Item Analysis of English Grammar Achievement Test

Sandar Lin[1]

Abstract

Item analysis is crucial in improving test items for the purpose of using them again in later tests; it can also be used to remove misleading items in a test. This work aims to analyze test items in terms of item difficulty and item discrimination of the English grammar achievement test.The study was carried out with 56 test takersdoing their studies at Mandalay University of Foreign Languages. One hundred items of cloze test type and ten multiple-choice items were analyzed by using classical test item analysis theory.The study shows that 41% of items have acceptable level of difficulty and 44% have acceptable item discriminating power.

Keywords: Item analysis, difficulty level, discrimination index

Introduction

Test is defined as a procedure for measuring ability, knowledge or performance, (Dictionary of Language Teaching and Applied Linguistics, 2010). Testing refers to an attempt to measure student's learning outcome in teaching-learning process. It is clear that teaching ought to be followed by testing so teachers should have capacity to arrange a good test and analyze it.

Tests are crucial tools in education. Teachers at all levels of education prepare and administer many formal teacher-made tests during the academic year as formative assessment or achievement tests at the end of semester as summative assessment. They have to design their tests to contribute scores that are appropriate to the ability level of the test takers or learners. This means that a test should not be prepared to be either too difficult or too easy for the learners. Thus teachers should plan, revise, and improve their test items by item analyzing.

Generally, there are two different frames of reference in measurement result interpretation: criterion-referenced interpretation and norm-referenced interpretation. Criterion-referenced testing is a test that measures a test taker's performance according to a particular standard or criterion that has been agreed upon. In other words, the test taker must reach this level of performance to pass the test, and a test taker's score is interpreted with reference to the criterion score, rather than to the scores of other test takers (Richards & Schmidt, 2010).The criterion is defined in terms of a specified domain of content. Domain-referenced is most commonly used with achievement tests, in which the domain of content is specified by the syllabus. In this approach, test performance is interpreted as the degree to which the individual has mastered the content in the criterion domain (Popham, 1978). In a criterion-referenced interpretation, an individual's score is interpreted with reference to a pre-determined criterion that is independent of the way test takers perform on the test (Bachman, 2004). In contrast to this, in a norm-referenced interpretation, an individual's score is compared, in terms of its relative standing with those of other individuals in a group that has taken the test.

Item analysis is a post administration process of a test. Item analysis begins after the test has been administered and scored. The quality of a test depends upon the individual items of a test. Brown (1971) mentioned that item analysis has two purposes: (i) to improve the test and evaluation procedures by identifying

[1] Daw, Lecturer, Department of English, Mandalay University of Foreign Languages

imperfect items (ii) to indicate which items or material students have mastered and have not mastered.According to Richards and Schmidt (2010), item analysis in testing is the analysis of the responses to the items in a test in order to find out how effective the test items are and to find out if they indicate differences between high and low ability test takers.

In writing effective test items, for the purpose of test validation, it is necessary to examine whether they are measuring the fact, idea, or concept for which they were intended. Selection of appropriate test items is not enough by itself to ensure a good test. Each item needs to function properly; otherwise, it can weaken the test. Fortunately, there are some rather simple statistical ways of analyzing individual item. This is done by studying the student's responses to each item. This is a highly formalized procedure in assessment. It is a scientific way of improving the individual items and the test as a whole. Therefore, Item Analysis is an important procedure to increase test validity and reliability. It provides a quick, simple technique for evaluating the effectiveness of individual test items.

Item analysis usually provides two kinds of information on items: item difficulty or item facility, which helps teachers decide if test items are at the right level for the target group, and item discrimination, which allows teachers to see if individual items are providing information on students' abilities consistent with that provided by the other items on the test.

Bachman (2002) stated that language assessment takes place in a wide variety of situations, including educational programs. In educational programs, the results of assessment are most commonly used to describe both the processes and outcomes of learning for the purposes of diagnosis or evaluating achievement, or make decision that will improve the quality of teaching and learning and of the program itself.

Statistical analysis is a set of tools to help teachers evaluate and improve the qualities of the tests they use. There are two statistical procedures or theories to better understand the characteristics of individual test items: Classical Test Theory (CTT) and Item Response Theory (IRT). Classical item analysis consists of calculating descriptive statistics for individual items and item response theory consists of a more sophisticated procedure for estimating the statistical characteristics of items.

This study focused on the English grammar semester end test paper and therefore it is an achievement test. The test measured how much the learners had successfully learned with specific reference to a particular course or textbook (Advanced Level, Oxford Practice Grammar by George Yule, 2006). Hence, this kind of testing is criterion-referenced one that measures a learner's performance according to a particular standard or criterion that has been approved. This means that the learner must reach the level of performance to pass the test, and a learner's score is interpreted with reference to the criterion score.

In the second year English grammar paper, ten multiple choice items and one hundred cloze type items are administered. These passages are taken form Macmillan English Grammar in Context, Intermediate level (2008) and Macmillan English Grammar in Context, Advanced level (2012).Test length is three hours and mark allocation is 80. Criteria reference is from 1 to 34 (Grade 1), from 35 to 49 (Grade 2), from 50 to 64 (Grade 3), from 65 to 74 (Grade 4) and from 75 to above

(Grade 5). Students who get Grade 1 and 2 would fail the exam. Students who get Grade 3 and 4 would pass the exam and Grade 5 scorers are regarded as qualifiers.

Aim and Objectives

The aim of this research is to evaluate the teacher-constructed test paper.The objectives are to find out the item difficulty level and the power of discrimination on multiple choice test and cloze test items in the English grammar question paper.

Methods

This study is categorized as descriptive analysis because it is intended to describe the objective condition about the difficulty level and discriminating power of students' achievement test. Besides, this study is called item analysis, because it analyses how well the items of English achievement test can discriminate between the students who have achievedwell and those who have achieved poorly. This study is considered as quantitative research, because the researcher used some numerical data which are analyzed statistically.

This paper adopted the model of classical item analysis in Statistical Analyses for Language Assessment by Lyle F. Bachman (2004). Individual items are typically scored as either right or wrong (R-W). If a test taker gets an item right, his/her score will be '1', while if he/she gets it wrong, his/her score will be '0'. The distribution of test takers' scores on items will thus form a dichotomous scale, comprising '1's and '0's. There are many different procedures for determining item analysis.

The study question paper is that of the semester end test and so it is an achievement test and criterion-referenced (CR) test. Based on Bachman (2004), in this study, just the two item characteristics are focused: item difficulty and item discrimination.

Item Difficulty

Item difficulty or item facility is the proportion of the test takers who answered the item correctly for R-W scoring. There are 56 test takers in this study and so the item difficulty is suitably calculated by including only the upper and lower groups. The proportion ('p-value') for R-W scored items which are dichotomously scaled (0, 1) is determined by the following formula:

Formula of Item difficulty index

$$P(value) = \frac{Ru+ Rl}{Nu+Nl}$$

where P is item difficulty index

Ru = the number of test takers in the upper group who responded correctly

Rl = the number of test takers in the lower group who responded correctly

Nu= Number of test takers in the upper group

Nl= Number of test takers in the lower group

Interpretation of Item Difficulty

If the item is dichotomously scored, the difficulty value of the item is equal to the proportion of persons who answered the item correctly relative to all the test takers tested. To calculate the item difficulty, the total number of test takers who score the item correctly in upper and lower groups is divided by the total number of people in these two groups. The proportion is usually denoted as '*p* value'. The larger the percentages, getting an item right, the easier the item. The higher the difficulty index, the easier the item.

The range is from 0% to 100%. The higher the value is, the easier the item is. P value above 0.90 items are very easy and might be a concept not worth testing. P-value below 0.20 items indicate being difficult and should be reviewed for possible confusing language or the contentneeds re-instruction. Optimum difficulty level is 0.50 for maximum discrimination between high and low achievers. For example, an item answered correctly by 70% examinees has a difficulty index of 0.70. If 90% of a standard group pass an item, it is easy; if only 10% pass, the item is hard or too difficult. Generally, items of moderate difficulty are to be preferred to those which are much easier or much harder.

RANGE = 0-100

easy item = above 70%

moderate (average)= 30-70%

difficult= below 30 %

Discrimination Index

It is a measure of whether an item discriminated between test takers who knew the material well and test takers who did not. That is the ability of the test to differentiate between good (Height Scoring) and poor (Low Scoring) test takers. In addition to knowing how difficult an item is, it is also important to know how it discriminates, that is how well it distinguishes between test takers at different levels of ability. If the item is working well, more of the top-scoring test takers should be expected to know the answer than the low-scoring ones. If the strongest test takers get the item wrong, while the weaker test takers get it right, there is clearly a problem with the item, and it needs investigating.

Formula of Discrimination Index

$$D = \frac{Ru-Rl}{Nu \text{ or } Nl}$$

In which

D = index of discrimination

Ru = the number of test takers in the upper group who responded correctly

Rl = the number of test takers in the lower group who responded correctly

Nu= Number of test takers in the upper group

Nl= Number of test takers in the lower group

Interpretation of Discrimination Index

The discrimination index, D, is the number of test takers in the upper group who answered the item correctly minus the number of test takers in the lower group who answered the item correctly, divided by the number of test takers either in the upper group or in the lower group. Item discrimination index ranges between 0.0 and 1.00. The higher the value, the more discrimination of the item is. A highly discriminating item indicates that the test takers who had high test scores got the item correct whereas test takers who had low test scores got the item incorrect.

Range

(+1)	---0---	(-1)
Maximum size	Zero	Minimum Size

Interpretation

- Good achievement test should have
- 50% of items = above 0.40
- 40% of items = 0.40 to 0.20
- 10% of items = 0.20 to 0.00
- zero % items = negative

Data Organization and Analysis

After scoring the test, the test papers were arranged in order from the highest score to the lowest score to conduct an item analysis (IA) by hand. It needs to be decided how many test takers or learners to include in two groups. In this study, there were 56 test takers and so it was a small group in which the upper and lower one-third are chosen. Total scores of the test takers were entered in Microsoft Excel sheet and they were arranged in descending order, then 18 (one-third) high and low test takers who responded correctly were selected for item analysis. The middle 20 test takers were excluded from the analysis. The formulae for difficulty level and discriminating index are as follows.

Formula of Item difficulty index

$$P(\text{value}) = \frac{Ru + Rl}{Nu + Nl}$$

Formula of Item Discrimination Index

$$D = \frac{Ru - Rl}{Nu \text{ or } Nl}$$

Ru = the number of test takers in the upper group who responded correctly

Rl = the number oftest takers in the lower group who responded correctly

Nu= Number of test takers in the upper group

Nl= Number of test takers in the lower group

The following figures show the selected items based on item difficulty and item discrimination.

Figure 1: Level of Item Difficulty

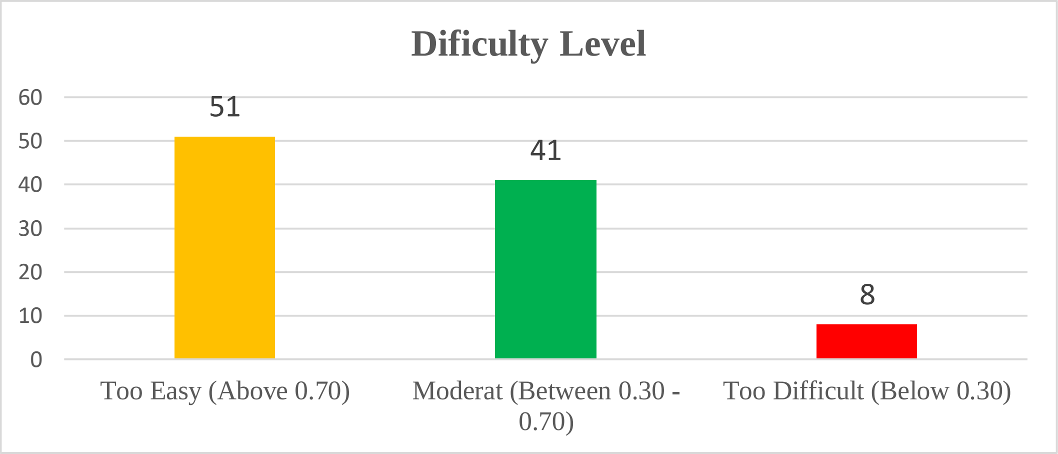

Figure 1 depicts the selected items in terms of difficulty level. Although too easy items are considerably more than too difficult items, moderate items are acceptable.

Figure 2: Power of Item Discrimination

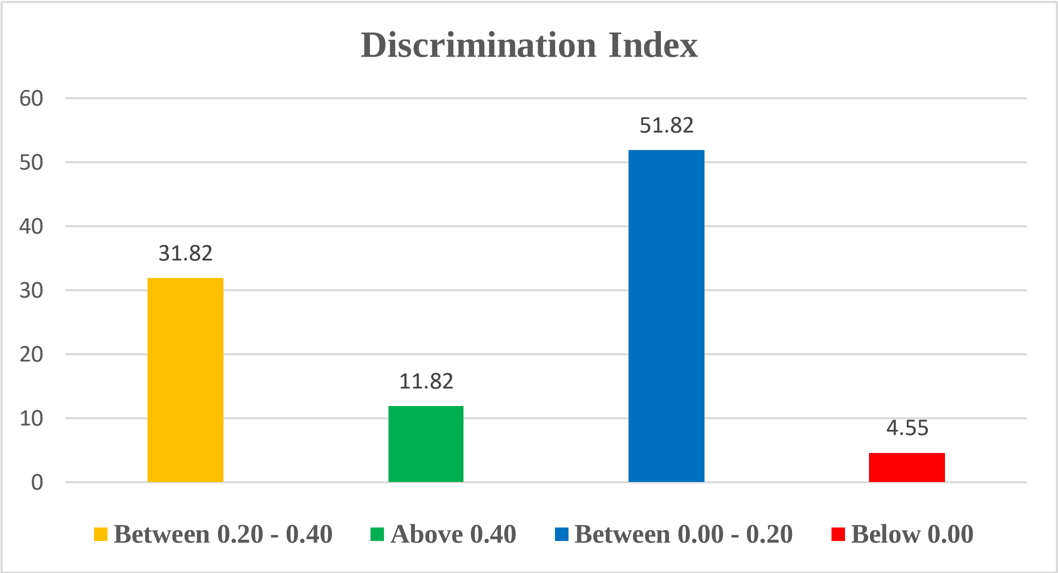

Figure 2 describes the selected items in terms of discrimination index. Items with lower discrimination index (between 0.00-0.20) are remarkably more than items in acceptable ranges (between 0.20-0.40 and above 0.40). There are 5 items (4.55%) with negative discrimination index and therefore these items need to be reviewed.

Findings and Discussion

In the diagram, the difficulty index, 9 items out of 110 (8%) are excluded due to difficulty level, 56 items out of 110 (51%) are too easy; so they are needed to be revised. Only 45 items (41%) are accepted without revision and moderate for achievement test. In the diagram, the discrimination index, 5 items (4.55%) are negative items, 57 items (51.82%) are between (0.00-0.20), 13 items (11.82%) are above 0.40 and 35 items are between (0.20-0.40). 62 items (56%) should be completely revised as they have very low discriminating power. 48 items (44%) are acceptable without revision because they have good to excellent discrimination indices.

Based on the two figures, it is seen that there are more too easy items and more zero-negative items in this achievement test. Though surface validity may be rated as being low, the cloze test type covers broad range of topics in a short span of time; in terms of practicality, one of the three criteria of testing, it is convenient to prepare and less time consuming. It can also effectively assess knowledge, comprehension and application levels in Bloom's taxonomy of cognitive domain. It covers a greater amount of contents than matching type tests does.There are 100 cloze question items but 10 multiple choice question types, and so these items need to be modified to be a good test.

Quality control is important for test expansion. It is, therefore, crucial for teachers to make item analysis or pursue support where they feel insufficient. Items should be modified if test takers consistently fail to select certain multiple-choice alternatives. Items with negative discrimination indices should be deleted or replaced. Classroom teachers ought to rewrite all items with zero discrimination indices. They also need to be considered to replace or rewrite all items with low positive discrimination indices.

Conclusion

It is found that it is generally too late to change the results after a test has been administered and scored. However, item analysis can improve tests by reviewing or excluding ineffective items. Item analysis can provide important diagnostic information on what learners have learned and what they have not learned. Classroom teachers may not try out the test with the same students who have taken it or they may try out the test with other classes of similar test takers or future test takers. Item analysis is a vital step in the test development cycle, as all tests are composed of items and good items are necessary for a good test. In short, item analysis is an essential part of analyzing the results of tests for improving their usefulness.

References

Bachman, L. F. (2004) ***Statistical Analyses for Language Assessment***. Cambridge: Cambridge University Press.

Best John W. & Kahn James V. (2010) ***Research in Education***. New Delhi: PHI Learning Pvt.Ltd.

Bonnel, A. M., & Boureau, F. (1985) ***Labor Pain Assessment*****:** ***Validity of a Behavioral Index*****.** Pain, 22, 81–90.

Boopathiraj. C. (2013) Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education, ***International Journal of Social Science & Interdisciplinary Research Vol.2 (2)***, February.

Brown, H. Douglas (2003) Language *Assessment*: ***Principles and Classroom Practices***. New York: Longman.

Ebel, R. L. & Frisbie, D. A, (1986) ***Essential of Education Measurement***. Englewood Cliffs, NJ: Prentice Hall.

Hoy Wayne. K (2010) ***Quantitative Research in Education: A Primer***. UK: Sage Publication.

Jack C. Richards & Richard Schmidt (2010) ***Dictionary of Language Teaching and Applied Linguistics***, 3rd ed., New York: Longman

Kennedy Quaigrain. (2017) *Using Reliability and Item Analysis to Evaluate a Teacher-developed Test in Educational Measurement and Evaluation,* Cogent Education, 4: 1301013

Lestari H. (2011) *An Item Analysis on Discriminating Power of English Summative Test,* (A Case Study of Second Year of "SMPN 87" Pondok Pinang), Jakarta.

Popham, W, J. (1978) *Criterion-referenced Measurement*, Englewood Cliffs, NJ: Prentice-Hall.

Vince, M. (2008) *Macmillan English Grammar in Context*, *Advanced*, Oxford: Macmillan Publishers limited.

Vince, M. (2008) *Macmillan English Grammar in Context*, *Intermediate*, Oxford: Macmillan Publishers limited.

Yule, G. (2006) *Oxford Practice Grammar, Advanced*, Oxford: Oxford University Press.

Internet Reference

https://www.blogger.com/profile/03032393678477163501

http://research-education-edu.blogspot.com/2011/11/test-items-analysis.html