ISSN 1799-2591
Theory and Practice in Language Studies, Vol. 5, No. 9, pp. 1901-1907, September 2015
DOI: http://dx.doi.org/10.17507/tpls.0509.19

A Comparative Investigation into Response Types in Listening Comprehension Test

Jun Shi
School of Foreign Languages, Shanxi Normal University, Linfen, China

Abstract—**There are two response types, multiple-choice response and constructed response, in testing the communicative language ability. This paper discusses the differences between them, and investigates which type of response proves to be closer to the testees' actual language proficiency level in the listening comprehension test. The research data were collected through the listening comprehension test, and cloze test. The results indicated that constructed response type was more valid in testing the students' language proficiency level.**

Index Terms—**response type, multiple-choice response questions, constructed response questions, listening comprehension test**

I. INTRODUCTION

Language testing has long been a focus of study for its educational and social importance. It may screen testees in their academic and career life; it may offer a ranking of testees based on which further decisions can be made; it may provide empirical evidences for bettering teaching. Some researchers see testing method as "an aspect of content" and claims "format and content are deeply intertwined (Bennet, 1993, p.23)". Due to its powerful educational and social consequences, researchers have been working on this field to ever perfect language testing theories and operations.

In order to provide an appropriate and fair basis for decisions and inferences as to testees' language proficiency, language testing researchers have embarked on an effort in examining factors affecting test performance. It is found that not only the testee's language ability, but also the test method used to measure it and characteristics of testees' have their due impact on test performance (Bachman, 1990, p.111).

In this light, test method has become a research concern. It is regarded as a vehicle that represents test designers' conception of language abilities to be tested, as well as an interface through which testees demonstrate their language proficiency. If test content deals with "what" to test, test method concerns "how" to test (Bachman, 1990, p.111).

With more and more teachers and scholars realize the significant role listening comprehension plays in language learning and communication, they recognize the importance of listening comprehension research, which resulted in an increase in the number of listening activities in students' textbooks and even in methodology texts designed specially for listening.

As for this background, this paper aims to explore the differences between the two response types, i.e. multiple-choice response questions (selected response) and constructed response questions in the context of communicative language testing of listening comprehension, from the perspectives of test validity, reliability, authenticity, and impact on language teaching and learning, and find out which response type could reflect the testees' actual language proficiency level. Thus, future test designers may have a deeper insight as to which response type can better reflect testee's language ability in real language use.

II. LITERATURE REVIEW

A. *Language Testing*

Language testing is basically a sampling of testee's performance on language related activities, from which language ability is judged and assumptions of future language performance are made. There are four parameters essential to the quality of a test, i.e. reliability, validity, authenticity and impact on language teaching and learning, which will be used as main considerations in exploring the two response types.

1. Reliability and validity

Reliability and validity are two major measurement concerns, which are essential to interpretation and use of language tests, thus "the primary qualities to be considered in developing and using tests" (Bachman, 1990, p.24).

Reliability is a major concern in the field of language testing, and educational and psychological measurement in a broader sense. It means the results of a test should be stable and consistent. In other words, no matter who or how to score the test, the results should be stable. If the standards and monitory of test can be well organized, such as the consistent standards of evaluating, the good conditions of the testing environment and the well-chosen testers, the reliability of a test will be convincing. It refers to "the degree to which test scores are free from errors of measurement" (American Psychological Association, 1985, p.19). It is the consistency of results of a test on a particular occasion with those which would have been obtained if the test had been administrated to the same students with same ability, but at a different time and rated again by the same rater or another one (Hughes, 1989, p.19). High reliability is a premise of validity.

By nature, validity concerns "specific use of a test", not "the test itself" (American Psychological Association, 1985, p.9). Bachman (1990:238) claims in the same vein: "in test validation we are not examining the validity of the test content or of even the test scores themselves, but rather the validity of the way we interpret or use the information gathered through the testing procedure." Linking validity to use of tests, American Psychological Association (1985, p.9) has laid out the definition as "the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores". In other words, it refers to the appropriateness of a given test or any of its component parts as a measure of what it purported to measure. A test is said to be valid to the extent that it measures what it supposed to measure. Any test then may be valid for some purposes, but not for others. Therefore, the validity is the centrality of a test.

Therefore, reliability and validity are two major criteria in evaluating a given test.

2. Authenticity

The study of authenticity in language testing has been carrying on for "the past quarter of a century (at least)" (Bachman, 1990, p.300) since its emergence in applied linguistics in the late 1970s when there was a prevailing interest in communicative methodology and teaching and testing real-life language and continues to be a major consideration in language testing.

The definition of authenticity has undergone three major modifications (Bachman, 1991, p.689). It is first defined as direct, in the sense of getting at the ability without going through a representation of the ability. In the second approach, authenticity is understood as similarity to real life. It is difficult and inefficient, if not impractical, to sample infinite real-life language uses in one single test. The third approach sees authenticity as equivalent of face validity, the problem of which lies in that it is too subjective on the part of evaluators.

3. Impact of testing on language teaching and learning

Testing is not an isolate issue from teaching and learning; it exerts influence on the two. The impact of testing on teaching is acknowledged as backwash. A test must try to obtain beneficial backwash, besides validity and reliability, in relation to language teaching and learning.

Beneficial backwash comes from the proper testing procedures that reflect the language construct that has been identified as the teaching objective of a particular curriculum. Also, "it should be supportive of good teaching and, where necessary, exert a corrective influence on bad teaching" (Hughes, 1989, p.2). If the test content and testing techniques are at variance with the objectives of the course, then there is likely to be harmful backwash (ibid: 1). In the framework of communicative language teaching, language testing has to have a corresponding communicative orientation.

In sum, the impact of testing on language teaching and learning needs "attention in the years to come" (Bachman, 1991, p. 679).

B. Listening Comprehension Testing

Listening is an indispensable ability in language communication, especially that through aural-oral channel. On this account, listening comprehension testing is included in most major language testing batteries, and the research into testing listening ability is ever progressing.

1. Nature of listening comprehension

Listening has often been called a passive skill. This is misleading, because listening progress demands active involvement from the speaker. Littlewood (2000) defined the active nature listening comprehension as that "the nature of listening encouraged to engage in an active process of listening for meanings, using not only the linguistic cues but also his nonlinguistic knowledge" and "the active nature of listening means that the learner must be to motivated by a communicative purpose" (p.67).

In communicative listening test, which aims at testing students' communicative competence in real life, the tasks are constructed by the purpose of communication. Therefore, communicative listening tests reflect the active nature of listening comprehension perfectly.

2. Characteristics of communicative listening test

Communicative tests are concerned primarily (if not totally) with how language is used, in communication and tasks are approximate as closely as possible to those facing the students in real life. So communicative listening test, as one part of communicative test is also concerned real-life listening, i.e. what we call authenticity, which includes two aspects: 1) the choice of listening materials and 2) the construction of listening tasks.

C. Response Type

1. Definition of response type

Many researchers have observed that testing method, including response type, is a factor intervening test performance. Response type is an important category of testing method, Bennet (1993, p.27) defines the response as "any physical activity on the part of the student in reaction to the stimulus materials".

Response type, test method in a broader sense, leads to the essential question of "validity", "reliability" and "authenticity" in communicative language testing (Bachman, 1990; McNamara, 2000), as it restricts and controls the nature of test performance elicited from testees.

Response type is not a superficially formal concern, but is closely related to test validity. It is linked with test validity in that it "affects the meaning of test scores by restricting the nature of the content and processes that can be measured" (Bennet, 1993, p.6). It is like a filter through which their language proficiency is demonstrated. The effects of response type used in a given language test may "reduce the effect on test performance of the language abilities we want to measure, and hence the interpretability of test scores" (Bachman 1990, p.156).

Response type also interacts with language teaching and learning. As testing is embedded in the whole educational cycle, it is interlocked with teaching and learning. Testing can serve as a feedback to teaching and learning, which are, in turn, influenced by the tests that explore their effectiveness. In particular, response type, the way in which testees will be required to respond to the test content, which is subject-specific, and its correspondent test-taking techniques involved, can influence classroom practice.

2. Classification of response types

A review of literature on classification of response types unanimously suggests two broad categories. The first category is termed "multiple-choice" (Bachman 1990, p.117), or "selected response" (Bachman 1990, p.129), or "fixed-response" (Bennet, 1993, p.2), which testes have to select the answer from several alternatives that have been "anticipated" (Bennet, 1993, p.2). As this category "characterizes multiple-choice tests" (Bachman 1990, p.129), it will be termed directly as "multiple-choice response questions" (MC for short) in the following discussion.

The other category is termed "free" (Bachman 1990, p.117), or "constructed response" (Bachman 1990, p.129), which means testes need to produce an oral or written response on their part. The author will adopt the term "constructed response questions" (CR for short) in the following part.

3. Strengths and Weakness of two response types

A. Strengths and Weakness of MC Response Questions

MC response question is a wide-used response type in major testing batteries worldwide. It has been the "mainstay of standardized testing programs in the United States" (Bennet, 1993, p.ix).

The benefits of MC response questions are widely recognized as being objective, economical, able to incorporate more test items, and consequently highly reliable (Hughes, 1989, p.59-62). While Hughes argues that the weaknesses of MC response questions can be listed as follows:

(1) this response type tests only recognition knowledge. The performance of MC test may give a quite inadequate picture of testes' productive skills. MC response questions feature a gap to be bridged between knowledge and use; as use is what communicative tests are intended to assess, that gap means that test scores are at best giving incomplete information;

(2) guessing may have a considerable but unknowable effect on test scores. In score interpretation, one can never know what part of the score has come about through guessing;

(3) cheating may be facilitated. The responses to the MC questions (a, b, c, d) are so simple that it is easy to communicate to other testes nonverbally.

B. Strengths and Weakness of CR Questions

CR questions can better assess the ability to use the language in performing authentic tasks. As testees have to construct their own inferences, recognitions, comparisons and summarizations themselves instead of picking from given options, CR questions are suitable to assess high-level abilities to make inferences, to recognizing a sequence, to compare and establish the main idea of a text, and to relate sentences with other items which may be some distance away in the text.

Bennet (1993) voices concerns about reliability of CR questions: "these tasks by their very nature will produce less reliable scores." (p.9) It is true that as CR questions require testes to write other than to choose, the load of spelling and writing may interfere with the construct being tested. Another problem comes from rating. As CR questions are subjective in nature, it might need more considerations to ensure reliability in scoring than the MC questions.

III. Research Methodology

In order to find out which type of answer could better reflect the language proficiency level, the researcher is conducting a test during the English major students. On this account, the material should contain both MC test and CR test, whose difficulty level is more or less the same. And also, the testing criterion should be settled down to judge the testees' language competence. The testing result should be compared with the criterion and tell which one is better.

A. Research Design

In order to clarify how the two different response types reflect the students' communicative ability, the researcher conducted a test. The study involves three sets of investigations:

(1) Exploring the subjects' listening comprehension performance in answering both MC and CR questions;

(2) Testing the authentic ability of the testes;

(3) Finding out which type of question would bring more beneficial effects to the development of students' communicative competence.

B. Subjects

The study was cited the experiment, which conducted among English major juniors in Shanxi Normal University. Totally, 184 students, 21 males and 163 females from six classes participated in the test. The test consisted of the following two parts: about thirty minutes' listening comprehension test, and a fifteen minutes' cloze test. During the test process, 160 students met all our sampling criteria and were selected as our subjects in the end.

C. Materials

1. Listening test materials

There are three most popular tests for the English learners in China: TOEFL (Test of English as a Foreign Language) test, CET (College English Test), TEM (Test for English Majors) and IELTS (International English Language Testing System).

An eminent change in these major testing methods is the modification of response type to be used. TOEFL, a test dominantly using MC questions, will include more CR tasks. Listening materials used in communicative listening tests are spontaneous and authentic, even if not totally they should be approximate as close as possible to the real situation. CET and TEM both include two response types, while the former is applied for non English major learners; the latter is for English majors. As the testees are in the junior grade, have just taken the Test for English Majors-4 (TEM-4), they would have been familiar with the authentic listening test of TEM-4, materials should take other consideration.

"As with all IELTS tests, authentic and general contexts are a central feature of the Listening and Reading tests, and a range of native speaker accents are used to record the lectures and dialogues." (http://www.ielts.org/researchers/research/ielts_reading_and_listening_te.aspx). In this paper, the listening test of IELTS has been chosen as the material for the testing.

Besides, the listening part of IELTS composed of both selected questions and constructed questions. Moreover, it ensures the two types of questions are on the same level, the number of the questions is the same, which is convenient to calculate. So the researcher picked one listening test from the authorized IELTS test publisher, Cambridge IELTS IV. (Cambridge University Press, 2013).

2. Cloze test materials

The cloze test would be used as the testing criterion. As it was easy to construct, administer and score. Reports of early research seem to suggest that it mattered little that passage, were chosen or which words were deleted. The result would be reliable and valid test of candidates' underlying language ability (Hughes, 1989, p.65).

2.1 Related research on cloze test

Oller in his collection work "Issues in Language Research", published in 1983, reprinted four papers in part Ⅲ talking about the pros and cons on cloze testing. Oller (1983, p.48) commented Shohamy's thesis "Interrater and intrarater reliability of the oral interview and concurrent validity with clone procedure in Hebrew" as following: "Elana Shohamy presents a somewhat different perspective on cloze tests. Her results show substantial convergent validity for her application of cloze procedure, from which high reliability can be inferred as well". James Dean Brown also answered the questions of the validity and reliability on cloze test, which were given by Alderson, J. Charles (1995). There is a common understanding that the cloze test can be used to test the convergent language ability.

2.2 The advantages of cloze test

If taking cloze test as the testing criterion to the listening comprehension, there are two obvious advantages.

1) Easy to prepare and score

The cloze test seemed very attractive. Cloze tests were easy to construct, administer and score. Reports of early research seem to suggest that it mattered little that passage, were chosen or which words were deleted. The result would be reliable and valid rest of candidates' underlying language ability (Hughes, 1989, p.65). However, we still need to pat attention to the designing principles. The passage for a cloze test is simple to select and prepare in that we can delete every words. Then we can use the exact-word method to mark. Many people can be scorers. So the cloze test can be widely used in the classroom because it is easy to prepare and score.

2) Superior to discrete point

Item analysis of cloze tests generally indicates their superiority to discrete point. More over, cloze tests and other integrative tests seem to correlate better with each other. It seems that integrative tests are simple better windows through which to view language proficiency than are discrete point tests. Harmer (1983, p.166) mentioned the disadvantages of discrete points tests as that it has been suggested that in real life people do not have to make the same kind of choice, and that the choice actually causes more confusion than it helps students to understand.

Research with integrative testing procedure has suggested a way to operationally define the "efficiency of the internalized expectancy grammar" of a particular learner. While discrete point tests focused on the minute details revealed by the grammatical analyses, integrative tests seek an index of global proficiency. Such an index can be defined operationally by comparing nonnative performance against the criterion of native performance. Therefore, the integrative test, such as the cloze test and dictation, can be employed to test the convergent language ability. Therefore, the researcher are choosing cloze test as the testing criterion to the listening comprehension.

D. Data Collection

The tests were managed in the multimedia classroom of Shanxi Normal University. The listening test lasted about 30 minutes. The cloze test lasted about 15 minutes. Altogether, there are 45 minutes. The listening teacher assisted in the process to introduce the researcher and helped organize the class.

The following is a detailed description of the testing procedures.

1) Introduction

When the class began, the teacher introduced the researcher and told the students that they will do an English listening test. The purpose of the test was to find out their weak points in listening comprehension.

2) Before the test

The teacher played a piece of news to warm-up students, then handed out the testing paper and answering sheet to the students, asked them to fill in the blank all the information required on the answering sheet, including name, gender, age, the length of learning English.

3) On the test

The teacher started the central part of the test, strictly followed the instruction of IELTS test, the tape was played once only for the students. After about 30 minutes, the listening part finished. The teacher gave the students another 15 minutes to answer the following cloze test. In all, after 45 minutes, the teacher asked to collect all the test papers. The researcher was standing by and monitoring the whole process of the test.

4) After the test

After careful scoring of the test papers, the marks of the 160 test papers of the students who have taken part in all the tests were recorded.

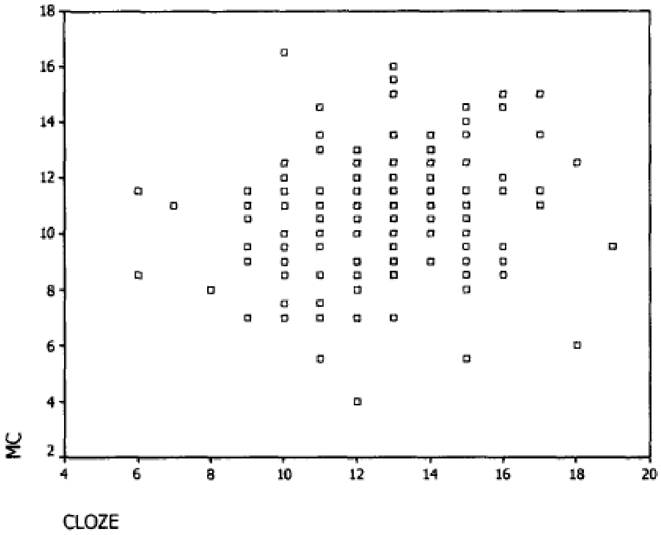
IV. Data Analysis and Discussion

A. *Description of the Test Results*

Previous explanation shows that cloze test could be used as the criteria for English proficiency level test. So if we could like to probe which response type could authentically reflect the person's English Language ability, we could keep an eye on which response type has closer relationship with cloze test.

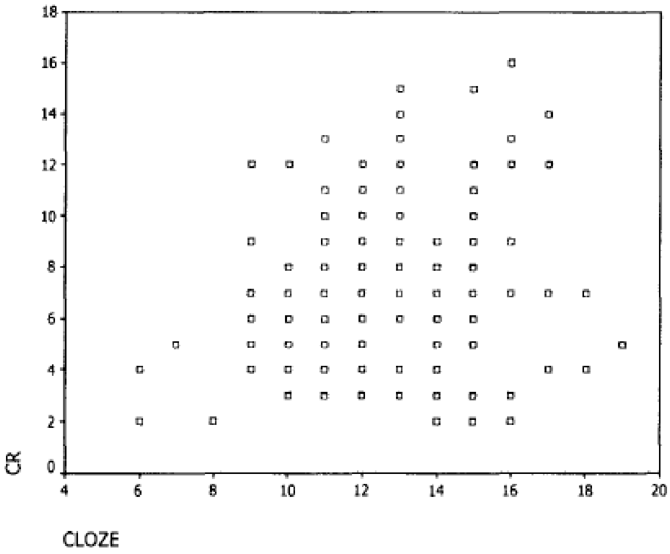
The relationship between the performance on multiple choice test and cloze test and between the performance on constructed response test and cloze test taken by the same group of students could be described graphically with a scatter spot. Each point on the graph or plot represents a particular person's performance on the two separate tests.

The scatter plot will tell whether there is a positive relationship between the variable, if there is any, what kind it is. As shown in the following graph 4-1, we can not visualize the relationship that exists between the two variables because the plots disperse.



Graph 4-1 The Relationship Between Multiple Choice and Cloze Test Results

From following graph 4-2, it is clear that there is a positive linear relationship between the constructed response and cloze test answers, which mean students who scored high on the cloze test tended have high scores on constructed response test.



Graph 4-2 The Relationship Between Constructed Response and Cloze Test Results

It seems that there might be a positive relationship between constructed response test and cloze test. But it is not easy to tell the relationship between multiple choice test and cloze test. From the graph, the relationship between multiple-choice questions and cloze test does not present as clearly as that of constructed response and cloze test. In order to find out which result is closer to that of cloze test, in other words, which response type could reflect the language proficiency level of the testees more efficiently, what we are going to do is to analyze the correlations of the variables.

B. *Variables Correlation Analysis*

Previous explanation shows that cloze test could be used as the criteria for English proficiency level test. Table 4-1 shows the correlations between the multiple choice, constructed response and cloze test. (Resulting from the Pearson product-moment correlation)

TABLE 4-1
CORRELATIONS BETWEEN MC, CR AND CLOZE TEST

Question Type	Cloze Test
Multiple Choice	0.145
Constructed Response	0.310(**)

** Correlation is significant at the 0.01 level (2-tailed).

As for the correlation between MC and cloze test score, the Pearson product-moment correlation coefficient is 0.145 for the association between scores on the MC test and the scores on the cloze test for the 160 subjects. While as for the correlation between CR and cloze test score, the Pearson correlation coefficient is 0.310 for the association between scores on the CR test and the scores on the cloze test for the 160 subjects. The correlation coefficient ranges from -1.0 ~ +1.0, while –0.3 ~ +0.3 means little or no correlation between the variables, and +0.3 ~ +0.7 stands for weak positive correlation. It is obvious that there is a significant positive correlation between the two sets of scores at the 1 percent significance level for a two-tailed test, for the computed value is marked by **.

The comparison above shows that if taking cloze test as the criterion, there is almost no or little correlation between MC and cloze test, while there is definitely correlation between CR and cloze test. Since the correlation is significant, it proves that the CR test could more effectively reflect the language proficiency level than MC test.

V. CONCLUSIONS

The English testing aims to find the way that combines the qualities of validity, reliability and authenticity. From both theoretical and empirical probes, a conclusion can be drawn that CR questions enjoy higher validity. In the aspect of listening comprehension performance in answering both MC and CR questions, the subject who gets high score in MC test might not have high language proficiency level, while the subject who obtains the high score in CR test proves better language proficiency level. In other words, CR proves to possess these three qualities.

Therefore, the study suggests that, English listening tests should adopt a greater proportion of CR questions, enlarge the non-selected-response type.

In this paper, the author picks listening comprehension, an indispensable section in almost all the test batteries, as the research area. Since the time and material is limited, there must be some demerits of this study. And in the future, more materials can be added. Moreover, other fields, like reading, speaking, and writing, may also display different effects of the response type variable.

ACKNOWLEDGMENT

I'm grateful to those who have done me a favour in the process of my article fulfilment. They include my supervisor Professor Lv Hong, the teacher Professor Luo Weihua, Professor Wang Haihua, and Associate Professor Sui Changhong. Moreover, I want to give many thanks to my sister, husband and my two-year lovely son.

REFERENCES

[1] American Psychological Association. (1985). Standards for Educational and Psychological Testing. Washington, DC: American Psychological Association.

[2] Alderson, J. Charles, et, al. (1995). Language test construction and evaluation. London: Cambridge University Press.

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

[4] Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, 25(4), 671-704.

[5] Bennet, R. E. (1993). On the meanings of constructed response. In R. E. Bennet & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Laurence Erlbaum Associates, 1-27.

[6] Cambridge IELTS IV. (2013). London: Cambridge University Press.

[7] Harmer, J. (1983). The practice of English language teaching. London: Longman Group Limited

[8] Hughes, A. (1989). Testing for language teachers. Cambridge: Cambridge University Press.

[9] Littlewood. W. (2000). Communicative language teaching. London: Cambridge University Press.

[10] McNamara, T. (2000). Language Testing. Oxford: Oxford University Press.

[11] Oller, J. W. (ed.) (1983). Issues in language testing research. Rouley, Mass: Newbury House Publisher.

[12] http://www.ielts.org/researchers/research/ielts_reading_and_listening_te.aspx (accessed 20/5/2015).

Jun Shi earned B.A. degree from Shanxi Datong University of English Teaching in 2006. After few years of EFL teaching in Shanxi Normal University Linfen College, she received a M.A. degree from Dalian Maritime University in 2011. Until now, she has been teaching in Shanxi Normal University.