ATLANTIS PRESS

Advances in Social Science, Education and Humanities Research (ASSEHR), volume 181

4th International Conference on Social Science and Higher Education (ICSSHE 2018)

A Comparative Study of Two Parallel Reading Comprehension Tests

Li Huijie*

School of Foreign Languages

Harbin Institute of Technology

Harbin, China

lihuijiehit@163.com

***Abstract*—Multiple-choice question is widely used in reading comprehension tests even though its disadvantages are well recognized. However, there is little literature available about the optimal alternatives to MCQ. Heaton criticizes that MCQ has no communicative feature and maintains that text itself should determine the types of questions. This paper adopts two texts from TEM4 reading comprehension subtest as prototype. Alternative items are written with an aim to make exploration of the effectiveness of multiple types of RC tests from the perspective of communicative testing approach. By comparing the score means of such types as *MCQ, SAQ, matching, and listing* used for the two texts respectively, the study shows that the *matching* type and the *listing* type stand out owning to the high correlation between the result scores and the whole comprehension amount. The research comes to the conclusion that MCQ should not the unique method to test RC ability, and those factors as topic; genre and readability should be taken into consideration at the stage of elicitation method design.**

Keywords—reading comprehension; multiple-choice question; testing method

I. INTRODUCTION

In most cases, reading comprehension (RC) accounts for the largest percentage of an EFL test paper. While the types of RC are of variety, multiple-choice question (MCQ) is the most widely used method in objective testing. According to Heaton [1], however, the chief criticism of MCQ is that 'it does not lend itself to the testing of language as communication'. Challenge to MCQ is accumulating, but a small quantity of research has been conducted on what type can be used as alternatives to MCQ.

Bachman emphasizes the impact of test methods on test performance. In his framework of language task characteristics, the relationship between input and response is highly stressed [2]. According to him, proper design of expected response format is an essential part of reading test construct validation. The Reading Comprehension part of the nationwide *Test for English Majors* Level 4 (TEM4) is usually satisfactory in terms of statistical validation analysis. Nevertheless, it is still worth investigating how much the MCQ type can elicit students' reading proficiency.

Based on the validity analysis of a TEM4, this paper aims to make an exploration of the effectiveness of multiple types of RC tests from the perspective of communicative testing approach. By comparing the scores of such types as *MCQ, SAQ, matching, and listing* used for two texts respectively, the study shows that the *matching* type and the *listing* type stand out owning to the high correlation between the result scores and the whole comprehension amount.

The paper is sponsored by the research project granted by HIT in 2017, Developing the Academic Vocabulary Module.

II. LITERATURE REVIEW

A. Reading process and reading skills

Throughout the developmental history of EFL, reading is regarded as the most important of the four areas, even though reading is given different definitions with varied focuses during the different periods of language learning theories. According to Carrell, "good reading comprehension has long been recognized as important as oral skills, if not more important" [3]. The psycholinguistic perspective, the interactive perspective and schema theory are frequently mentioned currently to explain the nature of reading process [4]. Goodman thinks of reading as "a psycholinguistic guessing game" and defines reading as a psychological process by which the reader reconstructs the message encoded by the writer as graphic display. Reconstruction is not a simple process of decoding, instead, it involves such process as "sampling, predicting, confirming and correcting" [5].

As far as reading skills are concerned, scholars give various taxonomies. Heaton proposed a 14-item reading skills list [1], and they can be classified into three categories: macro skills (mastery of discourse understanding), micro skills (mastery of grammatical and lexical use) and grammatical and lexical knowledge [6]. In a similar way, Hughes [7] proposes 3 levels of reading abilities: macro-skills, micro-skills and straight-forward grammatical and lexical abilities. Alderson [6] summarizes 8 abilities related to reading comprehension, and they are recognition of words, identification, discrimination, analysis, interpretation, inference, synthesis, and evaluation.

B. Types of reading comprehension test

Multiple-choice question (MCQ), short answer question (SAQ), cloze test, true or false (T/F), and matching are the main types of testing reading comprehension ability. Each pattern has its advantages and disadvantages [8] [9]. Hudges [7] summarizes the advantage of MCQ as the perfectly reliable technique, but he points out more disadvantages as follows:

the technique (MCQ) tests only recognition knowledge; guessing may have a considerable but unknowable effect on test scores; the techniques severely restricts what can be tested; it is very difficult to write successful item; backwash may be harmful; cheating may be facilitated.

Heaton designs word matching, sentence matching and pictures and sentence matching types for initial stages of reading; for intermediate and advanced stages, he analyzes the applications of various types, including matching tests, true/false reading tests, multiple-choice items for short/longer texts, completion item, rearrangement items, and open-ended and miscellaneous item [1]. He argues that "the text itself should always determine the types of questions which are constructed".

Hughes [7] puts forward the following possible techniques for RC items writing: multiple choice; short answer; summary cloze; information transfer; identifying order of events, topics, or arguments; identifying referents; guessing the meaning of unfamiliar words from context. He strongly opposes using only one multiple-choice type and claims that the C-test technique should be ruled out because "it is not clear that reading ability is all that it measures".

Alderson [10] suggests using a wide range of methods to test reading comprehension of texts and apply certain reading skills. The methods include discrete point vs. integrative methods; the cloze test and gap-filling tests; multiple choice tests; short answer tests, 'Real-life' methods, information-transfer techniques, and some alternative integrated approaches such as the C-test, the free-recall test, the summary test and the gapped summary.

C. Communicative Language testing

Communicative language testing arises after the practice of notional syllabus, and it characterizes itself more advanced than traditional tests with the traits of authenticity, interactiveness and feasibility [7]. According to Heaton [1], "communicative tests are concerned primarily (if not totally) with how language is used…. Success is judged in terms of the effectiveness of the communication which takes place rather than formal linguistic accuracy".

The Communicative language ability (CLA) model developed by Bachman [2] lays solid basis for communicative testing. The theoretical framework of CLA consists of "both knowledge, or competence, and the capacity for implementing, or executing that competence in appropriate, contextualized communicative language use". The model includes three components: language competence, strategic competence and psychophysiological mechanisms. He categorizes language competence into 4 groups: grammatical competence, textual competence, illocutionary competence and sociolinguistic competence. The former two are regarded as organizational competence and the latter two as pragmatic competence.

Bachman & Palmer use the term 'task' to refer to both target language use (TLU) tasks and test tasks [11]. They summarize the characteristics of communicative language tasks in facets of setting, test rubrics, input, expected response and the relationship between input and response.

D. Construct validity

From the perspective of second language acquisition, testing can be used as an elicitation device mainly owning to its construct definition and validation [12]. According to Bachman [2], construct validity can be viewed as 'definitions of abilities that permit us to state specific hypotheses about how these abilities are or are not related to other abilities, and about the relationship between these abilities and observed behavior'. *The Standards of Educational and Psychological Testing* (1999: 9) mentions that 'validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests'. In the contemporary definition of validation, *construct validity* is a unitary concept and the core of score interpretation [2] [6]. Bachman insists that the 'fundamental issue in construct validity is the extent to which we can make inferences about hypothesized abilities on the basis of test performance'.

Three steps are involved in the measurement of language competence: defining constructs theoretically, defining constructs operationally and qualifying observations. Two dimensions should be embraced in reading construct validation: definition of reading ability and eliciting methods. In proficiency reading test, the theoretical construct is built on experts' definition of reading skills, e.g. the skill lists proposed by scholars [13] [14]. In achievement tests, the reading construct is determined by both teaching syllabus and testing syllabus. The realization of construct validity also needs the safeguard of statistical validation and scientific item writing and scoring [6].

III. BACKGROUND ABOUT RESEARCH: VALIDITY ANALYSIS OF TEM4 READING COMPREHENSION TEST

A. General information

According to the TEM4 testing syllabus, The RC test of TEM4 aims to test whether candidates can meet the following reading ability requirements: able to understand the intermediate-difficulty-level materials published by English-spoken countries; able to get at the gist of the material by understanding the facts and details; able to understand both the literal meaning as well as the underlying meaning by making inferences; understand both the individual sentence meaning and the logical relationship between sentences. There are two sections (items numbered from 66 to 90) in RC part, careful reading and fast reading, with 15 items and 10 items respectively. MCQ is the only type of question presentation. Each correct answer earns one point and the full mark for this part is 25, which weighs 25% of the whole score.

B. Item analysis

Among the 25 items, item 85 ranks first in terms of facility index (.954); the lowest is item 68 (.184). Altogether 6 items get indexes over 0.8, which implies that candidates can easily work out the questions either because of their familiarity with topics or by finding answers through direct literal understanding.

Item 73 has the highest discrimination index (.53), and the lowest one is item 68 (.06). What is noticeable is that the

discrimination index of item 68 is low because of its low facility index, while item 85 is low (.12) because of its high facility index. It can be concluded that item 68 is not valid.

As far as distraction analysis is concerned, item 67, 68, 76, 77, 89 and 90 get lower pass rate, and the balance between the first distractor and the second one is less than 0.2

C. Reliability and validity analysis

The reliability of RC part is proved to be quite good, for the total alpha of the component is .679. If items 67 and 68 are deleted (α values are below 0.2), it rises up to .699. Text A has the lowest α value (.130), which consists with the analysis of item 67 and 68, for they both belong to text A.

Independent T-test between high score group and low score group shows that there is significant difference in means. Through making explorative factor analysis, we see the KMO value is as high as .863 at significant level.000. It means that there are desirable common factors and RC part is suitable for factor analysis. 4 factors are extracted after factor analysis twice, which correspond fasting reading, Text A &Text B, Text D and Text C. It is noticeable that item 81 and Text D belong to the same factor. The reason might be that candidates continue to use the careful reading skill to deal with that question. The confirmative factor analysis further proves the validity of RC test: the scree plot indicates that 2 factors are extracted, which correspond with careful reading ability and fast reading ability.

IV. Research Questions and Hypothesis

While the RC in TEM4 has a good differentiation between careful reading and fast reading, the fact cannot be neglected that all questions are in the form of MCQ. As Hudges points out [7], the technique of MCQ severely restricts what can be tested, for reasonable distractors are not always available. Xu [8] holds the viewpoint that candidates may get marks by ruling out impossible choices and therefore MCQ may encourage test takers to use test wiseness. In RC test, items 67 and 68 are the examples of unsuccessful design.

Here come the research questions as follows:

1) How much can MCQ scores reflect Ss careful reading competence?
2) What can be the optimal alternative types to MCQ?
3) How can the construct validity of RC test be improved?

V. Research design and conduction

The purpose of the research is to make an exploration of the effectiveness of multiple types of RC tests from the perspective of communicative testing approach. By comparing the scores of such types as *MCQ, SAQ, matching, and listing* used for two texts respectively, we can see which one is the best method to elicit Ss reading ability and so that draw some implications for RC testing.

A. Research design

1) Make validity study of RC part in TEM4, and design alternative test methods of Text B and Text C;
2) Choose subjects who can be regarded as TEM4candidates;
3) Divide the sample into 2 groups at random: the control group and the experiment group. Both of the groups are required to write a summary of each text before they handle the items, and the score of summary is used as a variable;
4) The control group does the original items, while the experiment group does the alternative items.

B. The sample: control group and experiment group

1) The independent T-test between the TEM4 candidates and the control group

The control group took 18 items of Text B, Text C and FR, and they got the mean of 11.00. Choose another sample of 26 students at random in virtue of SPSS 13.0, i.e. 0.02% from the whole TEM4 population, and we get their mean of the same 3 sections as 11.88. Make a comparison of their means, and obtain the following data:

TABLE I. THE INDEPENDENT T-TEST RESULT

	N	**Mean**	**SD**	**T value**
Control group	20	11.00	1.97	t=1.08 *p=0.29*
TEM4 sample group	26	11.88	3.52	

Statistics show that there is no significant difference between the two groups. Therefore, we can assume that the control group represents TEM4 candidates.

2) The control group and the experiment group

The control group (n=20) and the experiment group (n=22) derive from two classes of the same study programme. The distribution of the subjects is as follows:

TABLE II. THE DISTRIBUTION OF THE SUBJECTS

	Class 1	Class 2	Total
Control group	n=12	n=8	n=20
Experiment group	n=13	n=9	n=22

They entered the university in Sept. 2005 and study English-related courses including intensive reading, extensive reading, listening and speaking. By comparing the means of their midterm scores of extensive reading, we see there is no significance difference (*p=0.165*) between the groups, which means the two groups are comparable.

C. The procedures of experiment

1) Text analysis of Text B and Text C

Text B is a formal article of argumentation, which argues the reasons of shop-lifting and its consequences. Among the original 5 MCQ items, three ask for detailed information and two for implied information. According to Flesh readability formula, RE=206.84-0.85wl-1.02sl, Text B RE=63.04, i.e. of standard difficulty level. Text C is an extract from a literary work full of modifiers. This article aims to arouse readers' sympathy instead of giving viewpoints. 3 items are given, and

two for details and one for inference. Text C is also of standard difficulty level (RE=69.20), but the RE indicates that it is a little easier than Text B.

2) The alternative items

In the matching form, 4 subtopics and 8 statements are designed for Ss to match. There are 2 items in SAQ form, which test Ss inferential ability. Text C also has 2 substitution forms: one is word matching and the other is evidence listing. In the matching form, 5 word definitions are given and Ss are required to find the corresponding words from the text; the listing type asks Ss to list evidence to prove the woman's loneliness.

3) Paper administration and subjects' taking test

The original test papers and the alternative test papers are handed out to 43 students in the classroom. The control group (n=20) take the test of original types; while the experiment group (n=22) take the alternative types.

4) Data collection and analysis

After getting back the papers, one teacher gives marks all the summary (full mark is 10 of each text) for two purposes: one is to check whether Ss have read carefully and the other is to guarantee the consistence of the summary scores. Mark is given according to Ss understanding of the content; errors concerning spelling, grammar and expression are not taken into consideration. The addition of Ss self-assessing score and the teacher's mark is used as one variable (full mark is 20).

VI. FINDINGS AND DISCUSSIONS

By making independent T-test of summary score means between the control group and the experiment group, we see no significant difference.

The control group deals with MCQ type which reflects the non-reciprocal relationship between input and response; while the experiment group uses more reciprocal methods such as matching, SAQ and evidence listing. By comparing the correlation between the summary score means and the means of each type, we can make a judgment which method can elicit Ss reading ability more properly.

The statistics show that Ss have good understanding of Text B, but the pass rate is lower. That Text B MCQ means have a negative correlation to summary score means indicates that the MCQ type cannot reflect Ss exact understanding amount. When the statement matching type is substituted for MCQ, not only the pass rate becomes higher, but also the correlation is significant. The SAQ type gets a higher pass rate compared to MCQ, but it does not get at significant correlation to understanding amount. The result of data analysis indicates that statement matching should be the best elicitation method for Text B.

While Ss have a lower understanding degree for Text C, they have fairly higher pass rate (66.67%) when handling MCQ. Furthermore, there is no significant correlation between Text C MCQ and summary. It is noticeable that word matching has negative correlation to summary. The reason given by Ss is that they do not need to know the exact meaning of the words because they do not influence understanding. On the contrary, evidence listing takes on perfectly correlation to summary. The coefficient comes to .568 significant at 0.01 level. The findings show MCQ of Text C cannot truly reflect Ss reading ability at all; word matching has no close relation to overall understanding; evidence listing is the optimal substitution. According to Spark [16], whether a literary work can be understood is an important indicator to tell advanced learners from others. The findings suggest that the style should be taken into consideration in item writing.

VII. CONCLUSIONS AND IMPLICATIONS

According to the analysis of the data, we can keep the first null hypothesis and refuse the second hypotheses. As far as Text B and Text C are concerned, there is negative correlation or low correlation between score and the amount of understanding, while the alternative methods take on significant correlation. Therefore, we can come to the conclusion that MCQ scores cannot always reflect Ss careful reading competence and the relationship between input and response does have great impact on the construct of RC test.

Bachman insists that "validity is not simply a function of the content and procedures of the test itself" [2]. TEM4, as a major nationwide test, it is easy to have access to various input material, but how to guarantee the essential factor of validation—scientific elicitation methods—still needs improvement. Making reasonable choice of the relationship between input and response provides clues for the elicitation methods.

Heaton [1] points out that "the test itself should always determine the types of questions which are constructed". His words implies that the relationship does not need to be 'the more complex, the better'. The key point is that the style of the text plays important roles. Such factors as topic, genre and readability should be taken into consideration during the period of elicitation method design.

While this research gets some interesting findings, there exist some limitations in the process. For example, it is not a strict-sense concurrent validity study. In addition, choosing a sample of 26 from a large population of TEM4 candidates means the possibility of sample limitation.

REFERENCES

[1] J. Heaton, Writing English Language Tests. Beijing: Foreign Language Teaching and Research Press, 2000.

[2] L. Bachman, Fundamental Considerations in Language Testing. Oxford: OUP, 1990.

[3] P. Carrell, Metacognitive strategy training for ESL reading. TESOL Quartely (20), pp463-494, 1989.

[4] S. Siberstein, Techniques and Resources in Teaching Reading. Shanghai: Shanghai Foreign Language Education Press, 2002.

[5] K. Goodman, The Reading Process, in Carrell (eds) Interactive Approaches to Second Language Reading, pp.11-12, Cambridge: CUP, 1975.

[6] S. Zou. Language Testing, Shanghai: Shanghai Foreign Language Education Press, 2005.

[7] A.Hughes, Testing for Language Teachers. Beijing: Foreign Language Teaching and Research Press, 2000.

[8] Q. Xu, Communicative English Teaching and Assessing, Shanghai: Shanghai Foreign Language Education Press, 2000.

[9] Z. Wu, Theory and Practice of English Testing, Beijing: Foreign Language Teaching and Research Press, 2002.

[10] J. Alderson, Assessing Reading. Cambridge: CUP, 2000.

[11] L.Bachman & A. Palmer, Language Testing in Practice. Oxford: OUP, 1996.

[12] C. Chapelle, Construct Definition and Validity Inquiry in SLA Research, in L. Bachman& A. Cohen, (eds) Interfaces Between Second Language Acquisition and Language Testing Research, pp 32-33, Beijing: Foreign Language Teaching and Research Press, 2002.

[13] J. Munby, Communicative Syllabus Design. Cambridge: CUP, 1978.

[14] C. Weir, Understanding and Developing Language Tests. London: Prentice Hall, 1993.

[15] J. Alderson, Assessing Reading. Cambridge: CUP, 2000.

[16] R. Spark, Literature, Reading, Writing and ESL: Bridging the Gaps. TESOL Quarterly 19(4), pp703-25, 1985.