CSCanada
Higher Education of Social Science
Vol. 2, No. 1, 2012, pp.16-20
DOI: 10.3968/j.hess.1927024020120201.452

ISSN 1927-0232 [Print]
ISSN 1927-0240 [Online]
www.cscanada.net
www.cscanada.org

The Effects of Test Facets on the Construct Validity of the Tests in Iranian EFL Students

Zahra Shahivand[1,*]; Abdolreza Pazhakh[2]

[1] Department of English Language Teaching, Science and Research Branch, Islamic Azad University, Khuzestan, Iran
[2] English Language Department, Dezful Branch, Islamic Azad University, Dezful, Iran
[*]Corresponding author.

Received 15 October 2011; accepted 3 January 2012

Abstract

Language testing as a main device in assessing the learners' knowledge and language abilities plays a key role in training programs. Generally, the goal of language testing is to assure the extent to which learners have achieved the instructional goals during a course. The main objective of many studies in language testing has been to investigate whether test facets affect construct validity of the test or not. Therefore, in this study, we investigated whether the EFL Iranian participants' performances were different with respect to the different test facets and if these performances had some effects on the construct validity of the tests. In this investigation, the students were selected of 50 Iranian EFL students aged between 21 to 30 years, from two branches of Islamic Azad University, Dezful and Andimeshk, Iran. The 17 participants, placed at the low level in the Nelson proficiency test, received a test. The test facets included the integrative forms such as cloze-test, c-test, and discrete test items such as multiple choice and true/false. By statistics analyses, the significant differences were assessed in the test facets. Our results revealed that significant differences existed in the test facets among the performances of Iranian EFL students. Because of the integrity of the several abilities and mental strategies, the cloze-test was the most difficult form of testing.

Key words: Test facets; Construct validity; Integrative test items; Discrete test items

Zahra Shahivand; Abdolreza Pazhakh (2012). The Effects of Test Facets on the Construct Validity of the Tests in Iranian EFL Students. *Higher Education of Social Science, 2*(1), 16-20. Available from: URL: http://www.cscanada.net/index.php/hess/article/view/j.hess.1927024020120201.452 DOI: http://dx.doi.org/10.3968/j.hess.1927024020120201.452.

INTRODUCTION

Multiple choice tests are the most common type of tests used in evaluating the general English knowledge of the students in most universities in Iranian contextl; however, the efficacy of these tests are not examined precisely. We compare and examine the integrative tests and discrete point tests as measures of the English language knowledge of Iranian English major students. Besides, testing in general, and language testing in particular, as a main device in assessing the learners` knowledge and language abilities play a key role in training programs. Generally, the goal of language testing is to assure the extent to which learners have achieved the instructional goals during a course, so developing valid tests would be a troublesome task to be accomplished. Test facets are, in fact, of the greatest importance in determining the effects of the test on the learner's performance. In this regard, Rahimi (2007) indicates that when different test formats are used to measure certain ability, they lead to obtain different findings. In other words, the way of test administration may have some effects on the learner's performance and test results. In testing and the way of test administration, we deal with two types of test items, one is integrative test item, and the other one is discrete test item.

Giri (2002) expressed that the integrative tests such as cloze tests, are to practicalize the learners' knowledge of language, through the learners' use of more linguistic items to make the text meaningful. This includes the integration of a set of language items, for instance,

Copyright © Canadian Research & Development Center of Sciences and Cultures

eliciting information, knowledge of vocabulary as well as the ability to make conceptualization. Oller (1979) also claims that the knowledge of language cannot be measured in discrete forms, because it consists of an integrative set of items to assess the learner's competence. Besides, other researchers and applied linguists have different ideas. Weir (1990) believes that the integrative tests, such as cloze test and c-test only demonstrate a view of the learners' knowledge, and they fail to illicit the learners' language performance. Above all, Mousavi (2009) defined construct validity as "a form of validity which is based on the degree to which the items in a test reflect the essential aspects of the theory on which the test is based". (p.138). He added that when a test measures only the abilities it is supposed to measure, we can say the test has construct validity. "The construct validity is the most important type of the validity which can dominate all others" (Farhady, Ja'farpur & Birjandi, 2004, p.154) .

We reviewed pertinent works and the already prepared studies to support the above- mentioned ideas and arguments. Significantly, Ajideh and Esfandiari (2009) conducted a study on two groups of young freshman students at Tabriz University, in order to investigate and compare two tests formats, the multiple-choice test and cloze test. First, they administered a test to homogenize the participants. Following that, they designed two test forms, the multiple-choice tests of the lexical items, and cloze tests with ratio deletion. The contents of the two tests had been kept constant. After administering the tests, they used the statistical procedures to have the obtained scores analyzed. Finally, they concluded that in testing the proficiency of a group of learners, the achieved scores on the multiple-choice lexical tests were much similar to the cloze test scores. Although two tests were seemingly different, there was a high correlation between the two types of test formats on vocabulary-discrete-point item, and integrative cloze test. An interesting point was that those who acted better on cloze tests could also perform better on discrete-point tests.

In another study conducted by Grabowski (2008) to analyze the influences of the test facets on learner's scores, he worked with 60 adult English language learners, from the Teachers College, Columbia University, Community English Program (CEP), who participated in their study in different levels of age and gender groups. He used a model which consisted of both pragmatic and grammatical aspects to assess the participants' abilities in expressing and analyzing the implied meaning. The participants' answers were scored by two raters. A Rasch model of measurement was used to ascertain the trustworthiness of the nonnative speakers' pragmatic test scores and to support the claims of validity of the underlying test construct by recognizing the potential sources of variability in the participant's scores, also, to confirm the abilities of test-takers to show a fair estimate by comparing the test formats on the same scale . It was found that applying two test formats is an acceptable method to extract the learners` language competence, though each of them has different results in learners' performances.

In sum, although various researchers have tried to examine the test facet performance, the question has still remained vague as whether test facets affect construct validity of the test or not. Therefore, in this study, we investigated whether the Iranian EFL participants' performances were different with respect to the different test facets and whether their performances had anything to do with construct or validity of the tests. In other words, we wanted to see if test facets had any significant effects on the construct validity of the tests in question.

STATEMENT OF THE PROBLEM AND PURPOSE OF THE STUDY

Farhady (1979) claimed that the difference in learner's background knowledge overshadows the scores in some test categories such as discrete and integrative tests. A student, who is not experienced enough in various formats of testing, should not be expected to do well in unknown formats as opposed to more known ones. As the discrete tests are easily prepared and are frequently used to measure the learners' knowledge, teachers prefer this type of test; besides, applying other test formats, such as integrative tests, cloze tests c-test may lead to some confusions on the learners' part.

The purpose of this study is to investigate: the effect of test facets on the construct validity of the tests, the participants` performances in various test formats and the relationships among the results of each test facet compared to other test facets.

RESEARCH QUESTIONS

1. Would the test facets affect the construct validity of the tests?

2. Would the test facets differentiate the test-takers' performances in the tests?

3. Would the results of a test-takers' performances change across different test formats?

RESEARCH HYPOTHESES

H$_{01}$.The test facets do not leave or exert significant effects on the construct validity of the tests.

H$_{02}$. The test facets do not make significant differences among the test-takers' performances on tests.

H$_{03}$. The results of the each test do not differ significantly from the results of other test formats.

METHODOLOGY

Participants

The present study consisted of 17 students- 7 male and 10 female from two branches of Islamic Azad University in Khouzestan Province - Dezful, and Andimeshk, Iran. They were selected from a population of fifty EFL students of the two available classes; one class was third year students at B.A. program of English Translation, and the other one was third year students at B.A. program of English Language Teaching courses, during the Fall 2011. The age of the participants varied from 21 to 30. All of the sample population sat for the proficiency test to decide on their proficiency level. Accordingly, they were divided into three proficiency levels: low, intermediate and high- based on their scores on the Nelson proficiency test (Fowler & Coe, 1976). Following that, 17 participants were selected non-randomly, by applying purposive sampling to compare the test-takers' performances in the low level of proficiency.

Instruments

The instruments used in this study were: Nelson proficiency test (Fowler & Coe, 1976) in order to estimate the proficiency level of the sample population, also, to select homogenized participants. The test included 50 items; each item valued 1 point. Those students whose scores fell within the range of +1 SD above and -1 SD below the mean, they were considered as the mid-level ones. The scores which range below and above mid-level were regarded as low and advanced proficiency level, respectively. Another instrument used was a pre-test which was administered to the participants with the low level of the proficiency. In fact, this test comprised of a text chosen from "Exploring New Reading Strategies", level 1, by Birjandi and Mosallanejad (2010). All the participants performed on different types of this test facets. Four test facets-C-test, Cloze-test, Multiple choice and True/ False form were designed to assess the participants' performances.

Procedure

To select homogenous participants, all 50 participants in the study took the Nelson proficiency test (Fowler & Coe, 1976). To estimate the reliability of the test, a pilot test was done and the KR-21 formula was applied to the obtained data of 10 participants who had already taken the test. All the participants who were signified homogeneous in terms of their scores on the proficiency test were divided into three proficiency levels-high, intermediate and low. Then, among 50 participants, the 17 students who had been assigned to the low level of proficiency, took the four test facets designed based on the "Exploring New Reading Strategies", level 1 (Birjandi & Mosallanejad, 2010). It has to be stated that only those subjects with low proficiency level were selected, because the assumption was that subjects of this level are more sensitive to test facets and need to be given much attention in learning in general and testing in particular.

Statistic Analysis

Firstly, the reliability of the Nelson proficiency test which administered beforehand through a pilot study for 10 participants, was calculated on the base of KR-21 formula, and it was 0.86. Then, the descriptive statistics were calculated for all participants' scores in the Nelson Proficiency Test (Table1).

Table 1
Descriptive Statistics

	N	Mean	Variance
Score	50	27.8400	82.749

However, the descriptive statistics for the 17 participants participating in the study were calculated and presented in Table 2. An ANOVA test was used to see if there was any significant difference among the participants' performances. As Table 3 shows, the mean differences across all the four facets were significant (P< 0.05). This made the researcher claim that the meaningful differences could be attributed to the treatment of the study. So, the first and second null hypotheses were rejected, because the test facets imposed significant effects on the construct validity of the tests and the test-takers' performances.

Table 2
Descriptive Statistics for 17 Students in the Low Level of the Proficiency

Test	N	Mean	Std. Deviation	Std. Error
M.C	17	4.3529	0.78591	0.19061
True/False	17	4.7647	0.43724	0.10605
Cloze-test	17	3.5294	1.06757	0.25892
C-test	17	4.8235	0.39295	0.09531
Total	68	4.3676	0.87936	0.10664

As it can be seen in table 2, the c=test facet accounts for the highest mean score, while the cloze-test facet accounts for the lowest index of mean score.

Table 3
The One-Way ANOVA: Analysis of Variances for Tests

	Sum of Squares	df	Mean Square	F	Sig.
Between Test	18.162	3	6.054	11.515	0.000
Within Test	33.647	64	0.526		
Total	51.809	67			

Table 3 shows results acquired in the one-way ANOVA to find if there was a significant difference in the means of performances of the subjects across the four facets. Table 4, however, shows the redults from another analytic method 'a Scheffe test' to pinpoint the exact location of the difference among the means.

Table 4
Homogeneous Subsets: Scheffe Test

Test	N	Subset for alpha= 0.05	
		1	2
Cloze-test	17	3.5294	
M.C test	17		4.3529
True/False	17		4.7647
C-test	17		4.8235
Sig.		1.000	0.319

In Table5, by comparing the mean differences among 17 students in four test facets, the researcher found that the mean differences were significant at the level of 0.05. The difference was in a sense categorical. That is, cloze facet was found to be a different category than other facets which were relativelt homologus in nature. In fact, the data show that the performances of the students in Cloze-test were much lower than the students'performances in the other test forms. However, no significant difference was observed among the mean scores of the other test facets - multiple choice, True/ False, and C-test. Also, Table 5 indicates the means for four test facets, and the sequence of difficulty in the students' responses to the four test forms were in order: Cloze-test, multiple choice, C-test, and then True/False test forms. So, the third null hypotheses which claimed that the results of the each test do not differ significantly with the results of other test formats was rejected, because the mean differences are significant at $p<0.05$.

Table 5
Scheffe Test: Multiple Comparisons

(I) Group	(J) Group	Mean Difference(I-J)	Std. Error	Sig.
M.C tests	True/False	-.4118	.24870	.439
	Cloze-Test	.8235*	.24870	.017
	C-Test	-.4706	24870	.319
True/False	M.C Test	.4118	.24870	.439
	Cloze-Test	1.2353*	.24870	.000
	C-Test	-.0588	.24870	.997
Cloze-Test	M.C Test	-.8235*	.24870	.017
	True/False	-1.2353*	.24870	.000
	C-Test	-1.2941*	.24870	.000
C-test	M.C Test	.4706	.24870	.319
	True/False	.0588	.24870	.997
	Cloze-Test	1.2941*	.24870	.000

*P<0.05 Mean Difference was significant at the 0.05 level.

Significant of the Study

A few studies in the case of test facets' effects on the construct validity have ever been done in Iran as an EFL context. Students sometimes have the same understanding of a given test, but the way in which the test is administered leads to different consequences. In testing, by applying different test facets, we can examine much knowledge of the students. Through different test forms, students learn to study and understand the material comprehensively in different ways and it allows them to tap their strategies to various test facets in different administrations. Also, we can examine how various test forms lead to better or worse performances on the learners' part.

DISCUSSION AND CONCLUSION

According to the findings, the students' performances were different in different test facets. By comparing the obtained data and analyses of the results of the ANOVA and Scheffe test, it can be concluded that the most significant differences were seen in Cloze-test, because this form is an integrative test and students must integrate several abilities and mental strategies to complete the test. As these students didn't have enough experience in Cloze-tests, so the researcher found that this form of testing was difficult to be answered by the students. In other test facets-Multiple choice, True/False, and C-test, students were more familiar and could recognize the key to answer in discrete items, for example in multiple choice questions, students would find the answer among three or four options and sometimes would answer by guessing or cheating. Also, in True/False test items, the chance of answering is 50% to 50%, so students could answer the items by simplicity or by chance. And in C-test, one letter of the word was given, so students could complete the blank by this key. Another conclusion was that the discrete test items were simple to answer, because these types- Multiple choice and True/False forms- measure one aspect of the language, and students could answer the items more easily than integrative test items. So, the different ways of test administration made different performances in the students. The general view to results indicate that the students performed better in the discrete point test rather than the more integrative test. Our findings show

Copyright © Canadian Research & Development Center of Sciences and Cultures

that students perform better in non-productive rather than productive test. Since being competent English language user is an expected outcome of university language courses, it seems warranted to switch to integrative tests as a measure of English language competency.

REFERENCES

Ajideh, P. & Esfandiari, R. (2009). A Close Look at the Relationship Between Multiple Choices Vocabulary Test and Integrative Cloze Test of Lexical Words in Iranian Context. *Journal of English Language Teaching*, 2(3), 163-170.

Birjandi, P. & Mosallanejad, P. (2010). *Exploring New Reading Strategies*. Tehran: Sepahan Publication

Farhady, H. (1979). The Disjunctive Fallacy Between Discrete-Point and Integrative Tests. *TESOL Quarterly, 13*(3), 64-74.

Farhady, H., Ja'farpur, A., & Birjandi, P. (2004). *Testing Language Skills from Theory to Practice (11th Ed)*. Tehran: The Center for Studying and Compiling University Books in Humanities (SAMT).

Fowler, W.S. & Coe, N. (1976). *Nelson Proficiency Tests*. London: Butler & Tanner Ltd.

Giri, R, A. (2002). Approaches to Language Testing. *Journal of NELTA, 7*(1&2), 11-25.

Grabowski, K, C. (2008). Investigating the Construct Validity of a Performance Test Designed to Measure Grammatical And Pragmatic Knowledge. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6,* 131-179..

Mousavi, A. (2009). *An Encyclopedic Dictionary of Language Testing*. (4th Ed). Tehran: Rahnama, Press, I. R. Iran.

Norris, J., J. Brown, T. Hudson, W. Bonk (2002). Examinee Abilities and Task Difficulty in Task- Based Second Language Performance Assessment. *The Journal of Language Testing, 19,* 337-346.

Oller, J. W, Jr. (1979). *Language Tests at School: A Pragmatic Approach*. London: Longman.

Rahimi, M. (2007). L2 Reading Comprehension Test: Does the Language of Presenting Items Affect Testee's Test Performance?. *Journal of Social Sciences & Humanities of Shiraz University, 26*(4), 67-86.

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

Copyright © Canadian Research & Development Center of Sciences and Cultures