The Journal of Applied Linguistics Vol. 1, No. 1
February 2008

Task-Based Listening Assessment and the Influence of Construct-Irrelevant Variance

Arshya Keyvanfar
Mojgan Rashtchi
Islamic Azad University, North Tehran Branch

Task-based listening tests such as IELTS require testees to listen to some information on a CD and simultaneously answer the related items. To answer such items, testees are expected to comprehend, analyze, compare and infer pieces of information while listening to the incoming audio material. The present research attempted to investigate whether the two major characteristics of question type and consecutive/simultaneous performance have any impact on the listening performance of Iranian EFL learners. Findings indicated that participants had a significantly better performance when they tackled the tasks consecutively, and performed even better in listening MC items rather than in listening task-based items. The researchers, thus, concluded that task-based listening tests such as IELTS listening module may be under the influence of *construct-irrelevant variance*.

Keywords: Listening Task-Based Assessment, Test Method, Consecutive Performance, Simultaneous Performance, Construct-Irrelevant Variance

For decades, tests of second language reading, writing, and speaking have garnered large amounts of attention, research, and resources in the quest to create reliable, valid, and practical assessments. Listening, however, has traditionally been the forgotten skill when it comes to testing. Buck (2001) claims that it seems in practice, test constructors are obliged to follow their instincts and just do the best they can when constructing tests of listening comprehension. Obviously, this haphazard approach to testing listening presents serious implications for the validity of these assessments. Fortunately, in the last few years the assessment of second language listening has attracted increasing amounts of attention, and a great amount of research has been conducted on the subject.

Numerous researchers (e.g., Buck, 2001 and Richards, 2002), have described the necessity of defining the concept of second language (L2) listening comprehension, yet an adequate definition is still elusive, and there seems to be a general consensus that there is no widely-accepted definition. Part of the problem lies in the fact that because so many different processes and aspects are involved in L2 listening comprehension, providing a global, comprehensive definition may be impossible. Richards (2002) describes how L2 listening varies according to the purpose of listening (e.g. social interaction, exchanging information, academic listening, or listening for pleasure). Also, the process of L2 listening varies with the level of the learner and the context of the situation (Buck, 2001).

Also adding to the difficulty in formulating a widely-accepted definition of L2 listening ability is that the attempt to deconstruct language ability into four skills and distinguishing these skills in terms of channel and mode is considered to be misguided and inadequate. Bachman & Palmer (1996) argue that it is much more useful to see language use being realized as learners performing specific language use tasks. They insist that language skills should not be considered to be part of language ability, but to be the contextualized realization of the ability to use language in real-life situations. That is why they prefer “not to think in terms of ‘skills’, but to think in terms of specific activities or ‘tasks’ in which language is used purposefully” (pp. 75-76). Rather than considering listening to be a “skill”, they see it as a combination of language ability and task characteristics. Thus, when designing and using a test, it is necessary to define these listening tasks in terms of their features as well as the language ability and topical knowledge needed to perform them (Bachman & Palmer, 1996). Bachman and Palmer emphasize the importance of authenticity and interactiveness in creating tests that are construct valid. They argue that test tasks that have characteristics similar to those in the target language use domain (TLU) can provide the interactiveness necessary for authentic testing. In other words, tasks whose completion requires the test-taker to integrate their topical knowledge (and affective schemata) with their language ability are considered to be interactive (ibid.).

The use of authentic tasks should also serve to minimize sources of invalidity in a test (Messick, 1989, 1996). Bachman (1990) describes how a test-taker's performance is influenced by the characteristics of the methods used to elicit the test-taker's language performance. In other words, the way in which these "test method facets" are designed and controlled has a great impact on the test-taker's performance. Numerous studies have provided evidence regarding the effect of test method on test performance (Bachman, 1990; Bachman and Palmer, 1996). Bachman (1990) developed a framework to delineate the specific features or facets of test method that can affect test performance. His framework has five categories of test method facets, which include: facets of the input, the testing environment, the test rubric, the expected response, and the relationship between input and response.

Bachman and Palmer (1996) build on and slightly revise Bachman's (1990) framework. They use the term "task" in place of "test method" and "characteristics" in place of "facets". They state that the task characteristics always affect test scores to some extent. Since it is impossible to eliminate the effects of task characteristics, it is necessary to control them as much as possible so that the tests will be appropriate for the purpose they are designed. The goal, then, is for test developers to understand and be aware of what characteristics can be varied, and how they can be varied to best meet their objectives.

Listening As a Two-Stage Process

Recently, according to a number of scholars, listening has been divided into a two-stage process. Buck (2001) describes it as: “A first stage, in which the basic linguistic information is extracted, and then a second stage in which that information is utilized for the communicative process” (p. 51). He goes on to cite a number of researchers (Carroll, 1972; Clark & Clark, 1977; Rivers, 1966) that have hypothesized this two-stage process, and states that many *s*cholars seem to have arrived at similar conceptualizations of listening comprehension despite using different terminology.

He further goes on to describe the idea of identifiable listening skills, including lower skills that involve understanding utterances at the literal level, and higher order skills like inferencing and critical evaluation. One of the most commonly cited descriptions of listening involves the idea of both top-down and bottom-up processing. Bottom-up processing is the process in which the listener receives the input as sound and begins to interpret the meaning. The top-down processing involves the application of cognitive faculties in the attempt to give meaning to the string of sounds. The mind sets up the expectations and the sound provides confirmation. When enough information arises from both sources, then perception occurs. Thus, both types of processing occur simultaneously (Buck, 2001), although the contribution of both types is not necessarily constant and equal over the course of an utterance. He goes on to state that when the text and words are highly predictable, the listener does not need to rely much on bottom-up processing. When the listener’s expectations are low, however, he or she is forced to use the sensory level bottom-up processing. Because the words and texts are rarely predictable for beginning L2 listeners, they usually have low expectations of the upcoming spoken input, and thus are forced to rely mostly on bottom-up processing. This idea that learners with varying levels of proficiency process aural input differently has also been expressed by O’Malley, Chamot, & Kupper (1989).

While describing how the purpose and TLU situation determine the appropriate construct of listening to be used in the test, Buck (2001) gives a list of recommendations to be used when creating a listening construct, which he refers to as "default listening construct" (p. 113). He suggests that based on this default construct listeners should be evaluated on a variety of texts with a variety of topics. In addition, listening tasks that tap discourse, pragmatic, and strategic competence and require the testees to go beyond literal meaning are particularly appropriate for targeting this construct. Buck (2001) also gives a more formal definition of this default listening construct. This construct incorporates the ability to (a) process extended samples of realistic spoken language, automatically and in real time, (b) understand the linguistic information that is unequivocally included in the text, and (c) make whatever inferences are unambiguously implicated by the content of the passage. This default listening construct is useful, in that it is broad enough to be applied to task-based testing.

Listening and Task-Based/Performance-Referenced Testing

Ellis (2003) following Baker (1989) presents a general framework for classifying different language tests. First, he makes a distinction between *system-referenced tests* and *performance-referenced tests*. System-referenced tests assess knowledge of language as a system, without referring to any particular language use in any particular setting, while performance-referenced tests (also referred to as task-based tests) focus on the ability to use the language in specific contexts. He explains that "whereas system-referenced tests are more construct-oriented, drawing on some explicit theory of language proficiency, performance-referenced tests are more content-oriented… (ibid. 284). "

Then he explains that based on the relationship between "the test performance" and "the criterion performance" both system-referenced and performance-referenced tests can be *direct* or *indirect*. Direct test are constructed using direct samples of the criterion performance or TLU. Hence they are holistic in nature, requiring the measure of proficiency of the learner to be derived from it by obtaining an external rating. Indirect tests, however, are less contextualized and arguably more artificial, he contends. Indirect tests target the linguistic features that are essential in the composite of TLU. Table 1 provides a summary of this framework:

Table 1
Types of language assessment

	Direct (holistic)	Indirect (analytic)
System-referenced	Traditional tests of general proficiency: • free composition • oral interview Information-transfer tests: • information-gap • opinion-gap • reasoning-gap	Discrete-point tests: • multiple-choice • fill-in-the-blank Integrative tests: • cloze • dictation
Performance-referenced	Observing real-world tasks Simulating real-world tasks, e.g. IELTS speaking and writing modules	Measuring specific aspects of communicative proficiency : • tests of specific academic sub-skills, e.g. IELTS academic listening and reading modules • tests of performing specific real-world activities, e.g. IELTS general listening and reading modules

Based on Baker (1989), cited in Ellis (2003)

In ideal situations, where there is opportunity to observe real-life or simulated interactions of language learners with native speakers or other language learners, the evaluator will have the luxury of assessing all four skills at the same time. Nevertheless, in the majority of testing situations, evaluation becomes limited to the discrete measurement of specific aspects of testees' communicative ability as is the case with the IELTS listening component (ibid.).

It can be concluded that authenticity is best achieved when a "performance-referenced test" evaluates learners' communicative ability "directly". As a result, objectivity of evaluation comes only at the expense of sacrificing authenticity which is the outcome of indirect sampling of the criterion performance.

Construct Under-Representation, and Construct-Irrelevant Variance

In order for the results of a test to be generalizable to non-test language situations, the tasks on the test must be sufficiently representative of the TLU domain (Messick, 1996). Creating authentic test tasks (i.e. those that are representative of the TLU domain) is important because of the role authenticity plays in contributing to construct validity (Bachman & Palmer, 1996). Bachman and Palmer define authenticity as "the degree of correspondence of the characteristics of a given language test task to the features of a TLU task" (p. 23). If a test task (including the text used in the task) is authentic and corresponds closely to the TLU task, then it allows test users to generalize the test scores beyond the test itself, to similar non-test language uses, and "this links authenticity to construct validity, since investigating the generalizability of score interpretations is an important part of construct validation" (p. 24). Bachman and Palmer advise that when designing an authentic test task, the critical features of the TLU domain should be defined first, and then the test tasks should be designed so that they have these critical features.

Brown et al (2003) summarize the problems of validity into inadequate content coverage, lack of construct generalizability, sensitivity to performance-referenced tests (to test-method, task type, and scoring criteria), *construct under-representation*, and *construct-irrelevant variance.* Messick (1996) also believes that the main threats to valid assessment are construct under-representation and construct-irrelevant variance. Brown et al (ibid. 77) define under-representativeness as "the problem of generalizing from a few observations to the broad spectrum of real-life performances". As Bachman (1990) elaborates, in a real-life approach to authenticity, proficiency is viewed as the ability to perform particular tasks, which calls for direct rather than indirect testing. Within this framework, language proficiency is seen as the ability to carry out non-test situations linguistically. This naturally raises the concern of representativeness of any particular task, which in turn places restrictions on the generalizability of test results. Bachman notes that in the real-life approach, the particular examples of ability are treated as the construct. Messick (1996) suggests that utmost care should be given to the selection of authentic tasks that provide representative coverage of the content and processes of the construct domain to maximally minimize the threat of under-representativeness. That is, if the authentic tasks are used are sufficiently representative, then the score interpretation of the assessment should be generalizable to non-test language situations.

Brown et al further define construct-irrelevant variance as those "performance characteristics that have little or nothing to do with the students' language ability" (ibid.). In other words, in performance-referenced tests, one has to account for the non-linguistic factors that are in part responsible for success of the testees. Bachman (1990) argues that in task-based language testing it is difficult to distinguish between language ability and actual performance. Skehan (1996, as cited in Brown et al 2003) also notes that carrying out a task cannot be only based on the linguistic ability of the testees and there may be many other factors in the fulfillment of a task.

Factors Affecting Listening Tests Question Type

Among many factors, the role of question type in tasks is an important consideration in L2 listening comprehension testing. Buck (1997) claims that comprehension questions are the commonly accepted practice in listening exams, even though they are unrepresentative of the TLU domain. Nevertheless he maintains that comprehension questions are commonly accepted and achieved “respectability” for no better reasons than that they are similar to content-subject tests, and because students are very familiar with them. Perhaps most importantly, comprehension questions are relatively easy to create, and economical to administer in large-scale testing. But test-taker familiarity, ease of creation, and ease of administration do not alleviate the need to examine how exactly the task questions affect the listener’s comprehension of the text, or the need to examine the assessment for test method effect (Bachman & Palmer, 1996; Bachman, 1990).

Buck (2001) examined the feasibility of writing L2 listening comprehension questions that test higher-level processing, and found that it was very difficult to write such questions. He operationalized the distinction between lower-level processing and higher-level processing, and attempted to create two distinct question types, “those which asked for information clearly stated in the text, and those which required testees to make inferences based on that clearly stated information” (p. 76). The questions did not perform as Buck had anticipated, however. He attributes much of this to the effect of the short-answer format, in that test-takers could give different answers to the same question and thus questions meant to test lower-level processing sometimes had answers that required higher-level processing, and vice-versa. The data suggest that creating short-answer comprehension questions to test learners’ higher level processing skills is a very difficult task, for a number of reasons. Still, Buck feels that with skillful item writing and test piloting, it is possible to do so.

Consecutive/Simultaneous Performance

The researchers of the present study have chosen to coin two terms: *consecutive performance* as opposed to *simultaneous performance*. Consecutive performance can be rightly defined as a task in which first test-takers listen to the excerpt and afterwards are allotted ample time to respond to the questions. This, as you will recall, is in accordance to what Buck (2001) referred to as a two stage process. In contrast, simultaneous performance is the situation in which listening to the excerpt and using the incoming information for the communicative task should occur simultaneously. In other words, no extra time is given after the listening phase for answering the questions. This particular mode of testing can be readily seen in the listening section of the IELTS examination.

IELTS Listening Module: a Performance-Referenced Listening Test

Duration and Format

The Listening Module of IELTS takes around 30 minutes. There are 40 items in which candidates should (1) fill in the blanks with a certain number of words, (2) decide whether a particular piece of information is True, False or Not given, (3) choose the paraphrase of the incoming information, and (4) complete the information on a table. The Listening Module is recorded on a CD or a tape and is played ONCE only. There are four sections with approximately 10 questions in each. Before every section, there is a short introduction, giving information about the speakers, the situation, and its possible subsections. (This is not printed on the question booklet.) Answers are written on the question booklet as candidates listen. When the tape ends, candidates are given ten minutes to transfer their answers to an answer sheet.

Task Types

The first two sections are concerned with social needs. There is a conversation between two speakers and then a monologue. For example – a conversation about travel arrangements or decisions on a night out, and a speech about student services on a university campus or arrangements for meals during a conference.

The final two sections are concerned with situations related more closely to educational or training contexts. There is a conversation between up to four people and then a further monologue. For example – a conversation between a tutor and a student about an assignment or between three students planning a research project, and a lecture or talk of general academic interest. A range of English accents and dialects are used in the recordings which reflects the international usage of IELTS. A variety of questions are used, chosen from the following types:

- multiple-choice
- short-answer
- sentence completion
- notes/summary/diagram/flow-chart/table completion
- diagram labeling
- classification
- matching

Marking and Assessment

One mark is awarded for each correct answer in the 40 item test. A band score conversion table is produced for each version of the listening module which translates scores out of 40 into the IELTS 9-band scale. Scores are reported as a whole band or a half band. Care should be taken when writing answers on the answer sheet as poor spelling and grammar are penalized.

Research Hypothesis

The purpose of conducting the present research can be summarized in the following research hypothesis:

There are differences among three types of questions (task-based simultaneous, task-based consecutive, and multiple-choice questions) in terms of their effect on the listening performance of Iranian EFL learners.

The above research hypothesis was translated into the following null hypothesis:

There is no significant difference among the listening performance of the three groups of EFL learners evaluated through task-based simultaneous, task-based consecutive, and multiple-choice items.

Method

Participants

The participants of the study were 63 male and female Iranian adults, ranging from 23 to 48, matched in three groups of 21. They were selected from an initial number of 102 participants through a standard listening test. The participants were candidates of an IELTS preparation course of 160 hours. Since these courses are intended for individuals seeking to pursue their professional/academic career abroad, the researchers thought that language learners attending such courses can be the best target population for the present study as they usually enjoy an extraordinary amount of motivation and seriousness.

Instrumentation

Two instruments were used to fulfill the objectives of the present research:

- The first was the Brown, Carlsen, Carstens (BCC) Listening Test, which is a reliable and standard criterion for the evaluation of the general listening skill of EFL and ESL learners. As it was mentioned earlier, this test was used to select three groups of participants to function as the samples of the study. To compensate for lack of randomization, which might have resulted in biased samples, matching technique was used. In this technique, as Best and Khan (1989) explain, sets of individuals with identical or nearly identical scores are selected and assigned to two or more groups. In the present study, the participants were matched into three groups.
- Once their homogeneity was assured, three IELTS listening tests were administered. The IELTS listening CD used in the study was taken from the IELTS Specimen Materials 2006 (IELTS Specimen Materials are actual test previously used in the IELTS test).

1. The first IELTS listening test, given to Group One was exactly like the IELTS listening test, where students were to *simultaneously* listen and answer the questions elaborated on earlier in this article.

2. The second IELTS test given to Group Two somewhat differed from Group One in that there were pauses of five minutes to give test takers the time needed to respond to the questions *consecutively*. Thus, a total of four pauses were given to Group Two.

3. The third listening test administered to Group Three was a 40-item multiple-choice test based on the same listening excerpts taken from IELTS Specimen Materials 2006 used in the other two tests. It is worth mentioning that the multiple-choice questions were constructed by the researchers and standardized in a pilot study. The standardization process entailed:

- administration of norm-referenced item analysis procedure during which 40 items were selected from among 52 initial items,
- computation of KR-20 reliability index of r_{xx}= 0.74, and
- concurrent validation of the test with the listening section of a standard TOEFL, reaching an index of r_{xy}= 0.83.

With regards to scoring, all four listening tests were corrected based on an answer key.

Results

The data in this study consisted of three sets of scores which were obtained via the three listening tests of IELTS. The researchers believed that if the two aforementioned variables of question type and timing of answering had any impact on the listening performance of the participants, a comparison of the listening performance of the three groups through Analysis of Variance (ANOVA) would unravel the underlying difference among them. SPSS (the Statistical Package for Social Sciences) software was used to conduct the necessary computations of the study. Table 2 presents the descriptive statistics of the three groups henceforth called the task-based simultaneous (TBS), task-based consecutive (TBC) and multiple-choice (MC) groups:

Table 2
Descriptive statistics of the three groups

	N	Minimum	Maximum	Mean	Std. Deviation
group1 (TBS)	21	13	26	19.52	3.172
group2 (TBC)	21	17	29	22.29	3.273
group3 (MC)	21	18	35	25.71	3.423
Valid N (listwise)	21				

As it can be seen in the table, the three groups having very close variances manifest different means of 19.52, 22.59 and 25.71, respectively.

Graphs 1 to 3 show the histogram of the distribution of each group of participants.

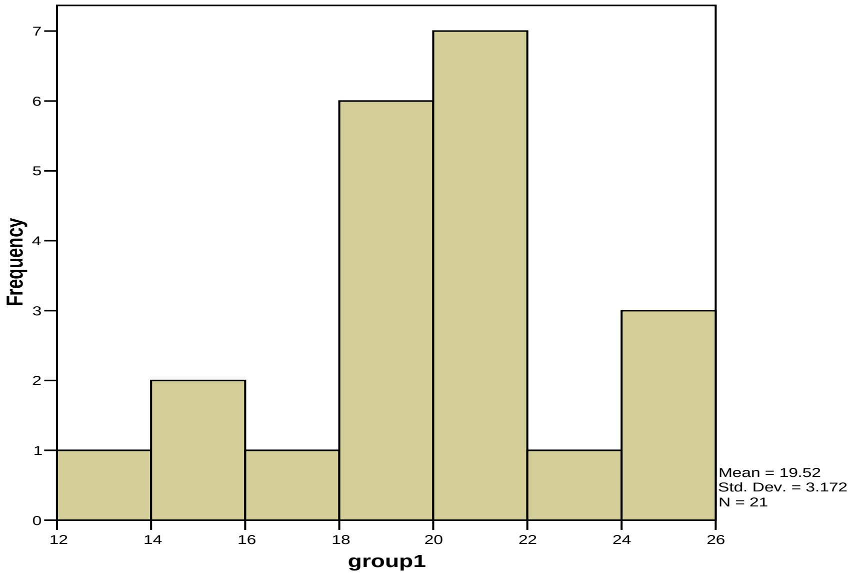

Figure 1. Distribution of the task-based simultaneous group

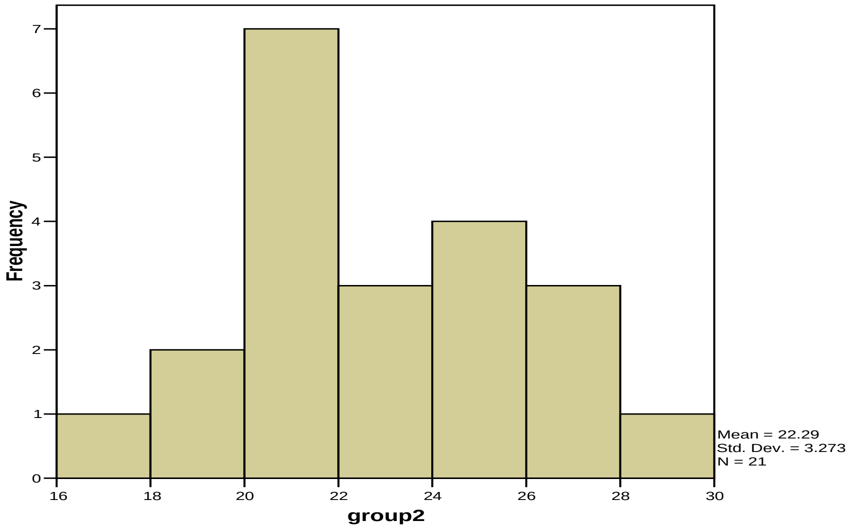

Figure 2. Distribution of the task-based consecutive group

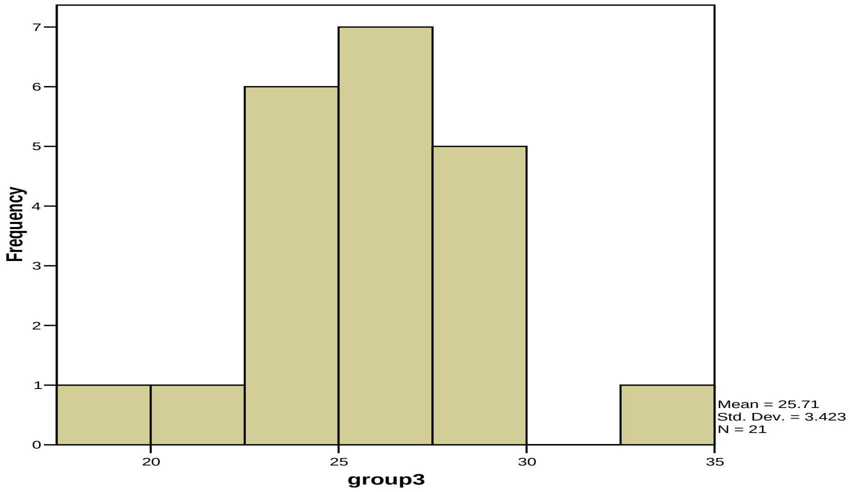

Figure 3. Distribution of the multiple-choice group

To verify the hypothesis of the study, having three groups at hand, the researchers conducted an analysis of variance summarized in the following table:

Table 3
ANOVA for the three groups

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	403.937	2	201.968	18.649	.000
Within Groups	649.810	60	10.830		
Total	1053.746	62			

Since the F-ratio of 18.649 obtained in the analysis of variance of the three groups is significant ($p< 0.01$), it can be concluded that the three groups do not belong to the same population anymore. This finding called for the administration of the post hoc analysis of the Scheffe test. The results are presented in Table 4 and 5:

Table 4
Scheffe post hoc test for the three groups

(I) group	(J) group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-2.762(*)	1.016	.031	-5.31	-.21
	3	-6.190(*)	1.016	.000	-8.74	-3.64
2	1	2.762(*)	1.016	.031	.21	5.31
	3	-3.429(*)	1.016	.005	-5.98	-.88
3	1	6.190(*)	1.016	.000	3.64	8.74
	2	3.429(*)	1.016	.005	.88	5.98

* The mean difference is significant at the .05 level.

Table 5
Means of the three groups

Group	N	Subset for alpha = .05		
		1	2	3
TBS	21	19.52		
TBC	21		22.29	
MC	21			25.71
Sig.		1.000	1.000	1.000

Means for groups in homogeneous subsets are displayed.
a Uses Harmonic Mean Sample Size = 21.000.

The Scheffe post hoc test revealed that the three mean differences of 2.762 between the task-based simultaneous and the task-based consecutive groups, 6.190 between the TBS and the multiple-choice groups and finally 3.429 between the TBC and the MC groups are all significant at .05. As a result, the null hypothesis was rejected with 95% confidence. So it can be claimed that the significant difference observed in the listening performance of the three groups was not due to chance and could be attributed to the variables of question type and timing of answering rather than the listening skill of the three groups.

Discussion

The purpose of the present study has been to investigate the construct validity of IELTS listening module with respect to two possible sources of construct-irrelevant variance, ie question type and timing of answering (consecutive/simultaneous). To this end, three matched groups of IELTS candidates sat for three different listening tests, the first being the regular IELTS listening test (where students were to *simultaneously* listen and answer the questions), the second being the same texts followed by pauses of approximately five minutes to give test takers the time needed to respond to the questions *consecutively* and finally the third being a regular 40-item multiple-choice test of listening.

Having three groups at hand, the researchers used one-way ANOVA to compare the listening performance of the three groups. The obtained F-value and the following Scheffe post hoc test indicated that the mean differences were significant and the null hypothesis could be rejected. In other words, although the three groups were matched in terms of their listening skill and belonged to the same population, when tested under three different conditions, the first group (multiple-choice) outperformed the other two, and the second group (task-based consecutive) outperformed the first (task-based simultaneous).

As it was discussed in the literature, unlike system-referenced general proficiency tests such as TOEFL, performance-referenced tests, even in their "indirect" form which are particularly designed to maximize the reliability necessary for large scale standardized general proficiency tests such as IELTS, evaluate testees' abilities to handle situations which very much resemble real-life interactions. For example, based on a conversation testees have heard between two individuals, they need to comprehend, analyze, compare, contrast, infer, generalize, etc. pieces of information in order to answer the related questions. This means that for testees to be able to answer the questions and in fact accomplish the task, they have to resort to their cognitive, social, and communicative skills besides their linguistic knowledge which is used only to decipher the linguistic code exchanged between the two interlocutors. In the present study, the differences observed between the two task-based groups and the multiple-choice group can be in part attributed to variables other than their listening skill.

Another construct-irrelevant variance could have sourced from an important factor that is usually present in real-life interactions involving listening skill. We are all familiar with the *simultaneity* of the incoming auditory data and all the analyses and decision makings that we have to perform during an interaction. Test constructors in their attempt to achieve authenticity have made this simultaneousness an integral part of listening tasks which naturally increases the difficulty load of task accomplishment. This source of variance can partially account for the significant difference between the means of the task-based consecutive and multiple-choice groups, in which testees were required to tackle each set of items after their related excerpt, and that of the task-based simultaneous group, which had a simultaneous performance.

Yet anther possible source of variance could be related to the *diversity* of tasks that the TBS and TBC groups had to deal with. In task-based listening tests, tasks to be done usually vary from one excerpt to the other, calling for re-adjustment to the new sets of questions. This is obviously not the case with multiple-choice tests where question (item) format remains the same and only the content of each excerpt varies.

Conclusion

This study set out to investigate the influence of construct-irrelevant variance on task-based listening assessment. Two potential variances of question type and timing of answering were examined through comparing the listening performance of three matched groups. Since task-based items required the participants to make analyses, comparisons, and inferences about the listening input, it can be argued that these abilities could have imposed additional sources of variance to the listening skill of the participants. The simultaneity of receiving input and the time of answering in task-based items was another potential source of variance influencing the performance of the participants. Finally, the different nature of tasks, which required the participants to adjust their cognitive strategies to the new task, added yet another possible source of variance to the listening skill of the task-based groups. On the whole, it can be concluded that unless the constructs of listening skill in particular and language ability in general are redefined, task-based listening assessment remains to be highly under the influence of construct-irrelevant variance.

The Authors

Arshya Keyvanfar and Mojgan Rashtchi are associate professors in Islamic Azad University, North Tehran Branch and have been teaching at undergraduate and graduate levels in the areas of English language teaching methodology, testing, and research. They are the authors of several articles and the book *ELT, Quick'n'Easy* which has been taught as the textbook of methodology and TTC courses throughout Iran and some neighboring countries.

References

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

Bachman, L. and Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Baker, D. (1989). *Language testing: A critical survey and practical guide.* London: Edward Arnold.

Best, J. W. & Khan, J. V. (1989). Research in education (6th Ed.). NJ: Prentice-Hall, Inc.

Brown, J. D. and Hudson, T (2002). *Criterion-referenced language testing*. Cambridge: Cambridge University Press.

Buck, G. (1997). The *testing of listening in a second language, Encyclopedia of Language and Education,* Volume 7: Language Testing and Assessment, 65-74.

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

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Messick, S. (1989). Validity, In R. Linn (Ed.). *Educational Measurement* (pp.13-103). Third edition. New York: American Council on Education and Macmillan.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13, pp. 242-256.

O'Malley, J., Chamot, A., & Kupper, L. (1989). Listening comprehension strategies in second language acquisition, *Applied Linguistics*, 10, pp.418-37.

Richards, J.C. (2002). *30 Years of TEFL/TESL: A Personal reflection*, Retrieved November 6, 2006 from http://www.professorjackrichards.com

The IELTS Handbook (2006). University of Cambridge Local Syndicate.

IELTS Specimen Materials Handbook, and Specimen Materials (2006).*University of Cambridge Local Examinations Syndicate*, The British Council, IDP Education Australia.