F	10	14
5. Cloze	A	MCQ	4	8
6. Structure	A,B,C,D	MCQ Correct form Put in correct order	8	8
7. Conversation	A	Matching	4	8
8. Pronunciation	A	MCQ	2	4
9. Picture	A	Full- answer	2	2
Total			40	60

The second instrument was two questionnaire surveys distributed among Iranian English language teachers (N = 40). The first one consisted of 8 questions and the second questionnaire included 11 questions based on Likert scale, which EFL teachers were asked to compare the 2000 and 2014 versions of final national tests of English for grade 3 senior high school students in Iran.

Data collection

On July 20, 2015, the 2000 test was distributed to participants. The examinees were given 90 minutes to finish that test and one week later they received the 2014 test papers and were given 120 minutes to finish the exam. After two weeks, two questionnaires were distributed among English language teachers in high schools in Iran.

RESULTS

Analysis of research question 1

The descriptive statistics for both tests are presented in Table 3. Table 3 shows that the 2000 test was easier than the 2014 test. Moreover, 2000 test had somewhat smaller standard variation, nearly twice as much as the 2014 test. Though the majority of students performed within a fairly tight score band in the 2000 test, the 2000 test also had a great range of overall score distribution.

Table3. Descriptive statistics of the 2000 and 2014 tests

	2000	2014		2000	2014		2014	2000
Overall correct answer rate	68.6%	48%	Mode:	14.0	10.5	Standard Deviation	6.40	2.58
High score	84.5%	77.5%	Median:	15.5	13.0	Range	36.5 Points	33 Points
Low score	47.8%	43.5%	mean	16.0	12	Variance	43.7	39.9

Sequentially, item discrimination and difficulty indices were employed. Item discrimination (ID) ascertains where a test taker's performance shows uniformity across the examined items and item difficulty or facility (IF) investigates the properties of individual test item appropriateness for the target group's level. Items should be rejected if the IF is <.33 or >.67.To calculated the ID, first a high group and low group must be established. As suggested by Brown (1995) it should be between %25-35% of the total group. For this study, 30 %(n=20) was used. If ID of item was>.67.It was rejected as this is the lowest acceptable cut-off point. All calculations are summarized in table 4 and 5.

Table 4. Acceptable item for final National Test grade 3 with this survey sample (N=40)

Grammar (10 items total)	Dictation (12 items total)	Vocabulary (12 items total)	Reading (7 items total)	Conversation (6 items total)	Pronunciation (2 items total)
	12 items total / 4 items acceptable	5 items acceptable	5 items acceptable	5 item acceptable	No items Acceptable

Table 5. Acceptable items for the 2014 Final National Tests grade 3 with this survey sample (N = 40)

Dictation (12 items)	Vocabulary (12 items total)	Reading (14 items)	Close – test (8 items total)	Grammar (8 items total)	Conversation (8 items total)	Pronunciation (4 items total)
3 items Acceptable	4 item acceptable	8 items acceptable	1 item acceptable	3 items acceptable	4 items acceptable	No items acceptable

Table 4 and 5 reveal that only 15% of overall items from the 2000 test were acceptable and only 3/2% those in 2014 test were acceptable for this surrey Sample. This indicates that the final national test of grade 3 in Iranian High schools may need significant revision.

Analysis of research question 2 and 3

As suggested by Alderson, Clapham and Wall (2000) one way of assessing the construct validity of a test is to correlate its various test components with each other. These correlations are generally low – possibly in the order to 3 – to .0.5.On the other hands Alderson, Clapham and Well suggested that in a Wall – designed test, the correlation between each subtest and whole test can be expected to be higher – possibly around +0.7or more .Since the overall score is taken to be a more general measure of language ability than each individual component score.

Tables 6 and 7 list the various correlations for the 2000 test. Those with a single asterisk were statistically significant at the P <.05 level and those with double asterisks were significant at the P<.05 level and those with double asterisks significant at the p<.01 level.

Table 6. Correlation coefficients of the score of the 2000 test with each subtest and the various subtests with each other

	List wise Correlation (n = 44)	Total Score	Dictation	vocabulary	Structure	Reading	pronunciation
Total score	Pearson Correlation	1	0.338	0.32	0.386	0.336	0.521
	Sig 2 – tailed	3	0.005	0.0008	0.001	0.005	0.00
Dictation	Pearson Correlation	0.338	0.216	0.288	1	0.80	0.292
	Sig. (2 - tailed)	0.001	0.80	0.18		0.521	0.017
Vocabulary	Pearson Correlation	0.325	0.186	1	0.288	0.175	-0.032
	Sig.(2- tailed)	0.008	0.136		0.19	0.156	0.755
Structure	Pearson Correlation	0.386	0.306	0.38	0.293	0.0807	1
	Sig.(2-tailed)	0.001	0.12	0.754	0.016	0.980	
Reading	Pearson Correlation	0.336	0.93	0.75	0.0805	1	0.88
	Sig.(2-tailed)	0.006	0.455	0.156	0.521		0.480
Pronunciation	Pearson Correlation	0.338	1	0.0185	0.216	0. 092	0.306
	Sig (2 -tailed)	0.005		0.136	0.079	0.455	0.012

Table 7. Correlation coefficients of the total score of the 2005 test with each subset and the various subsets with each other

			I	II	IV	V	VI	VII
	List wise Correlations n=44	Total score	Dictation	vocabulary	Grammar	cloze	reading	pronunciation
Total score	Pearson correlation	1	0.547	0.534	0.421	0.463	0.626	0.448
	Sig.(2-tailed)		0.000	0.000	0.000	0.000	0.000	0.000
Dictation	Pearson correlation	0.486	0.429	0.142	0.70	0.222	-0.025	0.488
	Sig.(2-tailed)	0.00	0.000	0.561	0.560	0.252	0.838	0.00
Vocabulary	Pearson Correlation	0.534	0.103	1	0.340	0.026	0.284	0.105
	Sig.(2-tailed)	0.000	0.404		0.005	0.830	0.20	0.404
Reading	Pearson Correlation	0.626	0.222	0.26	0.93	1	0.038	0.222
	Sig.(2-tailed)	0.000	0.016	0.820	0.456		0.755	0.14
structure	Pearson Correlation	0.452	0.187	0.284	0.280	0.38	1	0.177
	Sig.(2-tailed)	0.000	0.96	0.21	0.022	0.75		0.129
cloze	Pearson Correlation	0.463	0.438	0.34	1	0.02	0.28	0.087
	Sig.(2-tailed)	0.000	1	0.004		0.455	0.27	0.429
Pronunciation	Pearson Correlation	0.547	0538	0.04	0.096	0.292	0.187	1
	Sig.(2-tailed)	0.000		0.404	0.438	0.017	0.130	

Correlation significant at the 0/01 level (2- tailed). Correlation significant at the 0/05 level (2- tailed).Correlation of both tests corresponding sections are an effective way comparing their construct and seeing how consistent they are with each other.

The correlation for the 2000 and 2014 tests are summarized in Table 8. The overall correlation coefficient was 0.315 P <.05), suggesting only a moderate correlation between the two tests. According to Morgan, Griego and Gloeckner (2001) the effect size was medium. The correlation of the 2000 and 2014 vocabulary sections was 0.166 but the P was 0.179 – which was considerably higher than .05. The correlation of the 2000 and 2014 close sections was 0/145about this was not statistically significant (P =.0242). The correlation of the 2000 and 2014 reading section was .059, yet this too was not statistically (P=637) .The correlation of the dictation parts of these two exams was the highest (.0356) and it was statistically significant. Possible reason for these figures will be discussed in the next section of this paper.

Table 8. Correlations of corresponding sections of the 2000 and 2014 tests

	List wise Correlations	Total score	Dictation	vocabulary	Structure	cloze	reading	Pronunciation
Total score (2000 & 2015)	Pearson correlation	0.315	0.356	0.166	0.136	0.463	0.05	0.145
	Sig.(2-tailed)	0.009	0.002	0.049	0.233	0.000	0.637	0.242

Analysis of research Question 4

Now let us consider how the teachers felt about two different test examining the questionnaires which were administered after two weeks. Respondents were given a 5 point liker Scale to answer 8 questions in the first questionnaire and 11 questions in the second questionnaire.

Table 9. Survey responses for 2000 final national test in Iranian high schools

1 = very easy and 5 = very difficult For Q 1 – 4 1 = strongly disagree and = strongly agree For Q 5 – 8			
Survey Item	Number of responses	Mean	Std Deviation
Q1. How difficult was the structure section of this test?	44	3030	0.840
Q2. How difficult was the vocabulary section of this test?	44	3.70	0.875
Q3. How difficult was the dictation section of the test?	44	3.51	0.910
Q4. How difficult was the close section of the test?	44	3.21	0.910
Q5. The grammar section reflects students' English proficiency?	41	3.27	0.009
Q6. The vocabulary section reflects students' English proficiency?	41	3.53	0.885
Q7. The cloze section reflects students' English proficiency?			
Q8. The reading section reflects students' English proficiency?	41		0.984

Table 10. Survey responses for the 2014 final National test in Iranian high schools

Note 1 = very easy and 5 = very difficult For Q. 1 – 5 1 = strongly disagree 5 = strongly agree for Q. 6 - 11			
Survey Item	Number Of Responses	Mean	Std. Deviation
Q1. How difficult was the dictation section of this test?	44	2.76	0.818
Q2. How difficult was the vocabulary section of this test?	44	3.61	0.873
Q3. How difficult was the close section of the test?	44	3.36	0.670
Q4. How difficult was the Reading section of this test?	44	3.64	0.743
Q5. The grammar section reflects students' English proficiency?	43	3.30	0.827
Q6. The dictation section reflects students' English proficiency?	44	3.72	0.832
Q7. The vocabulary section reflects students' English proficiency?	44	3/38	0.925
Q8. The cloze section reflects students' English proficiency?	44	3.70	0.843
Q9. The Reading section reflects students' English proficiency?	43	3.67	0.826
Q10. The pronunciation section reflects students' English proficiency?	43	3.61	0830
Q11 The conversation section reflect students English proficiency?	44	3.62	0.874

As for question 11 in the second surrey about 42% (n = 44) respondents felt that the 2014 test was more difficult than 2000 Test. Since the scores for the 2000 Test tended

to be higher than in 2014 test. The quantitative data supports this. Interestingly 78/2% (n=44) of the respondents felt that 2000 test was more indicative of their English abilities than the 2014.

DISCUSSION AND CONCLUSION

As mentioned earlier, the three research findings were significant. The first research finding concerned differences of the test format , item facility , item discrimination and some descriptive statistics between the 2000 and 2014 tests . As for the response format and item types, it is fair to say that 2014 test differed significantly from the 2000 test. The 2000 exam attempted to measure dictation , vocabulary , grammar , reading , conversation and pronunciation while the 2014 test purported to measure dictation , vocabulary , grammar , cloze ,reading , conversation ,stress and pronunciation. The number of items increased from 50 items to 60 items in 2014 version .Points of items increased in 2014 version from 30 to 40 points. Tables 5 and 6 suggested that 2014 tests had many items which were performing poorly in terms of ID and IF. One possible reason for this was due to the tests format whereas vocabulary and dictation sections were fill in the blank in 2000 test ,in 2014 test Some sections in the 2014 exam were all in multiple – choice format .

The second research finding concerned the correlation between 2000 total score and its subtests. It is curious that reading part has lowest correlation coefficient because in the 2014 examination this section had the lowest correlation with the total scores. This suggested that the topic of the reading passage may have an important role in shaping performance since the examinees draw upon their background knowledge when writing (Clapham, 1996),so the text familiarity and task type has significant differences in subject overall and differential test and task performances (Salmani-Nadoushan,2003).

The 2000 reading passage was very simple rather than the passage was in the 2014 test. It contained open- end and true – false questions .By contrast 2014 reading passage which was about scientific topic, the reading passage included open –end, true – false and multiple-choice questions. The reading passage for the 2000 test was probably more familiar to the examinee rather than 2014 reading passage.

The number of questions in reading section was increased rather than 2000 reading passage questions. . Grammar section in 2000 versions was more easily than 2014 versions because numbers of items, format of items were changed. In dictation part, the number of items was increased than 2000 version. The close test items were added to the items in 2014 versions whereas this item was not in 2000 version. So this reasons makes students' scores be less rather than the 2000 test.

As Suggested by Alderson, Calpham and Wall (200) the correlation of subtests should be possibly in the order of +0.3 - + 0.05. In the 2000 test only sections 1, 2, 3, 4 and 5 had such correlations. This suggests that the 2000 test measured constructs which were quite easy. The 2005 test had only four sections which correlated within the parameters suggested by Calpham and Wall. Section 2 and 5 correlated moderately, as did sections

1 and 6. The fact that many of the subtests didn't correlate and it should focus and reflect on what this test is actually measuring (The dictation section, cloze test section, reading section and pronunciation section).

The third research finding stated that listening and speaking skills did not measure by 2000 and 2014 tests. The 2000 and 2014 tests try to focus on reading, vocabulary and grammar. According to the questionnaires which were analyzed, ID of 2014 is more than ID of 2000 version, so the teachers agree with it, the results show that the validity of final exam of grade 3 in Iranian high schools needs to be investigated further. The tests takers should be make a test which measure student's ability in every skill because skills like writing, listening and speaking don't measure student's ability perfectly. The numbers of items, test types, time of test seem not suitable with student's ability.

The results of our study indicated that final national test of English for grade 3 in Iranian high schools cannot measure what it supposed to measure and scores cannot be interpreted as real indicators of examinee's ability level in reading, writing, speaking and listening skills. The findings imply that the test cannot demonstrate construct validity in this section but in dictation, vocabulary and grammar findings demonstrate construct validity to some extent.

Although this research is preliminary, it has four practical implications. First of all, it points out the need to enhance the item discrimination and item facility ratings for the final national tests of English for grade 3 high schools in Iran. The 2014 had low percentage of acceptable items in regard to ID and If. In particular, cloze, reading, structure and pronunciation sections need improvement. Second, this study points out the need for a closer examination of conversation section of the 2000 and 2014 tests. The average mean score for this section was 12/2 so this section was too easy. Third the cloze section wasn't in the 2000 test but this section in added in the 2014 tests which it had a lower correlation in 2014 tests.

Fourth, this study also highlights the need for qualitative Feedback on the exam. In particular a well – triangulated analysis by students, teachers and test developers of what construct they believe this exam taps into and what they consider to be some of the biases inherent in the exam would shed valuable light on not only the way the exam is structured, but also the exam content. Test development is cyclical, not linear (McNamara 2000, P.23). That is, once a test is designed, constructed, trailed and operationalized its actual use generates evidence about its qualities (McNamara, P.32).

There are still some weaknesses with 2014 test. In that light, the Following three proposals are offered.

1) The cloze section need to have wider response format which include integrated and interactive test items rather than solely multiple – choice items. By incorporating a wider range of tasks and response Formats, more skills can be tested and hence the score can measure what it is supposed to measure.
2) Since 2014 test has a quite low percentage of appropriate items, by conducting piloting and /or pre – testing the ID and IF levels could be raised. That is, by having

a system by which the right statistical procedures are followed. Items which misfit or perform poorly would automatically be deleted. A Rasch analysis could be employed to do this.

3) Listening, speaking and writing in both 2000 and 2014 test needs to be improved. The validity needs to be investigated further.

REFERENCES

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

Bachman, L. F. (1990). *Fundamental considerations in language Testing*. Oxford: Oxford university press.

Bachman, L. F. & Palmer, A. (2000). *Language testing in practice*. Oxford: Oxford university press.

Birjandi, P. & Mosallanejad, P. (2010). *An overview of testing and assessment.* Tehran: Rahnama publication.

Breland, H. M., Kubota, M. Y. & Marilyn, W. (1999). A performance Assessment study in writing. *ETS Research Report Series, 2*, 1–18.

Brown, H. D.(2004). *language assessment principles and classroom practices*. White Plains, NY: Pearson Education.

Cronbach, L. J. (1990). *Essentials of psychological Testing*. New York. Harper Collins Publishers.

Hughes, A. (2003). *Testing for language Teachers.* Cambridge: Cambridge university press.

Jafarpour, A. (1999). *Statistics in linguistic science.* Shiraz: Shiraz university press.

Mousavi, S. A. (1999). *A dictionary of language Testing*. Tehran: Rahnama publication.

Praphal, K. (1990). *The Relevance of language Testing Research in the planning of language programmers*. London academic publication.

Marvin,L & Simner, C. (1999). *Postscript to the Canadian psychological Associations position statement* Retrieved Dec 5, 2015 from WWW.CPA.CA/Documents/TOEFL

Huong, T. (2001). *The construct Validity of the International English language testing system* (IELTS). London: academic publication.

Hanning, G. (1987). *A guide to language testing, evaluation*, research. Rowley, Massachuetts: Newbery House.

Thomas, B. (2006). *A closer Look of Construct Validity in C-test.* London: London academic publication.

Flucher, G. & Davidson, F. (2007). *Language Testing and Assessment*, London: Routledge.

Tavakoli, E. & Barati, H. (2011). *Investigating the construct Validity of the FCE and TOFEL.* Isfahan university press.

Yujie, J. (2007). *Evaluating the construct validity of EFL Test for PHD candidates.* Shanghai: Foreign Language Education Press.

Milton, E. S. & Gregory, T. S. (2009). Construct validity: Advance in Theory and Methodology. *Annual Review of Clinical Psychology, 27*(5), 1-25.