

The Feasibility of Measuring Reading Ability by the Format of Short Answer Questions

ZHOU Qingyan

University of Shanghai for Science and Technology, Shanghai, China

In recent years, Bachman's model of the Communicative Language Ability has become increasingly popular. The theories about the measurement of test usefulness put forward by him have aroused great interest and been adopted in the design and evaluation of language tests. Test usefulness incorporates six desirable qualities: validity, reliability, authenticity, interactiveness, impact, and practicality. They constitute an organic system, one quality interacting with another. The test developers' main task is to find an optimal balance among them so as to guarantee the usefulness of language tests. The Multiple Choice Question (MCQ) format is probably the only choice that test developers would like to turn to in their construction of reading tests. Its reliable scoring and efficient administration had once made people regard it as the only feasible as well as practical test format. However, it arouses, with the advancement of language testing theories, more and more skepticism among experts in this field. This paper sets out to propose a new test format—Short Answer Questions (SAQs)—for reading tests. The author chose some non-English majors from Shanghai University for Science and Technology as the subjects and made comparison between the two test formats—MCQs and SAQs—that appeared in their mid-term examination. By means of quantitative and qualitative measures, the author studied the formats in terms of their validity, reliability, interactiveness, practicality, and their impact. Conclusions are drawn and tentative suggestions provided in the hope that the problems with MCQs might be overcome, thus putting forward one more choice for the reform of CET (College English Test for non-English majors)-4/6 that has been put on its agenda.

Keywords: SAQs (Short Answer Questions), MCQs (Multiple Choice Questions), validity, reliability, interactiveness, practicality, impact

Qualities of Tests

The most important consideration in designing a language test is its usefulness, and this can be defined in terms of six qualities: reliability, validity, authenticity, interactiveness, impact, and practicality. These six test qualities all contribute to test usefulness, so that they cannot be evaluated independently of each other. Furthermore, the relative importance of these different qualities will vary from one testing situation to another, so that test usefulness can only be evaluated for specific testing situations. Similarly, the appropriate balance of these qualities cannot be prescribed in the abstract, but can only be determined for a given test. The most important consideration to keep in mind is not to ignore any one quality at the cost of others.

Theoretical Concerns of Reading Tests

The Nature of Reading

No one fully understands the extremely complex process we call reading. Most models of reading are partial in that they are concerned with specific aspects, stages, or modes and do not account for all phases of the reading process and thus cannot be called the most acceptable (Harris & Sipay, 1980, p. 6). There is a need to continue the search and make further study of the nature of reading. In doing so, we should first examine the importance of reading in society and synthesize the different definitions which relevant studies have produced.

Assessment of Reading Ability

Test materials. As we know, input is an essential component in the reading process. Accordingly, the materials for testing, i.e., texts and topics, are important and cannot be neglected in developing a communicative reading test. If the features of test materials correspond in specifiable ways to the reading materials involved in nontest language use, or more specifically, in the teaching program, a testee may fulfill all the required tasks perfectly, and his level of reading ability is said to be relatively high. Otherwise, he may or may not perform well. It depends on his knowledge of the world and of the language, his strategic competence of using reading skills, and other factors connected. In this case, it is difficult to judge his level of reading ability by his performance on a test. Thus, the selection of test materials becomes a critical step in assessing testees' reading ability.

Test formats. In addition to the test materials, the test formats that we employ have an important effect on test performance. The particular way test formats are designed and controlled, and the correspondence between these and the features of language use contexts determine the authenticity of the test and test tasks (Bachman, 1990, p. 112). If we are to understand the ways in which test formats influence test performance, it is necessary to examine various dimensions of test formats first.

Multiple-Choice Questions (MCQs). A multiple-choice test item is usually set out in such a way that the candidate is required to select the answer from a number of given options, only one of which is correct. As an objective-type test format, MCQ has both advantages and disadvantages (Weir, 1988, p. 46).

The advantages of MCQs. MCQs have perfect marker reliability. Unlike those in subjective formats, the personal judgement or bias of the scorer cannot affect candidates' scores. In addition, items can be pre-tested fairly easily so that the difficulty level of each item and the ambiguities in wording of items can be revealed and then be clarified or removed in advance. Most important of all, MCQ can avoid employing other unrelated skills such as writing which might affect accurate measurement of the trait being assessed. Thus, MCQ format is widely used today in national and international standardized test situations.

The disadvantages of MCQs. However, there are a number of disadvantages associated with the use of this format. According to Hughes (2000, pp. 60-62), the disadvantages are as follows:

- (1) The technique tests only recognition knowledge.
- (2) Guessing may have a considerable but unknowable effect on test scores.
- (3) The technique severely restricts what can be tested.
- (4) It is very difficult to write successful items.
- (5) Backwash may be harmful.
- (6) Cheating may be facilitated.

Short Answer Questions (SAQs). Short Answer Questions require the candidates to write down specific answers in spaces provided on the question paper. These answers are normally limited in length either by the space made available to candidates or by controlling the amount that can be written by deleting words in an answer that is provided for the candidates (Weir, 1993). This format has both advantages and limitations.

First, this format measures candidate's language ability directly.

Second, with careful formulation of the questions a candidate's response can be brief and thus a large number of questions may be set, enabling a wide coverage.

Third, activities such as inference, recognition of a sequence, comparison and establishing the main idea of a text, which require the relating of sentences in a text with other items some distance away in the text, can be done effectively through this format.

However, the limitations of SAQ format are not to be ignored. One of them is that it involves the candidate in writing and this may interfere with the measurement of the intended construct.

Another limitation is that the variability of answers might lead to marker unreliability.

Description of the Study

Purpose of the Study

The Testing Committee of College English Test (Band 4/6) has been constantly calling for reform in test formats. So this study is conducted in the hope that a new test format—short Answer Questions (SAQs)—may prove much better a format in testing reading ability, in particular.

Features of the Test

In the fourth semester, all the second-year students in University of Shanghai for Science and Technology (USST) will be given a mid-term examination, virtually a model test of CET-4 (College English Test), to see how well they have prepared for the coming CET-4. In this test paper, there are five parts, including Listening Comprehension, Reading Comprehension I (in the format of MCQ), Vocabulary & Structure, Close and Writing. I deliberately add another part—Reading Comprehension II (in the format of SAQ), in which the four passages were chosen from the original CET-4 test paper conducted in 1997 (Jan.), 1999 (Jan. and June), and 2003 (Sept. in Beijing).

Features of the Test Takers

The subjects of our study are a group of 45 second-year non-English majors from the Departments of Information Technology, Civil Engineering and Mechanical Engineering of USST. Having learned English for two years at college, they possessed a definite level of language proficiency and reading ability. Besides, the 45 students had widely varied topical knowledge due to their varying fields of study, together with the general topical knowledge in such areas as humanities and social science. They have even taken in vain CET-4 at the end of the third semester.

Results of the Study

General Picture of the Results

For the sake of making comparison between Reading Comprehension I (in the format of MCQs) and Reading Comprehension II (in the format of SAQs), the candidates' scores are listed in Appendix III and the frequency of each single score in Figure 1.

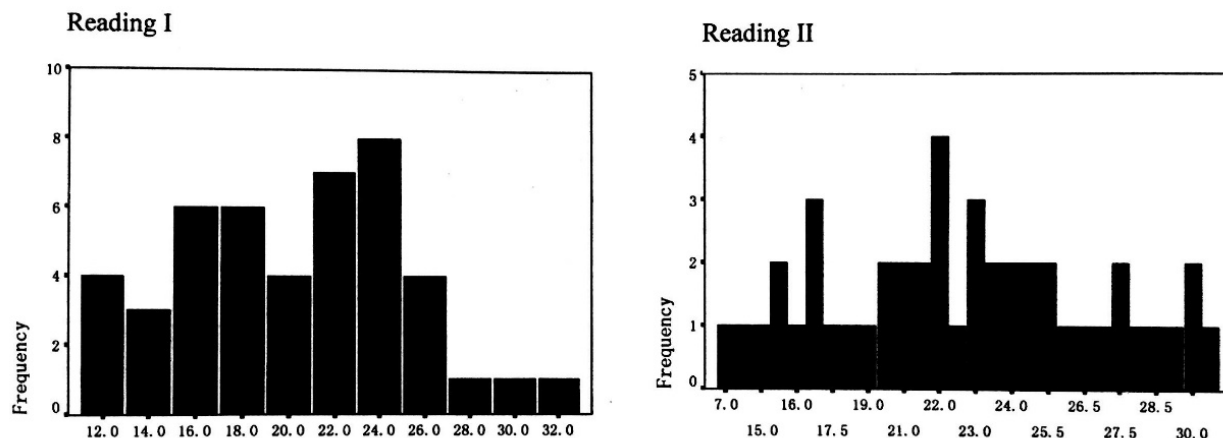


Figure 1. Score Distribution

To see the overall picture of the candidates' performance, the distribution of the total scores, the tendency and dispersion are to be closely examined.

Distribution of the scores. The data presented in Figure 1 visually present the information about how the candidates have performed on this test. Generally speaking, on the condition that the pass score was set at 24, 33.3% of the candidates could pass Reading I, a quite low percentage that is much beyond our expectation. However, the rate for Reading II—37.8%, though compatibly low, is a little higher than that of Reading I. Besides, the distribution of scores obtained in Reading II goes rather closer to the normal distribution.

To examine the details of the performance of the students, we must investigate their group behaviors.

Central tendency and dispersion. According to Brown (1988, p. 66), central tendency refers to the typical behavior of a given group, while dispersion refers to how the performance of those individuals may vary from that typical behavior. Central tendency involves three measures: mean, mode, and median; while dispersion two measures: range and standard deviation (SD). By means of the statistical methods provided by Brown (1988, p. 67), we have computed the central tendency and dispersion of Reading I and Reading II. The data are demonstrated respectively in Table 1.

Table 1

Central Tendency and Dispersion

| Central tendency | | | | |
|------------------|------|--------|-------|------|
| | Mean | Median | Mode | |
| Reading I | 20 | 20 | 24 | |
| Reading II | 22 | 23 | 22 | |
| Dispersion | | | | |
| | Low | High | Range | SD |
| Reading I | 12 | 32 | 20 | 4.97 |
| Reading II | 7 | 31.5 | 24.5 | 5.06 |

This table tells us the mean and median scores candidates got in Reading II (in the format of SAQs) are much higher than that obtained in Reading I (in the format of MCQs). That is to say the candidates' performance on Reading II is much better than that on Reading I.

Besides, the candidates' scores are rather widely dispersed for the range of Reading II (24.5) is much wider than that of Reading I (20) and the Standard Deviation of Reading II is also larger than that of Reading I. In other words, each candidate discriminates each other more noticeably in Reading II.

Summary. On the whole the candidates performed much better on Reading II than on Reading I in that the mean score they got in Reading II is higher, the score dispersion wider, and the standard deviation larger.

Validity

Content validity. According to the National College English Teaching Syllabus (NCETS), the requirement of reading ability in Band 4 includes: First, for a reading material on a general topic with intermediate level, students should be able to identify the main idea as well as supporting details and facts.

The reading skills listed in the NCETS (1999, pp. 165-166) are:

- R1: Understanding the topic and main idea.
- R2: Recognizing primary supporting details.
- R3: Distinguishing between facts and opinions.
- R4: Making inferences.
- R5: Drawing conclusions.
- R6: Skimming to get the gist of a reading material.
- R7: Scanning to find a particular piece of information.
- R8: Guessing the meaning of unknown words.
- R9: Guessing the meaning of unknown phrases through contextual clues.
- R10: Understanding the relationships within sentences.
- R11: Using reference skills.

Our investigation about content relevance indicates the passages chosen for both Reading I and Reading II show a considerable level of content relevance because all the items, either in the format of MCQs or SAQs, are relevant to the reading skills required by the NCETS. That's to say, these two reading tests, though in different test formats, demonstrate fairly little discrimination in terms of content relevance.

As for content coverage, we notice that all the passages in Reading I and Reading II cover very general topics as social science and humanities in order to avoid test bias.

Face validity. As far as the test formats are concerned, MCQs and SAQs are two common devices that testers often turn to elicit candidate's performance of obtaining the message. However, MCQs are increasingly unpopular in the testing world. The problem with MCQs is that the presence of a number of distracters presents test takers with possibilities they may not otherwise have thought of, which may result in an untypical picture of their understanding. The justification for SAQs is that the examiner can interpret candidates' response to see if they have really understood, whereas on MCQ items, candidates produce nothing, and certainly give no justification for the answer they have selected. Therefore, the interest and authenticity of SAQs give it comparatively high face validity.

Correlation with the total score. According to Han Baocheng (2000, pp. 124-125), we calculated their respective correlation coefficient. The data in Table 2 indicate that the scores candidates obtained in Reading II correlate much closer with Total Score II, thus contributing more positively to its validity.

Table 2

Pearson Correlation

| | Reading I/Total I | Reading II/Total II |
|---------------------|-------------------|---------------------|
| Pearson correlation | 0.783 | 0.885 |

Summary. On the basis of our investigation of the content validity, face validity, and the correlation with total score, we realize that the assessment of reading ability by the format of SAQs possesses an obviously higher level of face validity and the candidates' performance correlates much closer with the total score, when content validity is more or less the same.

Reliability

Reliability coefficient. With the support of SPSS, we calculated the reliability coefficient of Reading I (0.5582) and Reading II (0.7485). According to the criterion provided by Lado (1961, p. 332), in terms of objective tests, a good reading test should expect a reliability coefficient in the range of 0.9 and 0.99; while for subjective tests, the reliability coefficient beyond 0.7 will be considered satisfactory. Thus, the reliability of Reading I is not high while Reading II as a subjective reading test is quite reliable.

Marker reliability. While the perfect reliability of MCQs is not obtainable in SAQs, there are ways of making it sufficiently high for test results to be valuable.

To begin with, all criteria levels of performance should be given and agreed on at outset of marking. Only when the points given to each foreseeable answer are agreed on should real marking begin.

Besides, the candidates shouldn't be allowed too much freedom in the way that they respond. In the case of SAQs, it's one possible way to make the test more reliable by limiting the number of words for each response that the candidates have to give.

Above all, marker training deserves our special attention. The marking of subjective tests should not be assigned to anyone who has not learned to mark accurately from past administrations. After each administration, patterns of marking should be analyzed. Individuals whose marking deviates markedly and inconsistently from the norm should not be used again.

Interactiveness

As far as the test format is concerned, the reading test which employs MCQ format is not considered as highly interactive because, for one thing, it encourages the candidates to do an endless array of past paper, forgetting to develop their reading ability in an all-round way. For another, guessing may be facilitated. Worse still, we can never know what part of any particular individual's score has come about through guessing.

However, the reading test which adopts SAQ format will prove more interactive in that it requires the candidates to make use of their language ability more engagingly and directly, thus eliciting vivid demonstration of their language ability that test users are interested in.

Impact

We will focus our attention here on the impact on those individuals who are most directly affected by test use: test takers and teachers.

Impact on test takers. As far as the test format of reading test is concerned, the MCQ format might exert a negative effect on students. In order to get high marks, they're very likely to focus on recognition technique, and then performance on a multiple choice test may give a quite inaccurate picture of those candidates' ability.

Instead, the format of SAQs plays a more positive role in fostering students' productive ability as well as in measuring their reading comprehension. The SAQ format is subjective by nature. The process of producing the short answers requires not only the candidates' correct understanding of the passage but also certain level of expressing in the target language, a result that the skill of reading intends to motivate.

Impact on teachers. In the case of delivering reading courses, if the MCQ format is the only choice, the instruction is very likely to revolve around it. Otherwise, it will risk being challenged to have low test authenticity. Moreover, the test takers' little interest in their teachers' paraphrase and analysis of the reading passage might make the situation even worse. However, the subjective nature of the SAQ format will encourage the test takers to pay attention equally to developing their language proficiency in an all-round way.

Practicality

As we all know, all tests cost money and time to construct, administer, score, and interpret. If the format of Short Answer Questions, instead of that of Multiple Choice Questions, is adopted to test candidates' reading ability in the National College English Test (CET), it possesses the advantage of convenient construction of ideal items, if other aspects of the test are more or less similar.

In the foregoing parts, we've discussed the difficulties we might encounter in writing successful MCQ items. Saving in time for administration and scoring will be outweighed by the time spent on successful test preparation.

As for Short Answer Questions, instead of presenting candidates with four choices, one of which is supposedly better than the other three, one simply asks them a question which requires a brief response, in a few words. In this regard, the format of SAQs is at advantage as far as item construction is concerned.

Conclusions and Suggestions

Conclusions

Our investigation shows that the SAQ format is much better a format than the MCQ format is as far as the test's validity, reliability, interactiveness, impact, and practicality are concerned.

In terms of validity, though these two formats demonstrate more or less the same content validity, the SAQ format shows obvious higher level of face validity in that the format itself is subjective. In order to make response to the item, the candidates have to produce something in the target language instead of making but a choice, a process that is more "real-life", thus contributing positively to its validity. Besides, the correlation coefficient also tells us the SAQ format is preferable.

As for reliability, the correlation coefficient shows the SAQ format is more reliable. The candidates' scores of the SAQs disperse much wider and go closer to normal distribution. But there is one point that needs our elaboration. As far as the marker reliability is concerned, the scoring of SAQs is comparatively more effort-consuming because subjective judgement is required. Therefore, the training of markers should be given enough attention so as to achieve satisfactory marker reliability.

In addition, in response to those SAQ items, more interaction is involved between test takers and the test input for the construction of the short answers is the result of proper understanding and appropriate expression of the related information obtained from the very piece of reading. It is the combination of these two aspects that makes the correct answer possible.

Besides, compared with the MCQ format, the SAQ format exerts impact more positively on both the test takers and the teachers. TOEFL test is a case in point. In their preparation for TOEFL test, the candidates rush between different classes not to avail themselves of the required level of language proficiency but to master the strategy of doing MCQs for it is the only test format in TOEFL. For the same reason, the test format of CET should be varied so as to avoid the problems brought about by employing the MCQ format for too long a time.

Finally, in terms of practical concerns, the SAQ items are relatively easier to construct than the MCQ items which require distracters and distracters are not always available. Sometimes, it may not be possible to find three or four plausible alternatives to the correct answers. Saving in time for administration and marking will be outweighed by the time spent on successful test preparation.

Suggestions

According to our investigation, the SAQ format acts more effectively in assessing candidates' reading ability. By this statement, I do not mean the SAQ format is perfect. Nor do I mean that the test users will encounter no problems with this format. What I mean is the feasibility of measuring reading ability through which format provides us one more option when it comes to test construction. However, in attempting to implement this test format on large-scale tests, such as CET or TOEFL, several points should be borne in mind.

To begin with, a detailed marking scale should be provided. This should specify all acceptable answers and assign points for partially correct responses. This should be the outcome of efforts to anticipate all possible responses and have been subjected to group criticism. For high marker reliability the scale should be as detailed as possible in its assignment of points.

Also, the training of markers should be given much more attention to. Training will help markers to understand the rating scales that they must employ and should prepare them to deal with many problems, including the ones which could not be foreseen when the SAQ items were first designed. Training should give markers competence and confidence; however, it cannot on its own guarantee that markers will mark as they are supposed to. There are many factors which can interfere with a marker's ability to give sound and consistent judgements: problems with the rating scales, time pressures, professional worries, and so on. Even very experienced markers can be affected by these problems. It is the institution's responsibility to design quality control procedures to assure the users of the test that the marks are as reliable as possible.

Moreover, the standard in marking should be strictly kept. Only when all markers are agreed on the marks to be given should real marking begin. Once the rating scales are established, the markers are expected to mark as they were trained to mark instead of challenging or changing the rating scales according to their personal will.

In addition, multiple, independent marking should be employed. All scripts should be marked by at least two independent markers. Neither marker should know how the other has marked a test paper. Marks should be recorded on separate mark sheets and passed to a third, senior, colleague, who compares the two sets of marks and investigates discrepancies.

Moreover, it is essential to word the question in such a way that all possible answers are foreseeable. Otherwise, the marker will be left with a bewildering range of responses.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F., & Palmer, S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Brown, J. D. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.
- Carroll, B. J. (1980). *Testing communicative performance: An interim study*. Oxford: Pergamon Press Ltd.
- Carroll, B. J., & Hall, P. J. (1985). *Making your own language tests*. Oxford: Pergamon Press.
- Chomsky, N. (1965). *Aspects of the theory of syntax*. Massachusetts: The MIT Press.
- Clark, J. L. D. (1983). Language testing: Past and current status-directions for the future. *Modern Language Journal*, 67, 431-443.
- Davies, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell.
- Grellet, F. (1981). *Developing reading skills: A practical guide to reading comprehension exercises*. Cambridge: Cambridge University Press.
- Han, B. C. (2000). *The Statistical Approach in Foreign Language Teaching and Research*. Beijing: Foreign Language Teaching and Research Press.
- Harris, A. J., & Sipay, E. R. (1980). *How to increase reading ability*. New York: Longman.
- Hayes, B. L. (1991). *Effective strategies for teaching reading*. MA: Allyn and Bacon.
- Heaton, J. B. (1975). *Writing English language tests*. London: Longman.
- Heaton, J. B. (1988). *Writing English language tests*. London: Longman.
- Henning, G. (1987). *A guide to language testing*. New York: Newbury House Publishers.
- Hughes, A. (2000). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Kelly, R. (1978). On the construct validation of comprehension tests: An exercise in applied linguistics (Unpublished Ph. D. Thesis, University of Queensland).
- Lado, R. (1961). *Language testing: The construction and use of foreign language tests*. London: Longman.
- Li, G. Y. (1994). *The English Teaching Methodology with Chinese Characteristics*. Shanghai: Shanghai Foreign Language Education Press.
- Nuttall, C. (2002). *Teaching reading skills in a foreign language*. Shanghai: Shanghai Foreign Language Education Press.
- Oller, J. W. (1979). *Language tests at school*. London: Longman.
- Richards, J., & Rodgers, T. (1986). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.
- Ventry, I. M., & Schiavetti, N. (1980). *Evaluating research in speech pathology and audiology*. NJ: Addison Wesley Publishing Company.
- Weir, C. J. (1988). *Communicative language testing*. Great Britain: University of Exeter.
- Weir, C. J. (1993). *Understanding & developing language tests*. London: Prentice Hall International.
- Weir, C. J. (2000). *Studies in Language Testing 12: An empirical investigation of the componentiality of L2 reading in reading for academic purposes*. Cambridge: The Press Syndicate of the University of Cambridge.
- Zou, S. (1998). *The English Language Testing*. Shanghai: Shanghai Foreign Language Education Press.