

ITEM ANALYSIS AND ITS APPLICATION IN LANGUAGE ACHIEVEMENT TESTS AT THE TERTIARY LEVEL HOANG DUNG*

ABSTRACT: Given the widespread use of language assessments and their importance in the Vietnamese tertiary education, ensuring the quality of language assessments is a crucial activity to help them achieve their intended goals. Being a popular form of summative assessment in many language courses at the tertiary level, achievement tests also require careful attention. However, evaluating the quality of an assessment is a complicated task. This paper presents a guideline for a beginner-friendly test evaluation method called item analysis, which can also be used for small-scale achievement tests that involve high-stakes pass/fail decisions based on mastery/non-mastery classification. This item analysis procedure can be a practical choice for language instructors who need an accessible way to determine the quality of items in an important test but lack expertise to perform such an analysis. Although it has certain limitations, such as being incompatible with polytomous test items, the method would lend itself to a wide range of testing contexts at the tertiary level in Vietnam.

KEY WORDS: achievement test; language assessment; item analysis; item difficulty; item discrimination.

RECEIVED: July, 25nd, 2024.

EDITED: November, 25th, 2024

1. Introduction

1.1. *Language achievement tests and their uses in tertiary education*

In the field of language teaching at the tertiary level, as part of many language courses, the achievement test is a popular choice of summative assessment to evaluate students' progression as well as to inform language instructors of their teaching effectiveness (Bachman & Damböck, 2018). Although many achievement tests tend to play a rather limited role within a classroom context to effectively facilitate the process of teaching and learning, in multiple cases related to language courses at the tertiary level, results from achievement tests might be used to make more important decisions. To be specific, in such cases, there are often two layers of decisions involved. The first decision is to classify students into mastery and non-mastery groups by comparing their test scores to a predetermined cut-score. Students scoring above the cut-score are considered masters (i.e., having a good command of the course content), while students scoring below the cut-score are considered non-masters (i.e., having a poor grasp of the course content; Bachman, 2004; Brown & Hudson, 2002). The subsequent decision involves passing or failing students based on the first classification. Specifically, mastery students will be considered competent enough to pass the language course and/or advance to the next level, while non-mastery students may be required to receive additional language input at the same level (i.e., may have to retake the course) until they are considered masters (Bachman & Damböck, 2018). This second decision might also be influenced by factors other than the achievement test, such as the commonly-used midterm test or participation score. Despite being small-scale assessments in the sense that they are only used within a narrow classroom

*Ha Noi University; Email: dunghd@hanu.edu.vn

setting in comparison to a much wider setting targeted by large-scale tests such as IELTS and TOEFL, achievement tests involving classification and advancement decisions as illustrated above can be considered medium-stakes or even high-stakes assessments (Bachman & Damböck, 2018) under the consideration that failing such a test may hinder students' study progression and might subsequently extend the length of their study programme.

1.2. Evaluating the quality of a language achievement test

An achievement test is a criterion-referenced test (CRT) in the sense that the scores of the test takers are compared to a pre-determined cut-score in order to, as mentioned above, make the mastery/non-mastery classification and pass/fail decisions (Brown & Hudson, 2002). In this way, the cut-score is considered as a performance criterion that distinguishes masters and non-masters, and a CRT is therefore able to make absolute decisions (Brown & Hudson, 2002).

As for any language assessment in general, quality assurance is vital and should always be carried out (Bachman & Palmer, 2010). To be specific, there are several aspects of an assessment that need to be investigated before any conclusions about its quality can be reached. Some of the aspects may include the consistency of test scores (i.e., reliability) and validity of test interpretation (i.e., construct validity and content validity), and an effort aimed at investigating those aspects can be classified as a test validation study (Weir, 2005). Meanwhile, a more comprehensive endeavour at evaluating the quality of an assessment is called a test justification study, in which not only test score and test interpretation but also test use (e.g., decisions to be made from the test and their consequences) are investigated (Bachman & Palmer, 2010). It can be inferred that the more comprehensive a study is, the more resources (e.g., time, labour, expertise, budget, equipment) are required, and those resources may not always be available (Bachman & Palmer, 2010). The lack of resources for test evaluation can be said to be the case for achievement tests in general because of their small scale. Furthermore, within the field of language assessment, while a significant number of studies focus on such high-stakes, large-scale language assessments as IELTS and TOEFL (Chapelle & Voss, 2013), there is a limited number of studies that centre on the quality of CRTs such as achievement tests.

Given the limited attention that CRTs such as achievement tests receive despite the fact that they are widely used in the context of language teaching and bear significant impact on stakeholders such as language instructors and students (Bachman & Damböck, 2018), this paper attempts to bridge this gap in the existing literature of language assessment by focusing on small-scale CRTs that are used to make (relatively) important decisions. Furthermore, compared to large-scale language assessments which require considerable resources to be investigated, the evaluation of small-scale CRTs needs to be more efficient and practical given the more limited resources. Taking this into account, this paper proposes for language instructors the use of the procedure called item analysis (IA). Based on this procedure, language instructors can see how successfully an achievement test and the items within it serve the intended purposes of the assessment. Given the popularity of achievement tests in Vietnam and by targeting this particular type of assessment, this paper also aims to further improve the testing practice within this context.

2. Item analysis

2.1. Item analysis and its assumptions

The item analysis procedure to be introduced in this paper is informed by the work of Bachman (2004), Brown and Hudson (2002), and Malec (2018), which collectively provide a detailed explanation about the item analysis and the indices to be calculated in it. An item analysis (IA) can be conducted in Excel, a popular spreadsheet program, and the achievement test chosen to be studied

should be solely comprised of dichotomously-scored items. In other words, for any item within the test, there can only be either a right or wrong answer. A perfect example of this item type is a multiple-choice question or a dichotomously-scored open-ended question. An IA is mostly not compatible with polytomous items (Bachman, 2004) (e.g., essay writing), and the validation of this particular item type is not within the scope of this study.

In order to conduct an IA, the following data need to be prepared:

- The achievement test paper and keys.
- Each student's answer (i.e., right or wrong) to each test item.
- A cut-score for the achievement test. As mentioned above, a cut-score is used to classify mastery and non-mastery students and to make pass/fail decisions. It should be available for any achievement tests involving important decisions.

An IA is conducted under the following assumptions about the test to be analysed. Should any of the assumptions not be met, conclusions drawn from the IA might turn out to be biased and inaccurate.

- Assumption 1: The achievement test is administered fairly and consistently across different administrations; students' performance on the test is independent of each other.
- Assumption 2: The number of test takers to be included in the analysis should be at least 30.
- Assumption 3: The cut-score is able to effectively distinguish mastery and non-mastery students.

It can be challenging to check this assumption. A possible way for language instructors to test the validity of this threshold is their in-class observation of students' performance throughout the course. If the students that language instructors consider competent all pass the achievement test, this should be indicative of a well-chosen cut-score.

2.2. Item indices in an item analysis

An IA essentially involves the calculation and interpretation of two indices for every test item, namely the item difficulty index and item discrimination index. Those indices shed light on the quality of each item and, as a consequence, the test as a whole. Therefore, it is important to understand the meaning behind those indices and their interpretation.

Item difficulty index

The difficulty index for each item is calculated by dividing the number of students giving a correct answer over the total number of students taking the test. The value of the index can range between 0.00 and 1.00.

Example 1:

Given that the total number of students giving the correct answer is 70, and the total number of students taking the test is 100, the difficulty index for the item is $70/100 = 0.70$.

Essentially, this index indicates the proportion of students who answered the item correctly. A high figure indicates that a high number of students gave a correct answer, meaning the item is considered relatively easy for this particular group of students. Since the purpose of any achievement test is to test content that had been taught during the course rather than content from domains unrelated to the course, it is reasonable to expect that a high number of students who already studied the course should be able to answer the item correctly and a well-designed item should have a relatively high value. In fact, for an ideal achievement test, test takers are expected to score a high mark on average as the test score distribution should be negatively skewed (Malec, 2018). On the other hand, a very low value may be indicative of items testing wrong domains. According to Bachman (2004) and Brown and Hudson (2002), it is reasonable to expect the item difficulty index to

be as high as possible (e.g., up to 1.00) while the lower bound of the index should be within the relatively narrow range of the cut-score. Informed by these studies, it is advisable that the acceptable range for CRT item difficulty index should be within the following range:

- Lower bound: equal to cut-score expressed as a percentage minus 10%. For example, if the cut-score is 5/10 (50%), the lower bound for item difficulty index should be 0.40 (40%).
- Upper bound: equal to 1.00

Should the value of the difficulty index fall out of this optimal range, language instructors might need to revise those items. It should also be noted that evidence from the IA alone is not sufficient to conclude that those items are poorly-designed and therefore need to be eliminated. Possible reasons for such a low value shall be discussed in the “IA demonstration” section below.

Item discrimination index

Based on the calculation of the item difficulty index, it is possible to calculate *the item difficulty index for only mastery students (1)* and *the item difficulty index for only non-mastery students (2)*. These figures are needed to calculate the next index, *item discrimination index*. Examples showing the calculation of index (1) and (2) are shown below:

Example 2:

Given that the total number of mastery students giving the correct answer is 50, and the total number of mastery students is 60, *the item difficulty index for mastery students (1)* is $50/60 = 0.83$.

Given that the total number of non-mastery students giving the correct answer is 20, and the total number of non-mastery students is 40, *the item difficulty index for non-mastery students (2)* is $20/40 = 0.50$.

As can be seen, approximately 83% of mastery students gave a correct answer while it was only 50% for non-mastery group. Since mastery students performed better than non-mastery ones in terms of this item (83% compared to 50%), which should be a normal expectation for any well-functioning items, it can be inferred that the item is able to effectively distinguish masters from non-masters. On the contrary, if the percentage of mastery students giving a correct answer happens to be smaller than that of non-mastery students, this constitutes evidence for a problematic item. According to Brown and Hudson (2002) and Malec (2018), in an ideal scenario, a CRT item should have 100% of mastery students and 0% non-mastery students giving a correct answer to it.

In fact, the calculation of the *item discrimination index* is done by subtracting index (2) from index (1), and the obtained value gives information about the discrimination power of the item. The value of item discrimination index ranges between -1.00 and 1.00. In the second example, the item discrimination index would be a positive value (0.33). Meanwhile, a negative value suggests a faulty item on which non-mastery students performed better than mastery students. Therefore, a good item needs to achieve at least a positive discrimination value (Bachman, 2004). The optimal range of the discrimination index is thus between 0.00 and 1.00.

2.3. Item analysis procedure

With a good understanding of the item difficulty and item discrimination indices and their optimal range, language instructors can use the following steps for conducting an IA.

Step 1: The prepared item data (i.e., answer to each item by each student) is entered into an Excel worksheet. The data can be arranged following the layout suggested in Table 1. To be specific, the first row is used to record the item numbers, and the first column features the student numbers, after which an individual answer for each item by each student is entered in the corresponding cell. Item data should be transformed according to the following rules: A correct answer is entered as “1”; An

incorrect answer is entered as “0”; An unanswered item or an inappropriate answer is represented by a blank cell.

Step 2: The total test score for each student is calculated in a separate column, and the first student column is then reordered from the lowest to highest total test score. A blank row can be used to visually separate students from the mastery and non-mastery groups based on the cut-score.

Step 3: The difficulty index and discrimination index for each item are calculated.

Step 4: The indices from step 3 are compared with the optimal range. Items with index values within the optimal range perform well and therefore should be kept. Meanwhile, a variety of decisions can be made concerning the items with less-than-optimal values. Suggested actions regarding these items can be seen in the next section.

Table 1. *Item analysis table*

Student number \ Item number	item1	item2	item16	item54	Total score
2	1	0	1	0	25
20	1	1	1	1	44
40	1	1	0	1	50
The cut-score is 70%					
78	1	1	1	1	73
93	1	1	1	1	81
95	1	1	1	1	81
96	1	0	1	1	82
Item difficulty index	1	0.714286	0.857143	0.857143	
Item discrimination index	0	0.083333	0.333333	0.333333	

2.4. Item analysis demonstration

For the purpose of demonstration, the following data is extracted from the researcher's item analysis study of an achievement test used in a tertiary-level EFL course. The item difficulty and discrimination index for three selected items from the test have been calculated and presented in Table 2 below. This demonstration will use the aforementioned optimal range to draw conclusions regarding the quality of the three test items. The cut-score used for this test is 50%.

Table 2. *Comparison of index values*

Item	Item difficulty index	Item discrimination index	Optimal range for item difficulty index	Optimal range for item discrimination index
1	0.90	0.50	Between 0.40 and 1.00	Between 0.00 and 1.00
2	0.20	0.20		
3	0.60	-0.10		

Item 1:

Both indices fall within the optimal range. For an item with high item facility and item discrimination value and in the absence of evidence against its validity, language instructors can conclude with fair confidence that this item is of appropriate difficulty and can discriminate masters and non-masters. Given the limited resources available to language instructors, this is a fair

conclusion that one can possibly make. On the other hand, language instructors should not be fully confident in the conclusion due to an inherent weakness of IA. Specifically, indices such as item difficulty and discrimination may fluctuate across different groups of test takers (Bachman, 2004). For example, the group of test takers being studied might be highly competent compared to an average group, and therefore the indices might be skewed. To arrive at a more certain conclusion, language instructors will require more observations from other groups of students taking the same test, and the result from IA might need to be triangulated with results from other studies of the same test. For example, since IA only evaluates items in terms of their statistics, an additional study that investigates the content of the test items will provide a better understanding of the suitability of the test items.

Item 2:

Despite having an acceptable discrimination index, the item displays a very low difficulty value, which is a sign that the item was unnecessarily challenging for this particular group of test takers. Such a difficult item may also have demotivating effects on students (Malec, 2018). There could be different reasons for this low value: (1) the item is poorly designed (e.g., it tests the wrong domains); (2) the item is well-constructed but the language instructors did not teach the content targeted by this particular item to students during the course; (3) this particular group happens to be less competent than an average group, and the result could be skewed. Regarding reason (1), as Brown and Hudson (2002) put it, the purpose of an achievement test is to test knowledge that had been transferred to students, and the use of tricky questions or questions that test irrelevant domains should be minimized or avoided. If language instructors believe reason (1) is the root of the problem, they can further investigate the content of the test item and make necessary revisions. However, a different solution is required if reason (2) is the cause. Overall, it is evident that item 2 is problematic, and it is up to the language instructors to take the most appropriate action about it.

Item 3:

Even with an acceptable difficulty index of 0.60, a negative item discrimination value must mean that the test item is not useful and should be removed from the assessment since a higher proportion of non-masters than masters answered it correctly. Possible reasons for this low value might be that the key for the item is wrongly set or that the item is badly constructed (e.g., wrong domain is tested). Again, it is the task of language instructors to decide the most appropriate action regarding the test item based on their awareness of the situation.

3. Conclusion

In conclusion, item analysis is an accessible method especially for those inexperienced in evaluating language assessments as it does not require extensive expertise. It is also simple to carry out with minimal effort and resources since there is no need for a dedicated software, and it can be done even with a single test administration involving a smaller number of test takers. Furthermore, since IA is often the first step towards a more comprehensive test justification investigation, learning how to conduct it is reasonable if one wishes to further improve their testing practice, especially in a context where assessments such as achievement tests are widely used. However, IA is not without flaws. As can be seen, it is only compatible with certain test formats. In addition, an inherent weakness is that it is only able to draw conclusion from a certain group of students and that the result may fluctuate across groups of varying competence or occasions of administration. Therefore, if one wants to draw consistent conclusions from IA, multiple IAs are much needed. However, a single instance of IA will still benefit users decidedly with the information it can provide.

REFERENCES

1. Bachman, L. F. (2004), *Statistical analyses for language assessment. Cambridge language assessment series*. Cambridge University Press.
2. Bachman, L. F., & Palmer, A. S. (2010), *Language assessment in practice*. Oxford University Press.
3. Bachman, L., & Damböck, B. (2018), *Language assessment for classroom teachers*. Oxford University Press.
4. Brown, J. D., & Hudson, T. (2012), *Criterion-referenced language testing*. Cambridge University Press. <https://doi.org/10.1017/CBO9781139524803>
5. Chapelle, C. A., & Voss, E. (2013), Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1079–1097). Wiley. <https://doi.org/10.1002/9781118411360.wbcla110>
6. Malec, W. (2018), *Developing web-based language tests*. Wydawnictwo KUL.
7. Weir, C.J. (2005), *Language testing and validation: An evidence-based approach. Research and practice in applied linguistics*. Palgrave Macmillan.

**Phân tích câu hỏi trắc nghiệm và ứng dụng trong bài thi cuối học phần
môn học ngoại ngữ ở bậc đại học**

Tóm tắt: Với việc đề thi đánh giá năng lực ngoại ngữ được sử dụng phổ biến và đóng vai trò quan trọng trong giáo dục đại học tại Việt Nam, việc đảm bảo chất lượng đề thi là một hoạt động cần thiết nhằm đem lại kết quả mong muốn. Vì được sử dụng phổ biến như một hình thức kiểm tra hết môn đối với nhiều khóa học ngoại ngữ ở cấp đại học, bài thi cuối học phần vì thế cũng cần được chú trọng về mặt chất lượng. Do đánh giá đề thi là một công việc phức tạp, bài viết hướng dẫn cách áp dụng một phương pháp dễ tiếp cận có tên gọi là phân tích câu hỏi trắc nghiệm. Phương pháp này có thể áp dụng với dạng bài thi cuối học phần thường được sử dụng để đánh giá mức độ nắm vững nội dung môn học của người học để đưa ra các quyết định quan trọng (ví dụ như người học đủ điều kiện qua môn học hay không qua môn học). Đây là một lựa chọn thiết thực đối với giáo viên bộ môn ngoại ngữ có nhu cầu đánh giá chất lượng các câu hỏi trong đề thi nhưng chưa có nhiều kinh nghiệm để tiến hành công việc như vậy. Mặc dù có một số hạn chế nhất định như không áp dụng được với các định dạng câu hỏi ứng đáp đa phân, phương pháp này có tính ứng dụng cao trong bối cảnh giáo dục đại học ở Việt Nam.

Từ khóa: bài thi cuối học phần; đánh giá năng lực ngoại ngữ; phân tích câu hỏi trắc nghiệm; độ khó của câu hỏi; độ phân biệt của câu hỏi.