Research and Review Insights

Research Article ISSN: 2515-2637

Appraising lecture room achievement assessment: A documentary perspective using item analysis technique

Aloysius Rukundo*

Department of Educational Foundations and Psychology, Mbarara University of Science and Technology, Mbarara, Uganda

Abstract

Introduction: The aim of this paper was to explore item difficulty and discriminating power for sample objective and essay items of an undergraduate achievement test. In this paper, results of item analysis of sample multiple choice and essay items are presented.

Methodology: This paper involves a review and analysis of test scripts, using the item difficulty index (P), the discriminating index (D), and distractor analysis. One item (5%) from the 20 multiple-choice items was selected for analysis. All three essay items (100%) were involved in the analysis. Respective formulae for P and D were used during the analyses.

Results: The difficulty and discrimination index respectively for the multiple choice item were 84% and .24. The only functional distractor was option "A". The first essay item had P of 78% and D of .28. The second essay item had P=.63 and D=.37. For the third essay item, P was 81% while D was .16.

Conclusion: The items had from low to moderate difficulty indexes and from poor to good discriminating power. These results indicate fair test items that were accommodative of all candidates of different abilities.

Introduction

Assessment at all levels of learning is a cardinal process that signals the stakeholders about achievement of the teaching–learning objectives [1]. Further, assessment is a valuable tool for identifying the extent to which learning occurred among learners of different abilities. Moreover, assessment is necessary for provision of feedback regarding the teaching–learning process [2]. For any assessment tool, quality is a central aspect, as the quality of an instrument usually determines the accuracy and credibility of its results. When the credibility of an assessment device suffers, the decisions based on its results suffer as well. As a result, concerns regarding the credibility of tests and examinations among universities will often arise.

Assessment is mainly realised in classroom tests and examinations [1,3]. In Uganda, testing among some tertiary institutions, secondary schools and lower levels of learning is governed by examinations bodies. Apart from helping in coordinating the examinations processes, the examinations bodies serve as van guards for quality and equity of the examination process [4].

However, {terminal} assessment in institutions of higher learning, specifically at academic institutions of higher learning, remains a relatively liberal process. Different academic institutions have different examination systems [3]. Therefore, they use different test and item types in evaluating their learners. Common item types in institutions of higher learning include multiple choice questions (MCQs) and essays. Consequently and to a larger extent, individual examiners have the autonomy to set, score and mark their examinations. To an extent and following statutory guidelines, universities have put some quality regulatory mechanisms such as internal and external test moderation, external examination and quality assurance committees. While different mechanisms are often put in place to safeguard the quality of examinations at universities, such mechanisms may not be robust and in-depth at quality checks and balances. Partly, time and other logistical constraints may not allow comprehensive external processes intended to keep tests and examinations free from ambiguities and other quality constraints.

Often, item constructors make some "gut level" decisions regarding the quality and credibility of the items, especially when the items have to be kept for reuse [5]. When such decisions are made, examiners who reuse test items risk using both "good" and "bad" items. Fortunately, certain more quantitative mechanisms exist that could be used to estimate the quality of stored test items. Such estimates exist for individual examiners to self-check regarding the quality of their tests. One such a mechanism is item analysis. Item analysis is appraising of test item to ascertain its effectiveness in assessing the learning outcome for which it was constructed [6]. Item analysis considers many aspects of a test, although two central aspects–item difficulty (P) and item discriminating power (D) tend to override the debate about test effectiveness.

Therefore, item analysis is usually purposed at evaluating the functioning of specific test items [7,8]. To that end, items suspected to be flawed tend to have negative discriminating power or considerably appear easier or tougher than items from the same subject domain specification [7]. With such an importance at the back of the mind, item

Correspondence to: Aloysius Rukundo, Department of Educational Foundations and Psychology, Mbarara University of Science and Technology, Mbarara, Uganda, E-mail: arukundo@must.ac.ug, ORCID: https://orcid.org/0000-0002-6518-4360

Keywords: *item analysis, discriminating and difficulty power, achievement tests, assessment*

Received: August 11, 2025; **Accepted:** September 08, 2025; **Published:** September 15, 2025

analysis data serves a number of functions, especially creation of an item pool for improvement of future test items, and enhancing validity of future tests, through mechanisms that eliminate test items that show poor functioning during the analysis [9].

During item analysis, students' responses are counted. Counting then provides a means of testing items and then "compiling statistical data on the number of examinees who answer the item correctly". In item analysis, a number of considerations are made. This paper, however, will concentrate on two common aspects of item difficulty and item discriminating power. Item discriminating power is the ability of an item to differentiate between high and low achievers in a test [7]. Item difficulty indicates how easy or difficult an item was for examinees [1]. The higher the difficulty index, the easier the item [6]. It is necessary to appreciate that items with moderate difficulty index are more likely to have good discriminating power. Common causes of poor discrimination are usually those related to technical or test writing problems, poorly taught or untaught subject material, ambiguity in item phraseology, grey areas of opinion and controversy, comprehension, wrong or mistaken keys [1,8].

Item analysis for essays and multiple choice items usually considers calculation of difficulty and discriminating indices. However, in addition to analysing difficulty and discrimination indices for multiple choice items, examiners usually appraise the distractors. Appraising of distractors happens in terms of evaluating their functionality [5]. A functional distractor is one that is chosen by 5% or more of the candidates [1]. That selected by less than 5% is regarded as a non–functioning distractor. Usually, non–functional distractors make the test easier, reducing the discriminating power of the item. The reverse is true [1].

There can be two main causes citable for the presence of non–functional distractors. The obvious cause relates to the training and item construction ability of the examiner. The second is associated with the mismatch between the subject content that was targeted for the test. Other less important factors are first and foremost related to the cognitive domain or level at which an item is constructed [1]. Accordingly, distractors of items at lower cognitive levels could have higher chances of non–functionality. Secondly, items from irrelevant subject areas and low number of distractors could be responsible for non–functionality. The third aspect is connected to the item construction ability of an examiner, the presence of logical cues in the item options in relation to the item stem. To an extent, masterly of the subject matter among candidates may have its foot in the poor functioning of the item distractors. That means candidates can easily identify the distractors.

In certain circumstances, a distractor may be more frequently selected than the correct alternative. That could be due to possible poor construction of the stem, which would eventually mislead the candidates. In such scenarios, the item could be corrected by moderating the test and/or double–checking the item. That phenomenon is commonly manifested when more candidates in the high achieving group fail the item than expected.

Item analysis has been applied in appraising achievement tests among different subjects. In analysing a 40-item test for 120 medical students, the mean item difficulty index was 50.16 ± 16.15 while mean discriminating power was 0.34 ± 0.17 [8]. In that analysis, most items were found to have moderate difficulty and discriminating power. In appraisal of another test comprising 90 multiple choice items, the majority of the items, 74 (82%) were found to have a good or an acceptable difficulty level (Mean= 55.32 ± 7.4) [4]. In the same study, 7(8%) of the items were very difficult and 9 (10%) were too easy for students. Further, 72 (80%) of the items in the same analysis were in the "excellent" to "acceptable" discriminating index, while 18 or (20%) of the items had a poor discriminating index (Mean= 0.31 ± 0.12). The analyses reported above, however, were based on objective type items only. Moreover, no case–by–case easy demonstration of the analyses was done for beginners in assessment.

As earlier noted, the knowledge of item analysis could benefit a multitude of university faculty that have no basics of teaching and assessment. Oftentimes, universities recruit into teaching graduates with excellent grades (the cream) but without the basic skills in assessment. Yet, limited arrangements are made in tooling such faculty with assessment skills. Studies that tried to demonstrate item analysis concentrated on one type of assessment, usually of objective format. Moreover, they mainly concentrated on high school tests, using complex weighting methods. Such techniques could be comprehensible to assessment experts but incomprehensible to ordinary examiners, who may be the majority in the Ugandan higher education system. There was a need to document techniques that demonstrate analysis of both multiple choice and essay types of assessment. This paper aims at testing difficulty and discriminating power of a sample of multiple choice and essay items of a university achievement test. In addition, an appraisal of distractors for the multiple choice item will be made.

Materials and methods

Technique: The study involves documentary review of test papers, using item analysis (IA) as the major analytical strategy. Item analysis denotes a method of statistical computation involving students' responses to a particular test item [1]. IA is one of the techniques used in ascertaining the quality of classroom test-items in a given subject assessment. The technique aims at analysing the relationship between students' responses to an item. It serves as a basis for providing constructive feedback regarding appropriateness and effectiveness of the item [1].

Materials: The paper considered 122 marked and graded examination papers of a university course called measurement and evaluation in education and psychology. One item (5%) from the 20 multiple choice items was selected for analysis. Further, all the three essay items were considered for analysis. The decision for the selection of the items selected from the different sections was based on the fact that both measured application of knowledge.

Procedure: A test comprising 20 multiple choice items, one short essay item and two long essay items was set and moderated in May 2024. After moderation, the test was processed and administered to 122 finalist undergraduate students of Bachelor of Science with education students, in May 2024. Test scripts were marked and scored in June 2024, and marks recorded in appropriate excel mark sheet. Marks were serially arranged, in a descending order. Using the mark sheet, the test scripts were also arranged in the same order, with the highest score to the top and lowest score to the bottom. The assumption was higher marks indicated high achievers and lower marks low achievers. In analysing the objective item, twenty five scripts (20.5%) were picked from the upper group (high achievers) and 25 scripts (20.5%) picked from the lower group.

Data Analysis: The aim was to compute item difficulty and item discriminating power of a sample of items from the test. The computations were done manually and using Excel spread sheets. One multiple choice item was considered for analysis. Item number 12 was selected as an example from the multiple choice section for

Table 1. Data for difficulty and discriminating power for multiple choice item

Group	Alternatives				Omitted	Indices	
	A	*B	C	D		P	D
Upper (25)	1	24	0	0	0	0.84	0.24
Lower (25)	5	18	1	1	0		
Total	6	42	1	1	0		
*Correct alternative; D=Discriminating index; P=Difficulty index							

Table 2. Data for difficulty and discriminating power for essay item 1

Item score	High group frequency	*f*U	Low group frequency	*f*L
15	07	105	00	00
14	14	196	00	00
13	06	78	03	39
12	03	36	00	00
11	00	00	05	55
10	00	00	09	90
9	00	00	07	63
8	00	00	04	32
7	00	00	00	00
6	00	00	01	06
5	00	00	00	00
4	00	00	01	04
3	00	00	00	00
2	00	00	00	00
1	00	00	00	00
0	00	00	00	00
Summary	**N=30**	**∑*f*U=415**	**N=30**	**∑*f*L=289**
Indices	**P=0.78=78%**		**D=0.28**	

analysis. This particular item was selected for demonstration because it combined testing of higher-order and low order cognitive abilities. One item was considered in order to allow detailed and extensive analysis of the item. In addition to analysing the difficulty and discriminating powers of the item, consideration was made, of distractor evaluation as well. Distractors that were selected by less than 5% of the candidates considered for analysis were regarded as non–functional [1].

All the three essay items were included because each of the items combined and tested higher order thinking abilities, application, analysis and evaluation of knowledge. This paper applied analysis estimate from [5], as it is relatively simpler and comprehensible by non–expert assessment users. Moreover, the analytical strategy uses simple, easily accessible and applicable tools.

For the multiple choice item, the item difficulty was obtained using a formula $P = \frac{H+L}{T}$ where P is the difficulty index, H was the number of scripts picked from the upper category (high achievers), L was the number of scripts picked from the lower category (low achievers), and T the total number of scripts used in the analysis (sum of upper and lower categories) [5]. The discriminating index (D) was obtained using this formula: $D = \frac{H-L}{T/2}$. The acronyms in this formula have the same meaning as in the computation for item difficulty above.

For the essay items, the difficulty index was computed using the formula: $P = \frac{\Sigma fu + \Sigma fL}{2N(Score\ Max)}$. In the formula above, ΣʃU is the sum of the frequencies in upper group multiplied by scores in that group. In the same formula, ΣʃL is the sum of the frequencies in lower group multiplied with the scores in that group, while N is the sum of frequencies either in the upper group or lower group. The discriminating power, D was computed using the following formula: $D = \frac{\Sigma fU - \Sigma fL}{N(Score\ Max}$. The terms in the formula for item discriminating power are similar as those in the item difficulty formula. The formulae gave statistics as recorded among the tables in the results section.

Results

The aim of this paper was to explore item difficulty and discriminating power for objective and essay items of an undergraduate achievement test. Using the formulae for item difficulty and discriminating power, the results in Tables below were achieved.

Table 1 indicates that 24 students in the upper group got the correct while 18 from the lower group got it right. Applying the numbers in the formula for difficulty and discrimination index respectively, the difficulty index was 84% and the discriminating power was .24. The difficulty index of 84% indicates that the item was easy for the candidates. The discriminating power of .24 was fair and acceptable. Further, Table 1 shows that only one student from the upper group and five from the lower group picked alternative or distractor A. None of the students in the upper group selected alternatives C and D. However, distractors C and D were each selected by one student from the low achievers. Accordingly, A was the only functional distractor, as it attracted the highest number of candidates (12%). There was no script omitted, since the item was answered by all students involved in the analysis.

In Table 2, it is observed that the difficulty index of the first essay item is 78% and the discrimination index is .28. Therefore, the item was easy but good at discriminating high from low achievers. The first essay item had five subsections, each testing a different cognitive ability. Possibly, the different abilities were accommodative of candidates from both the high achievers' and low achievers group. Moreover, a similar example had been given during the lectures.

The data in Table 3 shows that the difficulty index of the second essay item was 62%, while the discriminating power was .37. That shows the item had average difficulty and excellent discriminating power. The respective item had two sections, one that required understanding of concepts, and the other, application of knowledge. The section for application of knowledge could have pushed away most of the candidates from the appropriate responses, hence the excellent discrimination power of the item. The average difficulty index could be attributed to the section of the item that needed understanding of concepts.

Table 4 shows that the difficulty index, P of the third essay item was 81%. The discriminating power, P, was .16. The difficulty index of 81% manifests that the item was too easy for candidates from both

Table 3. Data for difficulty and discriminating power for essay item 2

Item score	High group frequency	*f*U	Low group frequency	*f*L
10	05	50	01	10
9	07	63	00	00
8	06	48	00	00
7	10	70	02	14
6	02	12	04	24
5	00	00	11	55
4	00	00	04	16
3	00	00	03	09
2	00	00	02	04
1	00	00	01	01
0	00	00	02	00
Summary	**N=30**	**∑*f*U=243**	**N=30**	**∑*f*L=129**
Indices	**P=0.63=63%**		**D=0.37**	

Table 4. Data for difficulty and discriminating power for essay item 3

Item score	High group frequency	*f*U	Low group frequency	*f*L
15	12	180	00	00
14	05	70	01	14
13	04	52	06	78
12	03	36	07	84
11	04	44	05	55
10	02	20	05	50
9	00	00	00	00
8	00	00	05	40
7	00	00	01	07
6	00	00	00	00
5	00	00	00	00
4	00	00	00	00
3	00	00	00	00
2	00	00	00	00
1	00	00	00	00
0	00	00	00	00
	N=30	**∑*f*U=402**	**N=30**	**∑*f*L=328**
Indices	**P=0.81=81%**		**D=0.164**	

the high achievers and low achievers groups. As a result, it had poor discriminating power. The item had two sections, "a" and "b". Section "a" that carried higher marks needed knowledge of the "current" obstacles facing measurement and evaluation in Uganda, thus testing lower cognitive ability. Possibly, this section favoured even candidates from lower achievers, which could have negatively affected the discriminating power of the item. The section "b" of the item that required more advanced cognitive ability carried lower marks. It is necessary that this particular item needs revision for future use. The revision could be making the first section more advanced in terms of the cognitive abilities tapped into, or revising the marks allotted to the first section in the negative direction.

Generally, the difficulty and discrimination indices recorded in the Tables above indicate fair or moderate test items. That implies that the test in general was fairer. There is a possibility that some of the items could be revised, if in future such items were to be reused, and the future assessment was to achieve maximum effectiveness.

Discussion

This paper aimed at exploring item difficulty and discriminating power for some multiple choice item and essay items of an undergraduate achievement test. The difficulty index of the multiple choice item was over 70%, indicating the item was very easy for the candidates [1]; suggesting the item could be revised. However, the discriminating power was between 0.20 to 29, indicating moderate discriminating power [1] for which the item could be kept for future use.

Generally, the discriminating power and difficulty index of the items considered for this paper varied. This shows a similar track as different studies that have found varying levels of item difficulty and discriminating power, implying varying strengths of the respective tests. Rao, et al. [8] for example found moderate indices of difficulty and discriminating power. To the contrary, Kumar, et al [4] established that most items in their analysis were of good difficulty while others were very difficult. In the same study, a small percentage of the items were of very low difficulty. Therefore, students found such items too easy to answer. The same study concluded that another minority of items had poor discriminating index, while others were in the "excellent "to "acceptable" discrimination. Nevertheless, the results of such analyses depended on the selection type of items.

In the analysis for this paper, the difficulty power of the MCQ was moderate. That was below the views of scholars on item appraisal. For instance, [1] asserted that for most of the selected-response tests, the difficulty power was in the middle or moderate range, of about 55% to 75%. The discriminating index of the multiple choices item was lower than that of any of the essays. This observation is comparable with the views in literature [2]. Previous studies show that essay items show better discrimination indices than multiple choice items, because they are less susceptible to guesswork [2,3]. Multiple choice items are influenced by a number of factors, such as candidates' ability to guest the key, arrangement of the alternatives that make the correct answer obvious, and presence of clues in the entire item. Such factors make multiple choice items easier to answer [3]. On the other hand, candidates invest a high level of thought and concentration when answering essay items, which significantly impacts on the ability of distinguishing students in high achieving category from lower achievers [3].

Two distractors were not functioning as intended. While the item was deemed well by the test moderators and the distractors matched the subject are from which the item was constructed, it is possible that the mid-level cognitive domain enabled candidate to easily spot the two distractors [1], especially the candidates in the high performing (upper) group, who did not select the distractors. Further, the test was the first paper of the semester, which could have enabled masterly of the subject matter by the candidates. It is advisable that a non–functional distractor is removed or modified, as it does not contribute to the functioning of the item and the overall test. In the case of the item analysed for this paper, it is logical to modify the distractors, probably to answers as or more attractive than the key [10-13].

Conclusion

The multiple choice item was easy and had low but acceptable discrimination. The first essay item was easy with low but acceptable difficult index. On the other hand, the second essay item had a moderately acceptable difficulty but good discriminating power. The third item was too easy with poor discrimination. These results indicate fair test items that were accommodative of all candidates of different abilities. Determining of item discriminating and difficulty index would be a key step in improving the quality of items in higher education.

Recommendations

The mixed indices of P and D indicate that it is not until a teacher appraises her/his test that they could ascertain the extent to which the test is effective. Therefore, examiners at higher education institutions of learning need to make item appraising a routine for improvement in learning and assessment. Routine appraisal of tests in higher education institutions could institute a safeguard for test quality and effectiveness. Further, there is a need for inclusion of test appraisal as part of the statutory requirements for examiners at higher education institutions, for the reason articulated above.

Funding

The study was not funded by any organization.

Conflicts of interest

The author declares no conflict of interest.

References

1. Downing, S. (2010) Test Development. In S. Downing, P. Peterson, E. Baker, & B. McGaw (Eds.), International Encyclopedia of Education. Elsevier: 159-165.

2. Eldakhakhny B, Elsamanoudy AZ, Elsamanoudy A (2023) Discrimination power of short essay questions Versus multiple choice questions as an assessment tool in clinical biochemistry. Cureus 15: e35427. [Crossref]

3. Khan G, Ishrat N, Khan AQ (2015) Using item analysis on essay types questions given in summative examination of medical college students: Facility value, discrimination index. Int J Res Med Sci 3: 178-182.

4. Ingale AS, Giri PA, Doibale MK (2017) Study on item and test analysis of multiple choice questions amongst undergraduate medical students. Int J Community Med Public Health 4: 1562-1565.

5. Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V (2021) Item analysis of multiple choice questions: A quality assurance test for an assessment tool. Med J Armed Forces India 77: S85-S89. [Crossref]

6. Kunjappagounder P, Doddaiah SK, Basavanna PN, Bhat D (2021) Relationship between difficulty and discrimination indices of essay questions in for mative assessment. J Anat Soc India 70: 239-243.

7. Lahza H, Smith TG, Khosravi H (2023) Beyond item analysis: Connecting student behaviour and performance using e-assessment logs. Br J Educ Technol 54: 335-354.

8. Al Sulaim LS, Salati SA (2024) Item analysis of multiple-choice questions in an undergraduate surgery course-an assessment of an assessment tool. Sanamed 19: 163-171.

9. Mehrens WA, Lehmann IJ (1991) Measurement and evaluation in education and psychology.

10. Office of Educational Assessment (2024) Understanding item analyses. Seatle: University of Washington.

11. Rao C, Kishan Prasad HL, Sajitha K, Permi H, Shetty J (2016) Item analysis of multiple choice questions: Assessing an assessment tool in medical students. Int J Educ Psychol Res 2: 201-204.

12. Rezigalla AA (2022) Item analysis: Concept and application. Medical education for the 21st century 7: 1-6.

13. Tobin MA (2024) Guide to item analysis. Schreyer Institute for Teaching Excellence.