The Journal of Applied Linguistics Vol. 4, Issue 1
Spring 2011

Test Method Facet and the Construct Validity of Listening Comprehension Tests

Roya Khoii[1]

Sara Paydarnia
North Tehran Branch, Islamic Azad University, Tehran, Iran

The assessment of listening abilities is one of the least understood, least developed and, yet, one of the most important areas of language testing and assessment. It is particularly important because of its potential wash-back effects on classroom practices. Given the fact that listening tests play a great role in assessing the language proficiency of students, they are expected to enjoy a high level of construct validity. The present study was dedicated to investigating the construct validity of three different test formats, namely, multiple-choice, gap filling on summary (also called listening summary cloze), and fill-in-the-blank, used to evaluate the listening comprehension of EFL learners. In order to achieve the purpose of the study, three passages with relatively similar readability levels were used for the construction of 9 listening tests, that is, each appeared in three formats. Following a counter-balanced design, the tests were administered to 91homogeneous EFL learners divided into three groups. The statistical analysis of the results revealed that the multiple-choice test enjoyed the highest level of construct validity. Moreover, a repeated measure one-way ANOVA demonstrated that the fill-in-the-blank task was the most difficult with the MC test as the easiest for the participants.

[1] Corresponding Author. Email: roya_kh@yahoo.com

Keywords: Construct Validity, Factor Analysis, Gap Filling On Summary (Listening Summary Cloze), Multiple-Choice Items, Fill-In-The-Blank Task

Accordng to Davies (1990, p. 57), "Testing lies at the center of language teaching". There could be no science without measurement, so testing is one form of measurement for educational assessment. Like teaching, testing deals with all four language skills. Listening has long been recognized as an essential part of communication and an important trigger of language acquisition (Rubin, 1994). Thus, testing listening comprehension is one of the major concerns of language testing, and for decades , it has been considered in high-stakes tests and other examinations. However, there has been relatively little research on how to measure listening comprehension in a reliable and valid manner (Shin, 2008). In fact, listening, as one of the major language skills, has been neglected for years by most language teachers and test designers.

As the process of listening performance itself is the invisible and inaudible process of internalizing meaning from the auditory signals being transmitted to the ear and brain, most language teachers prefer to ignore it in their classes. As Brown (2004) argues, we cannot observe the actual act of listening, nor can we see or hear an actual product. We can observe learners only when they are listening, so all the teachers can do is to assess listening comprehension on the basis of observing the test taker's speaking or writing. In other words, we can only observe the result of the test taker's auditory processing in the form of spoken or written responses. Thus, the assessment of listening can be done only by drawing inference from the test takers' speaking or writing in responding to aural passages. Hence, it is particularly important in listening tests to ensure that the questions actually measure the construct of listening comprehension (Shin, 2008).

Second language listening comprehension tests are very common in language education; yet, a review of the literature on listening suggests that there is no generally accepted theory of listening comprehension on which to base these tests. It seems that

in practice test constructors follow their instincts and just do their best when constructing tests of listening comprehension. Therefore, there is an urgent need for research into the listening process and the best ways of testing it. One crucial principle for assessing a learner's competence is to consider the fallibility of the results of a single performance, such as that produced in a test. As Brown states (2004, p.117), "We must rely as much as possible on observable performance in our assessments of students". Observable means being able to see or hear the performance of the learner. For this reason, a number of tasks are in widespread use, varying from the more closed, objective tasks of completing an outline of a talk/lecture, to the more open, subjective tasks such as writing a summary of a lecture.

Factors Affecting Test Performance

In the field of language testing, there is a steadily growing interest in the identification and characterization of those factors which affect the test performance of language learners with the objective of achieving more informed construct validation results (Bachman, 1990; Foster & Skehan, 1996). Bachman (2002) points out that we should clearly distinguish among three sets of factors that can affect test performance: 1) characteristics inherent in the task itself, 2) attributes of test takers, and 3) interactions between test takers and task characteristics. Bachman (1990) argues that various variables that are not part of test takers' language ability may affect test performance, including personal attributes such as test takers' cognitive style, random factors such as test taker's emotional state at the time of taking a test and the characteristics of the test methods such as test formats, test question types, and test organizations.

In'nami and Koizumi (2009) proposed that among the many existing variables which affect language test scores, one central issue is the effect of test formats on test performance (e.g., Alderson, 2000; Bachman & Palmer, 1996; Brantmeier, 2005; Buck, 2001). As In'nami and Koizumi (2009) suggest in their study, different formats or methods such as cloze , c-test, gap-filling, matching , multiple-choice, open-ended, ordering, recall ,

summary , and summary gap filling have been employed in language testing (e.g. Alderson, 2000; Buck, 2001).They add that because there is no prefect test format that functions well in all situations, researchers must understand the characteristics of each format and select the best format which most appropriately serves the purpose of a test in each context.

Ying-hui (2006) claims that task features can be further categorized into those related to task input (or text) and those to test item. A review of studies examining task features and test performance suggests that variations in the specific characteristics of task input and test item affect the difficulty of items. It is, therefore, vitally important for language testing researchers to determine what the nature of the relationship between test tasks and test performance is, and how it affects the interpretation of test results. The information can be used as the basis for the improvement of test reliability and validity, and more specifically, for the design of tests for particular populations.

The scholars in the field of language testing have continuously tried one means or another to find a reliable, valid and practical measure of different aspects of second or foreign language. Eykyn (1992) examined the impact of four test item types on the listening comprehension of French beginners and found that multiple-choice tests resulted in the best scores. Teng (1998) studied the effects of test item types and question preview on listening comprehension tests for 187 freshmen in Yullin Technology University. The three test item types selected for the research were: multiple choice, cloze test, and short answer questions. She concluded that although the content of the three types of tests were identical, different types of questions resulted in different test scores. The multiple-choice test resulted in the highest scores followed by the short answer questions whereas the cloze test resulted in the lowest scores.

Different Methods of Testing Listening

Listening comprehension is usually measured in different ways using a wide variety of techniques. Associated with each of them are a set of theoretical notions about language and what it

means to comprehend spoken language. Each of these methods emphasizes particular aspects of language ability, and when we examine the testing techniques associated with each, we might see considerable overlap between them. What is certain is that test-takers' performance would be affected by the type of response that is required of them. Buck (2001) argues that test-takers may have to mark on a score sheet or simply make addition or alteration to a drawing or diagram (selected responses). They may be required to write one word in the case of a gap filling test, or one or more sentences in the case of comprehension questions (constructed responses). Other response formats are drawing pictures or creating diagrams.

Based on the focus of this study, four of the main test formats are briefly explained below. The multiple-choice, fill-in-the-blank, and gap-filling-on-summaries are the main test formats studied in this research; however, the listening cloze format is also explained because of its relation to the two fill-in-the gap tests.

Multiple-Choice Questions

Selected responses can be of many types, but the most common is the multiple choice item with three, four, or even five options. The construction of such items demands a high level of professional skill, which takes considerable time and training to do well. All items ought to be pretested before being used in any high-stakes assessment, but this is particularly the case with multiple choice items. Questions can be presented aurally (after the text has been heard) in this kind of test, so the test takers will not know what to listen for before they hear the text, and there is the problem of the candidates having to hold in their heads four or more alternatives while listening to the passage and, after responding to one item, of taking in and retaining the alternatives for the next item (Hughes, 2003). Questions may also be printed on paper, and so test takers may have time to scan the items first, which means they should be able to listen for specific information. Multiple-choice items are difficult to write because of their complexity. Brindley (1998) believes that they have a strong method effect. Hanson and Jenson (1994) also claim that they make considerable

processing demands on the test-taker. According to Nissan, DeVincenzi and Tang (1996), they can force test-takers to readjust their interpretation if it does not agree with the options. Wu (1998) found that they favored more advanced listeners, and that misinterpretations of the options led to test-takers selecting the wrong options, and conversely, test-takers selecting the correct options for the wrong reasons.

Listening Cloze

Listening cloze tasks (sometimes called cloze dictations or partial dictations) require the test-taker to listen to a story, monologue, or conversation and simultaneously read the written text in which selected words or phrases have been deleted. The cloze-procedure is most commonly associated with reading only. In its generic form, the test consists of a passage in which every n^{th} word (typically every 7^{th} word) is deleted and the test-takers see a transcript of the passage that they are listening to and fill in the blanks with the words or phrases that they hear (Brown, 2004, p.125). One potential weakness of listening cloze techniques is that they may simply become reading comprehension tasks. Test-takers who are asked to listen to a story with periodic deletions in the written version may not need to listen at all, yet may still be able to respond with the appropriate word or phrase. One can guard against this eventuality if the blanks are items with high information load that cannot be easily predicted simply by reading the passage (Brown, 2004, p.126).

Gap-filling Tests

We can make a gap-filling test by giving test-takers a transcript of a spoken text, with words deleted, then play a recording of the text and ask test-takers to fill in the blanks based on what they have heard. However, the problem here is that test-takers could treat the passage as a normal cloze test and fill in the blanks without listening to the passage at all, in which case it is no longer a listening test at all but perhaps a perfectly good test of reading comprehension or general language ability. Henning et al.(1983) have tried to solve this problem by putting the blanks on

content words, which have a high information load, and which are the least predictable words in the passage, and hence, the most difficult to guess. They call this a listening recall test and report that it enjoys a higher level of reliability, discrimination power, and validity comparing to the other sections on a battery of tests. However, it is difficult to claim that the listening-recall test provides evidence of comprehension because some test-takers would surely listen specifically for the individual words, in which case the test will be a test of word recognition.

Gap-filling on Summaries

When test constructors wish to prevent test-takers filling in blanks without actually understanding the meaning, they can ask them to fill in the blanks on the summary of the passage. This forces them to process the meaning in order to fill in the blanks. Here, test-takers are given a summary of the passage they are going to hear, in which some of the content words have been replaced by blanks. After looking at the summary for a while, test-takers listen to the original passage. They are recommended not to write anything while listening to the passage. Their task is to use their overall understanding of the heard text in order to fill in the blanks. Lewkowicz (1991) found that the technique could be used with a wide variety of texts and topics to develop a large number of items on one passage, with nearly objective marking.

Validity

In investigating validity, we examine the extent to which factors other than language abilities affect test scores. According to Angoff (1988), before 1950 or so, it was generally understood that it was necessary for anyone who was supposed to use a test for an announced purpose to show that the test was in fact useful for that purpose. This was known as test validity. The validity of a test is traditionally defined as the extent to which – or in Garrett's words "the fidelity with which it measures what it purports to measure" (Garrett, 1947, p. 394). Messick (cited in Linn, 1993) defines validity as "an integrated evaluative judgment of the degree to

which empirical evidence and theoretical rationales support the adequacy and appropriateness and actions based on test scores or other modes of assessment (p.13)". Messick's view of validity is referred to the degree to which we are justified in making an inference to a construct from a test score, rather than a property of a test. That is, the behavioral inferences that one can conclude from test scores is of immediate focus (Swaim, 2009). For validating an inference, not only the validation of score meaning is required but also the validation of value implications and action outcomes for particular applied purposes and of the social consequences of their use are important (Messick, cited in Linn, 1993).

According to *Standards for Educational and Psychological Testing* (American Psychological Association, 1985), validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences.

Construct Validity

Construct validation helps to substantiate the extent to which a testee's performance on a particular test can be indicative of his/her underlying competence. Messick (1988) claims that several changes in orientation toward validity have taken place in the last 35 years. The most important of these changes is the articulation of Cronbach and Meehl (1955) and later by Cronbach alone (1971) of the concept of construct validity which reflects a verification of the inferences and interpretations to be drawn from the test scores and a corresponding modification of the instrument or the theory underlying the construct. A second change in emphasis, related to the conception of construct validity, is that there has been a new recognition of that which is to be validated. The earlier view was that it was the test whose validity was being sought in a specific sense and context. Through years, it has become clear that it was the subject's responses to the test, and even more, the inferences and interpretations to be drawn from those responses, that were to be validated. Thus, the responsibility for the validation falls to a considerable degree to the user, or perhaps more generally, to the

person willing to claim that certain inferences may be validly drawn from the test scores.

According to Angoff (1988), construct validity is a major innovation in the conception of validity, already understood as the most fundamental and embracing of all the types of validity. In terms of test validity, the major problem with psychological constructs is that testers cannot take a construct out of a student's brain and show that a test is in fact measuring it. The only recourse is to demonstrate indirectly through some kind of experiment that a given test is measuring a particular construct. Since such demonstrations are always indirect, the result must be interpreted vary carefully (Brown, 2005, p.227).

Regardless of how construct validity is defined, there is no single best way to study it. In most cases, construct validity should be demonstrated from a number of perspectives. Hence, the more strategies used to demonstrate the validity of a test, the more confidence test users have in the construct validity of that test, but only if the evidence provided by those strategies is convincing (Brown, 2000). For example, taking the unified definition of construct validity, we could demonstrate it using content analysis, correlation coefficients, factor analysis, ANOVA studies demonstrating differences between differential groups or pretest-posttest intervention studies, factor analysis, multi-trait/multi-method studies, etc.

One of the most extensively used approaches in the construct validation of language tests is factor analysis (Bachman, 1990). Factor analysis attempts to identify underlying variables, or factors, that explain the pattern of correlations within a set of observed variables (Farhady, 1983a; Oller & Hinofotis, 1980). As methods of statistical analysis progress with advanced computer equipment, a distinction has been made between exploratory and confirmatory factor analysis in the discussion of the traditional term (Lu, 1999). According to Stevens (1996), exploratory factor analysis is used to explore data to determine the number or the nature of factors that account for the covariation between variables when the researcher cannot form a hypothesis about the number of factors underlying the data. In contrast, confirmatory factor

analysis earns strong theoretical and empirical foundation that enables the researcher to invent an exact model to specify factor loadings and correlations.

As discussed previously, listening comprehension is a construct that can be assessed through different test methods. In order to assess L2 learners' listening ability, we need to use a test which enjoys a high level of construct validity. Therefore, the incentive for the present study came from the need for investigating the construct validities of different listening test formats and identifying the most appropriate test for the measurement of this skill. In line with the purpose of this research, the researchers concentrated on three of the most common formats used for testing listening comprehension in Iran, namely, MC questions, gap filling on summaries, and fill-in-the-blank tasks.

Research Question

The present study aimed at providing an answer to the following research question:

Do different test methods bear a statistically significant effect upon the construct validity of the tests used for the evaluation of EFL learners' listening comprehension?

Participants

The participants of this study were 91 undergraduate male and female Iranian students majoring in English translation at two different universities in Iran: Islamic Azad University, North-Tehran branch, and Islamic Azad University, Karaj branch. All of them were attending Oral Translation II classes. The sample was homogeneous with regard to nationality, background, and educational level. Their age range was 21–30.

As indicated by the results of a sample TOEFL test, there were no outliers in the study, and the participants comprised an almost homogenous sample in terms of language proficiency. The sample included three intact classes who were randomly assigned to three groups each taking one format of the listening tests (MC

tests, gap-filling on the summary of the listening text, and fill-in-the-blank test) constructed on basis of each text following a counter-balanced design.

Instrumentation

Ten tests were used in this study:

1. A 110- item multiple-choice test including 30 listening comprehension, 30 grammar, 30 vocabulary, and 30 reading comprehension items used in order to locate the outliers.
2. Three 10-item multiple-choice listening tests constructed on the basis of 3 texts.
3. Three 10-item gap filling tests on the summary of 3 listening passages.
4. Three 10-item fill-in-the-blank tests constructed based on three listening passages.

It is emphasized that the main nine tests were constructed on the basis of three passages, that is, each passage was tested using three test formats.

Procedure

At the outset of the study, a multiple-choice proficiency test consisting of 110 items derived from Nelson and Barron's TOEFL test was administered to 91 university students. After scoring the papers, the item facility (IF) and item discrimination indexes (ID) of the items were calculated. Items with IFs between 0.25 and 0.7 and IDs beyond 0.19 were considered to be acceptable. After discarding the poor items, 67 items were left on the test. Then, the test was rescored on the basis of the remaining items. The results indicated that there were no outliers and the sample was almost homogenous regarding language proficiency. Later, the students were randomly assigned to three groups.

Then three different listening passages were chosen from Cambridge IELTS 3 and 4 to construct the required tests. In order to make sure that the level of difficulty and the content of the texts were appropriate, the readability levels of the listening texts were

estimated using the Flesch Reading Ease readability score. A careful inspection of the passages also revealed that they included no difficult words to hinder the subjects' comprehension. The passages were each 5 minutes long on average and were played once. Each listening passage was divided into two parts, and there was half a minute pause for students to answer each part. As mentioned before, the tests were administered following a counter-balanced design. All the participants received all the three types of listening tests but in different orders. The researchers administered the tests in the following order:

Group 1	**Group 2**
Passage 1: Listening Summary Cloze	Passage 1: Multiple-choice test
Passage 2: Fill-in-the-blank task	Passage 2: Listening Summary Cloze
Passage 3: Multiple-choice test	Passage 3: Fill-in-the-blank task
Group 3	
Passage 1: Fill-in-the-blank task	
Passage 2: Multiple-choice test	
Passage 3: Listening Summary Cloze	

The exact word method was used for scoring the papers. The participants received one point for each correct response, with no points awarded to incomplete responses. There was no penalty for incorrect responses. The face and content validity of the constructed tests were confirmed by two testing experts, and the reliabilities of all the tests were estimated through the KR-21 formula.

The design of this research was a counter-balanced design in which the three groups received the three different formats of listening tests but in different orders. Finally, all the results were subjected to a series of statistical tests to study the effects of the three test formats on the construct validity and level of difficulty of the tests.

Data Analysis and Results

As mentioned before, a 110-item proficiency test was given to all the participants in order to locate the outliers among them.

The results indicated that there were none. Table 1 shows the descriptive statistics for the homogenizing test.

Table 1.
Descriptive Statistics for the Proficiency Test

	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
experimental	90	47.00	13.00	60.00	32.8444	1.06093	10.06486	101.301
Valid N (listwise)	90							

The reliability of the proficiency test calculated through the KR-21 formula was equal to 0.85, which was quite satisfactory. Then, the nine main tests (multiple-choice, fill-in-the-blank, and listening summary cloze tests) of the study were given to the three groups. The related descriptive statistics are given in Tables 2, 3, and 4.

Table 2
Descriptive Statistics for the Multiple-Choice Tests

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
MC1	33	6.00	4.00	10.00	7.5152	1.75216	3.070
MC2	27	7.00	3.00	10.00	7.5926	1.64689	2.712
MC3	31	6.00	4.00	10.00	7.6129	1.45321	2.112
Valid N (listwise)	27						

Table 3
Descriptive Statistics for Listening Summary Cloze Tests

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Cloze1	33	8.00	2.00	10.00	6.5758	2.20837	4.877
Cloze2	27	7.00	2.00	9.00	5.2222	1.73944	3.026
Cloze3	31	7.00	3.00	10.00	5.9677	2.07338	4.299
Valid N (listwise)	27						

Table 4
Descriptive Statistics for Fill-In-The-Blank Tests

	N	Range	Minimum	Maximum	Mean	f	Variance
Fill-in-the-blank 1	33	7.00	1.00	8.00	5.0606	1.76670	3.121
Fill-in-the-blank 2	27	5.00	3.00	8.00	5.1111	1.36814	1.872
Fill-in-the-blank 3	31	6.00	2.00	8.00	5.0968	1.32551	1.757
Valid N (listwise)	27						

The descriptive statistics of each test format for all the participants were also calculated. Tables 5, 6, and 7 illustrate the results.

Table 5
Descriptive Statistics for Multiple-Choice Tests

	N	Minimum	Maximum	Mean	Std. Deviation
MC	91	3.00	10.00	7.5714	1.60653
Valid N (listwise)	91				

Table 6
Descriptive Statistics for Listening Summary Cloze

	N	Minimum	Maximum	Mean	Std. Deviation
Cloze	91	2.00	10.00	5.9670	2.08407
Valid N (listwise)	91				

Table 7
Descriptive Statistics for Fill-In-The-Blank Tests

	N	Minimum	Maximum	Mean	Std. Deviation
Fill-in-the-blank	91	1.00	8.00	5.060	1.83774
Valid N (listwise)	91				

The reliability indexes of the tests were also calculated as a precondition for their construct validity. The Cronbach's alpha reliabilities of the tests (each three taken as one for each format) are given in Table 8.

Table 8
Cronbach's Alpha of the Three Tests

Test	Cronbach's alpha	N of items
MC	0.85	30
Cloze	0.76	30
Fill-in-the-blank	0.66	30

As shown in Table 8, all the three tests enjoyed acceptable to good levels of reliability. If a test enjoys strong internal consistency, most measurement experts agree that it should show only moderate correlation among items. For exploratory purposes 0.60 is accepted; for confirmatory purposes 0.70 is accepted; and 0.80 is considered good (Garson, 2010). Here, the multiple-choice test had the highest level of reliability and the fill-in-the-blank the lowest.

In order to investigate the construct validity of the three measures used in this study (MC, gap filling on the summary of the listening text, and fill-in-the-blank), the scores of the subjects on these measures were subjected to a factor analysis. It is worth mentioning that the purpose of all these three tests was to measure the listening comprehension of the subjects. Table 9 shows the result.

Table 9
Total Variance Explained by Factor Analysis

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.976	65.865	65.865	1.576	52.533	52.533
2	.653	21.782	87.647			
3	.371	12.353	100.000			

Extraction Method: Principal Axis Factoring.

In order to determine how many factors to extract, the eigenvalue-greater-than-one was selected as the extraction rule. This rule suggests that those factors whose eigenvalues (sum of squared loadings) are less than unity will be excluded from the analysis. As Table 9 reveals, only one factor with eigenvalue more than one (1.97) was extracted, and all the tests loaded on the same underlying factor, that is, factor 1 (Table 10).

Table 10
Results of Factor Analysis

Factor Matrix[a]

	Factor
	1
MC	.938
Cloze	.598
Fill-in-the-blank	.582

Extraction Method: Principal Axis Factoring.

The results of factor analysis, as illustrated in Table 10, revealed that almost all the measures had high loadings on Factor 1 (i.e., had high correlations with it). The highest belonged to the multiple-choice test (.93), and the lowest to the fill-in-the-blank task (.58).

Mean Comparison

In order to compare the participants' mean scores on the three listening test types, a one-way analysis of variance (ANOVA) was performed. The results are given in table 11.

Table 11
ANOVA Results for Comparing the Means on the Three Test Formats

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	288.615	2	144.308	47.258	.000
Within Groups	824.484	270	3.054		
Total	1113.099	272			

The obtained F ratio (47.258) was significant at p<.000 level suggesting that there was at least one significant difference between the means of at least one pair of the groups compared. However, it was necessary to perform a Tukey's HSD test in order to identify the exact location of the difference (s), that is, to find out which two means were significantly different from each other. Table 12 provides the results of the Tukey's test.

As shown in Table 13, there was a significant difference between the means of all the three groups. In other words, different methods of testing listening comprehension produced significantly different results. Here, the participants performed better on the multiple-choice test, followed by the gap filling on summary or listening cloze, and finally in the fill-in-the-blank test.

Table 12
Multiple Comparisons of the Means of the Three Listening Tests

(I) VAR00002	(J) VAR00002	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	1.60440*	.25906	.000	.9939	2.2149
	3	2.48352*	.25906	.000	1.8730	3.0941
2	1	-1.60440*	.25906	.000	-2.2149	-.9939
	3	.87912*	.25906	.002	.2686	1.4897
3	1	-2.48352*	.25906	.000	-3.0941	-1.8730
	2	-.87912*	.25906	.002	-1.4897	-.2686

*. The mean difference is significant at the 0.05 level.

A summary of Table 12 is given in Table 13.

Table 13
Summary of Tukey's Results

VAR00002	N	Subset for alpha = 0.05		
		1	2	3
3	91	5.0879		
2	91		5.9670	
1	91			7.5714
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

Discussion

The accurate assessment of all language skills, including listening comprehension, is crucial for empirical research, and it is also important that we evaluate and question the adequacy of our assessment instruments and procedures and consider possible confounds in our measurement of each skill. The findings of this study support the claim that test-takers perform differently across different response types. The data analysis and consequent results indicated that different test formats produce different results, confirming the results from numerous research studies which have

Table 9
Total Variance Explained by Factor Analysis

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Factor	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.976	65.865	65.865	1.576	52.533	52.533
2	.653	21.782	87.647			
3	.371	12.353	100.000			

Extraction Method: Principal Axis Factoring.

In order to determine how many factors to extract, the eigenvalue-greater-than-one was selected as the extraction rule. This rule suggests that those factors whose eigenvalues (sum of squared loadings) are less than unity will be excluded from the analysis. As Table 9 reveals, only one factor with eigenvalue more than one (1.97) was extracted, and all the tests loaded on the same underlying factor, that is, factor 1 (Table 10).

Table 10
Results of Factor Analysis

Factor Matrix[a]

	Factor
	1
MC	.938
Cloze	.598
Fill-in-the-blank	.582

Extraction Method: Principal Axis Factoring.

The results of factor analysis, as illustrated in Table 10, revealed that almost all the measures had high loadings on Factor 1 (i.e., had high correlations with it). The highest belonged to the multiple-choice test (.93), and the lowest to the fill-in-the-blank task (.58).

Mean Comparison

In order to compare the participants' mean scores on the three listening test types, a one-way analysis of variance (ANOVA) was performed. The results are given in table 11.

Table 11
ANOVA Results for Comparing the Means on the Three Test Formats

	Sum of Squares	Df	Mean Square	F	Sig.
Between Groups	288.615	2	144.308	47.258	.000
Within Groups	824.484	270	3.054		
Total	1113.099	272			

The obtained F ratio (47.258) was significant at p<.000 level suggesting that there was at least one significant difference between the means of at least one pair of the groups compared. However, it was necessary to perform a Tukey's HSD test in order to identify the exact location of the difference (s), that is, to find out which two means were significantly different from each other. Table 12 provides the results of the Tukey's test.

As shown in Table 13, there was a significant difference between the means of all the three groups. In other words, different methods of testing listening comprehension produced significantly different results. Here, the participants performed better on the multiple-choice test, followed by the gap filling on summary or listening cloze, and finally in the fill-in-the-blank test.

Table 12
Multiple Comparisons of the Means of the Three Listening Tests

(I) VAR00002	(J) VAR00002	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	1.60440*	.25906	.000	.9939	2.2149
	3	2.48352*	.25906	.000	1.8730	3.0941
2	1	-1.60440*	.25906	.000	-2.2149	-.9939
	3	.87912*	.25906	.002	.2686	1.4897
3	1	-2.48352*	.25906	.000	-3.0941	-1.8730
	2	-.87912*	.25906	.002	-1.4897	-.2686

*. The mean difference is significant at the 0.05 level.

A summary of Table 12 is given in Table 13.

Table 13
Summary of Tukey's Results

VAR00002	N	Subset for alpha = 0.05		
		1	2	3
3	91	5.0879		
2	91		5.9670	
1	91			7.5714
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.

Discussion

The accurate assessment of all language skills, including listening comprehension, is crucial for empirical research, and it is also important that we evaluate and question the adequacy of our assessment instruments and procedures and consider possible confounds in our measurement of each skill. The findings of this study support the claim that test-takers perform differently across different response types. The data analysis and consequent results indicated that different test formats produce different results, confirming the results from numerous research studies which have

demonstrated that the methods which are used to measure the language ability influence performance on the related tests (Bachman, 1990).

Based on the results of the factorial analysis in this study, only one factor with eigenvalue more than one (1.97) was extracted, and all the tests loaded on the same underlying factor, that is, factor 1. From among all the three test formats, the multiple-choice test had the highest loading on this factor suggesting that it was capable of fulfilling many of the requirements of a suitable test in terms of construct validity. It also enjoyed the highest level of reliability. The comparison of the means of the students by a repeated measure one-way ANOVA and a follow-up Tukey's test demonstrated that there was a significant difference between the means of each pair of tests. Therefore, it was concluded that students performed best on the MC-test and weakest on the fill-in-the-blank test.

However, in practice, it is doubtful if better performance or higher scores on a test can account for the supremacy of that test. In recent years, MC-tests have been the targets of attacks from many practitioners regarding their low potential for the measurement of language proficiency, particularly in communicative contexts. Nevertheless, the results of this study cast some suspicion on such accusations and demand a revision of the capabilities of this test format in the assessment of the listening skill.

The common testing device in Iran's educational system is the multiple- choice test. This method is used because it is seen as being cheap, efficient and reliable. Therefore, students have greater familiarity with this type of test than with other test methods. The results of the present study also suggest that the low scores on the listening summary cloze and fill-in-the-blank task may be reflective of the participants' unfamiliarity with such test formats. Furthermore, of all the question types, fill-in questions may be the most feared because they do require students to produce language while in multiple-choice questions students select a response rather than construct their own, which may lower test anxiety for test-takers, allowing them to make the best use of their knowledge.

The findings of this study may help listening scholars to have more confidence in the measures they use and consequently the results and conclusions they draw. Besides theoretical contributions, the present research can have some practical applications for different parties involved in the field of ELT. In a language classroom, teachers can benefit from different types of listening tests for learners at different levels of proficiency and with different backgrounds.

Nevertheless, the researchers admit that the domain of application of the results of this study is limited to the effect of using different test formats on the assessment of listening comprehension. They did not try to exert any control over the participants' personal characteristics and could not limit their study to students belonging to a strictly controlled level of language proficiency. They acknowledge that these factors could have played significant roles in determining the findings of this study. Moreover, they also believe that replicating this research with a greater number of items in each test format might produce different results or, otherwise, consolidate its findings to a higher extent.

The Authors

Roya Khoyii is assistant professor at IAU, North Tehran Branch. She holds a PhD in TEFL and has been teaching the related courses at BA and MA levels for 20 years. She has written and translated more than 30 books and papers and spoken in several International conferences. Her research interests include CALL and teaching and testing foreign language skills. The paper she presented at the International Canadian Conference on Education in 2012 at the University of Guelph won the Best Paper Award.

Sara Paydarnia holds an MA in TEFL. She has been teaching in different language institutes for more than five years. Presently, she is also working as a trademark agent in the field of intellectual property in a legal service office providing legal services for various clients in different parts of the world. Her

main research interests include language testing, in general, and the role of tasks in EFL testing, in particular.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Angoff, W. H. (1988). Validity: An evolving concept. In Wainer, H. and Braun, H. I (eds.), *Test validity*. (pp.19-32). Hillsdale, New Jersey.

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

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

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19* (4)*,* 453-476.

Brantmeier, C. (2005). Effects of reader's knowledge, text type, and test type on L1 and L2 reading comprehension. *The Modern Language Journal*, *89* (1), 37-53.

Brindley, G. (1998). Assessing listening abilities. *Annual Review of Applied Linguistics*, *18*, 171-191.

Brown, J. D. (2000). What is construct validity? *JALT Testing & Evaluation SIG Newsletter 4*(2), 8-12.

Brown, H. D. (2004). *Language assessment: Principles and classroom practices*. New York: Longman.

Brown, J. D. (2005). *Testing in language programs*. New York: Mc Graw-Hill.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.). *Educational measurement* (2nd Ed.). Washington, D. C: American Council on Education.

Eykyn, L. E. (1992). *The effects of listening guides on the comprehension of authentic texts by novice learners of French as a second language*. Diss., University of South Carolina.

Farhady, H. (1983). On the plausibility of the unitary language proficiency factor. In J.W. Oller (Ed.), *Issues in language testing research* (pp.11-29). Rowley, Mass: Newbury House.

Foster, P. & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition 18*, 299-324.

Garrett, H. E. (1947). *Statistics in psychology and education*. New York: Longman, Green.

Garson, D. (2010). *Factor Analysis*. From Statnotes: Topics in multivariate analysis. Retrieved June, 12, 2010 from http://www2.chass.ncsu.edu/garson/PA765/factor.htm

Hansen, C., & Jensen, C. (1994). Evaluating lecture comprehension. In J. Flowerdew (Ed.), *Academic listening* (pp. 241-268). New York: Cambridge University Press.

Henning, G., Gary, N., and Gary, J. (1983). Listening recall: A listening comprehension test for low proficiency learners. *System*, *11*, 287-293.

Hughes, A. (2003). *Testing for language teachers*. ***Cambridge: Cambridge*** University Press.

In'nami, Y. & Koizumi, R. (2009). A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing*, *26* (2), 219-244.

Lewkowicz, J. (1991). Testing listening comprehension: A new approach. Hong Kong. *Papers in Linguistics and Language Teaching 14* (1015-2059).

Lin, R. L. (1993). *Educational measurement*. Phoenix: American Council on Education and the Oryx Press.

Lu, C. H. (1999). Application of computer technology: Exploratory/confirmatory factor analysis to promote quantitative research. Paper presented at the *National Conference of American Association of Physics Teachers* (AAPT), San Antonio, Texas.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer and H. I. Braun (Eds.), *Test Validity*. (pp.33-48). Hillsdale, New Jersey.

Nissan, S., DeVincenzi, F., & Tang, K. L. (1996). *An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension.* (TOEFL Research Rep. No. 51). Princeton, NJ: ETS.

Oller J. W., Jr. & Hinofotis, F. (1980). Two mutually exclusive hypotheses about second language ability: indivisible or partially divisible competence. In J.W. Oller & Perkins, K. (eds.), *Research in language testing.* Rowley, Mass.: Newbury House Publishers, Inc.

Rubin, J. (1994). A review of second language listening comprehension research. *Modern Language Journal, 78* (2), 199–221.

Shin, S. (2008). Examining the construct validity of a web-based academic listening test: An investigation of the effects of constructed response formats in a listening test. *The Spaan Fellowship Working Papers in Second or Foreign Language*. : 95-129. English Language Institute. University of Michigan.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Swaim, V. S. (2009). *Determining the number of factors in data containing a single outlier: A study of factor analysis of simulated data.* Unpublished dissertation. Louisiana State University and Agricultural and Mechanical College.

Teng, H. C. (1998). The effect of text and question type on English listening comprehension. *English Teaching, 23* (19), 5-18.

Wainer, H. Braun, H. I. (1988). *Test Validity*. New Jersey: Lawrence Erlbaum Associates.

Wu, Y. 1998. What do tests of listening comprehension test? A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing*, *15*(1), 21-44.

Ying-hui, H. (2006). An investigation into the task features affecting EFL listening comprehension test performance. *Asian EFL Journal*, *8*(2).

روش آزمونی و روایی آزمونهای شنیداری

رویا خویی
سارا پایدار نیا
دانشگاه آزاد اسلامی واحد تهران شمال

در حیطه آزمون سازی و اندازه گیری، ارزیابی مهارت های شنیداری، علیرغم اهمیت بالای آن ، کمتر مورد توجه و درک متخصصین قرار گرفته و از پیشرفت کمتری نسبت به آنها برخوردار بوده است. اهمیت این حیطه بیشتر ریشه در تأثیرات بازگشتی آن روی تمرین های کلاسی دارد. با توجه به نقش شایان توجه آزمون های شنیداری در اندازه گیری معلومات زبانی فراگیرندگان، از آن ها انتظار میرود که از روایی سازه ای بالایی برخوردار باشند. پژوهش حاضر به منظور بررسی روایی سازه ای سه نوع متفاوت از آزمون های شنیداری شامل آزمو ن های چهار جوابی، پر کردن جای خالی در خلاصه متن، و پر کردن در متن دست نخورده با هدف اندازه گیری مهارت شنیدن فراگیران زبان انگلیسی به عنوان یک زبان خارجی انجام شد. برای دستیابی به هدف تحقیق، سه متن با سطح دشواری نسبتا یکسان برای ساخت نه آزمون شنیدن برای درک مفهوم مورد استفاده قرار گرفتند. سپس این آزمون ها بر اساس طرح موازنه ای به 91 زبان آموز همگون داده شدند. تحلیل آماری داده ها نشان داد که آزمون چهارجوابی از بالاترین میزان روایی سازه ای برخوردار بود. بعلاوه، بعد از مقایسه ی نتایج آزمون ها با استفاد هاز آزمون ANOVA یکطرفه، نتیجه گرفته شد که آزمون پر کردن جای خالی در متن دست نخورده از همه ی آزمون ها دشوارتر و آزمون چهارجوابی از همه ی آن ها برای شرکت کنندگان در پژوهش آسان تر بود.

کلید واژه ها: روایی سازه ای، تحلیل عاملی، فعالیت پر کردن جای خالی، پر کردن جای خالی در خلاصه متن (آزمون شنیداری cloze خلاصه، آزمون چهار جوابی