HNUE JOURNAL OF SCIENCE DOI: 10.18173/2354-1075.2023-0062
Educational Sciences, 2023, Volume 68, Issue 3, pp. 62-67
This paper is available online at http://stdb.hnue.edu.vn

HIGH SCHOOL TEACHERS' TEST ITEM DESIGN: FROM THEORY TO PRACTICE

Trieu Tuan Anh
Faculty of English, Hanoi National University of Education

Abstract. Due to the profound impacts of the national high school graduation exam, high school teachers tend to design test items that are similar to the sample items designed by the Ministry of Education and Training. Employing the principles of test item design by Brown (2004), this study attempted to analyze the quality of the objective multiple-choice grammar and vocabulary test items written by high school teachers. The research drew its data from 24 test papers submitted by teachers from three provinces in Vietnam. The findings revealed that a number of test items violated the principles, having unclear testing purposes or having more than correct answers while no needless redundancies in the stems and distractors were recorded. Also, the participants gave reasons for their difficulties in test design. The paper concluded with several resolutions to address the issue.

Keywords: multiple-choice items, grammar and vocabulary items, item design principles.

1. Introduction

Grammar and vocabulary have been central to language teaching and assessment. Many grammar and vocabulary tests are currently in use, both on large-scale and in classroom assessment (Purpura. J.E, 2013) [1]. In the context of Vietnam, grammar, and vocabulary play a significant role in both teaching and testing. Together with phonetics and reading comprehension, grammar and vocabulary are tested in virtually all exams, the most important of which is the national high school graduation exam. And due to the tremendous effects the graduation exam exerts on the education system, high school teachers tend to design tests that resemble the national high school graduation exam with regard to the test construct so that they can prepare their students best for this exam.

Upon the conduct of this research, the author reviewed and referred to a number of existing studies. Quang (2017) published an article entitled "Evaluation of the quality of multiple choice test bank for the module of Introduction to Anthropology by using the RASCH model and QUEST Software" [2]. He applied classical and modern test theories to analyze and evaluate the difficulty level of the items, the degree of difference among the test questions, and the correlation factors between the test score and the whole score. Canh (2021) applied the item response theory for 2–a parameter model to analyze and evaluate question items in English 1 exam papers in Dong Thap University from 2017 to 2021 [3]. He identified the satisfactory items which met the exam requirements and unsatisfactory ones for further improvement. Phuong et al (2010) presented a conceptual framework and the methodology for a validation study on the interpretation and use of the 2018 university entrance examination English test scores in selecting students for the English Department of the College of Foreign Languages, Vietnam National University [4].

Received May 21, 2023. Revised June 14, 2023. Accepted July 5, 2023.
Contact Trieu Tuan Anh, e-mail address: trieutuananh@hnue.edu.vn

However, despite such an extensive body of literature, to the best of my knowledge, little research was conducted to examine the quality of test items developed by high school teachers at their institutions; hence, this research is an attempt to minimize the gap.

This study is carried out with a view to exploring the quality of test items designed by high school teachers. In order to achieve that goal, this research will seek the answers to the following questions:

1. What are the common issues confronting high school teachers in terms of test item design?
2. What are the contributory reasons for these problems?
3. What solutions are suggested to tackle these issues?

2. Content

2.1. Theoretical background

2.1.1. Test types

Hughes (2003) classified tests into four types according to the test purposes, namely placement, diagnostic, achievement, and proficiency tests [5].

Placement tests are used to place students into groups or classes appropriate to their language levels. Diagnostic tests are often carried out at the beginning of the course to help teachers gain a deep insight into students' language competence, their strengths, and weaknesses, which enables teachers to adopt suitable teaching methods and materials. Achievement tests are conducted at the end of the semester or the academic year to wrap up and decide how much students have achieved. Achievement tests may also take place after several units or lessons to identify how much knowledge students have acquired and how much progress they have made. Proficiency tests are used to determine the language levels of students, and language learners can sit for proficiency tests at any time in their life without any required previous training.

This research has no intention of evaluating test items in all test types; instead, it will only look at the test items in the achievement tests since these are the popular tests used in high schools in Vietnam.

2.1.2. Objective or Subjective tests

Heaton (1990) defined subjective and objective tests as terms used to refer to the scoring of the tests [6]. Objective tests are those with only one correct answer or a limited number of correct answers, so they can easily be marked by any examiner or scored automatically. Subjective tests, on the other hand, require an examiner to give an opinion or a judgment. This study will take a close look at the quality of objective multiple-choice items since this is the most frequently used test type in high schools in Vietnam.

2.1.3. Principles of designing test items

Leading experts in the field of testing and assessment have devised a number of basic rules for designing test items, which are proven to exert significant impacts on the reliability and validity of the tests.

Brown (2004) suggested four major principles of designing multiple-choice items, which are as follows [7]:

a. Design each item to measure a specific objective.

Each item should be designed to test a particular grammatical point. In other words, the testing purposes need to be clear and avoid ambiguity.

b. State both stem and options as simply and directly as possible.

It is of great importance that item writers eliminate needless redundancy from the stems and the distractors.

c. Make certain that the intended answer is clearly the only correct one.

One drawback of multiple-choice grammar items is that with only a minimum of context in each stem, a wide variety of responses may be viewed as correct. Therefore, item writers need to ensure that there is only one correct answer to each question.

d. Use item indices to accept, discard or revise items.

The appropriate selection and arrangement of suitable multiple-choice items on a test can be best accomplished by measuring items against three indices: item difficulty, item discrimination, and distractor analysis.

This study applied the principles of multiple-choice items suggested by Brown (2004) to evaluate items designed by high school teachers. However, due to its limited scope, this study did not look at the item indices since it was impossible to collect the data after the tests were conducted.

2.2. Methodology

This is primarily a qualitative study in which test items are analyzed based on the principles of test design.

The participants of this study include 96 in-service high school teachers from Yen Bai, Ha Nam, and Phu Tho provinces who took part in the training workshops delivered by the Faculty of English, Hanoi National University of Education. All the participants possess the C1 English language certificate and have at least five years of teaching experience.

In terms of data analysis procedures, each group of four participants was required to submit a 15 or 45-minute test that was used at their schools. The items were subsequently evaluated by the researcher who was also the trainer of the course. In the second phase, short interviews and questionnaires were delivered to investigate contributory reasons for their test design problems.

2.3. Findings and discussions

2.3.1. Common issues confronting teachers

In this part, sample items that violate the principles suggested by Brown (2004) will be presented and analyzed. The asterisk signals the correct answer which was provided by the item writers.

a. Ambiguous testing purposes

This is the most common issue in the collected test papers with 39 items having unclear testing purposes. Below are the typical examples of such items.

Example 1: Phong often ..... his bike to school every morning.

A. is driving B. drive C*. drives D. drove

Presumably, distractors (A) and (D) are designed to lure students who do not have a good understanding of basic tenses in English; therefore, they serve as effective distractors. However, the distractor (B) aims to test students' understanding of the subject-verb agreement; consequently, no assessment has been made of tenses in this distractor. Hence, it is unclear whether this item aims to test tenses or subject-verb agreement.

Example 2: Hurry up, Jane! We..... for you.

A. are waited B*. are waiting C. wait D. waiting

Similarly, in this item, the writer may desire to test tenses, so distractors (A) and (C) are appropriate and serve this purpose. Nonetheless, distractor (D) needs to be revised since it tests verb patterns rather than verb tenses like in other options.

Example 3: You will be surprised at how _____ Joe is in French after a year.

A. fluently B*. fluent C. fluency D. influence

It is hard to determine the purpose of this item. While options (A), (B), and (C) test students' ability to distinguish different parts of speech, options (C) and (D) check students' understanding of meanings in context. It means that both grammar and vocabulary are tested in this item.

Another issue that needs to be taken into account is that in some cases, the testing purposes of the items do not match the instructions. It leads to the ambiguity of the testing purposes. A typical example is presented below.

Example 4: Choose the sentence having the same meaning as the given one.

It is no use arguing with him.

A. It is no good to argue with him.

B*. It is no good arguing with him

C. It is no use of arguing with him

D. It is use for him to argue.

As can be seen from the instruction, testees are required to choose a sentence that has the same meaning as the given one. However, all these four distractors virtually have no difference in meaning, and what makes them distinct from the others is the grammar. The focus of this item is, therefore, on the grammar, not the meaning.

b. More than one correct answer

In the national high school graduation exam, students can only choose one correct answer as required in the instructions. Therefore, all the participants did similarly, having their students choose only one correct answer. If test takers had circled two, three, or four options, they would have earned no point for that question. However, the three items below are seen to have more than one correct answer, and this is not a common phenomenon in the collected test papers.

Example 1: Choose the correct answer (A, B, C, or D) to complete the sentences.

The number of car accident deaths is continuing _____.

A. decline B. to decline C*. declining D. having declined

This item has a clear purpose, which is testing verbs as complements or verb patterns, and all four distractors serve this purpose well. The problem is, however, the verb "continue" can be followed by a gerund or a full infinitive without any differences in meaning. As a result, both (B) and (C) are accepted, and alteration needs to be made to either of these distractors to ensure that the intended answer is the only correct answer.

Example 2: Choose the correct answer (A, B, C, or D) to complete the sentences.

A car is more _____ than a bicycle.

A. convenience B. inconvenience C*. convenient D. inconvenient

According to the answer key provided by the item writer, (C) is correct. However, distractor (D) is also possible on the grounds that it fits the sentence grammatically. Besides, as far as meaning is concerned, (D) may be appropriate in certain contexts. For example, a car is more inconvenient than a bicycle in big cities where traffic congestion occurs on a frequent basis. It is impossible for test takers to decide whether the answer is (C) or (D) owing to insufficient contextual clues.

Example 3: Choose the correct answer (A, B, C, or D) to complete the sentences.

I find participating in _____ activities very interesting.

A*. volunteer B. voluntary C. voluntarily D. volunteerism

In this item, test takers are required to select the suitable part of speech. It is evident that distractor (A) is a correct answer to form a compound noun, yet distractor (B), which functions as an adjective to complement a noun, is also possible.

2.3.2. Contributory reasons

After the test items were analyzed, a short interview was conducted to gain a better understanding of the teachers' background and perception of test item design.

First and foremost, all the participants said that they lacked basic knowledge about test item design. 93% of the participants confessed that they used to be taught the subject of testing and assessment at the undergraduate level with a focus on assessing four skills and assessing three language components, but little attention was paid to test item design. In contrast, the rest of the participants revealed that they were not offered the course of testing and assessment at the university at all. In addition, although training courses for in-service teachers were organized by the department of education and training annually, they mostly focused on teaching methodology or information and communication technology in English language teaching. Test evaluation and item design were not included in the training programmes.

Also, the teachers revealed that they designed items with unclear purposes due to their desire to increase the difficulty level of the questions. For example, they were inclined to test tenses and subject-verb agreements simultaneously so as to make the question more challenging to their students. Besides, the majority of the participants ensured there was a correct answer without taking much notice of the other distractors.

2.3.3. Solutions

To address these issues, it is highly recommended that test evaluation and design should be included in the English language teaching subjects at the undergraduate-level programme. Apart from theoretical backgrounds to ELT methodology, teaching methodology, textbook evaluation and classroom management, pre-service teachers need to be taught basic principles of writing test items.

Furthermore, regular intensive training workshops with a focus on test item design should be organized by the department of education and training. In these workshops, they will have an opportunity to cultivate in-depth knowledge about test design, practice writing items, and be given feedback by experts.

3. Conclusions

In this research, it was found out that high school in-service teachers encountered a number of problems in the process of designing test items, namely ambiguous test purposes and double keys. No needless redundancies in the stems and the distractors were recorded in the collected tests. There were several factors leading to such issues. The major cause was that they lacked basic knowledge of test item design, and another reason was their misconception about item design.

This study shows a number of limitations due to its scope. Initially, merely the quality of multiple-choice questions was taken into consideration while the other task types were omitted. Moreover, it did not explore the quality of speaking, reading, listening, or writing test items; instead, only grammar and vocabulary items were examined. For further research, it is suggested that the other task types in the test papers designed by high school teachers should be examined. It is also necessary that further research on the quality of speaking, reading, listening, and writing test items be carried out so that the test papers will become more reliable.

REFERENCES

[1] Purpura, J. E., 2013. *Assessing Grammar*. Cambridge University Press.

[2] Quang, B.N., 2017. Evaluation of the Quality of the Multiple Choice Test Bank for the Module of Introduction to Anthropology by Using the RASCH Model and QUEST software. *Science and Technology Development*, Vol 20, No.X3.

[3] Canh, N.V. & Tac P.V., 2021. Analysis and Evaluation of Multiple-choice Test Items and Test Design: A Study on Application of Item Response Theory. *TNU Journal of Science and Technology.*

[4] Phuong, H.T., Griffin, P., Cuc, N., 2010. Validating the University Entrance English Test to the Vietnam National University: A Conceptual Framework and Methodology. *Procedia Social and Behavioral Sciences 2*.

[5] Hughes, A., 2003. *Testing for Language Teachers*. Cambridge University Press.

[6] Heaton, J. B., 1990. *Writing English Language Tests*. Longman.

[7] Brown, H.D., 2004. Language Assessment: Principles and Classroom Practices. Longman.