DOI: 10.36315/2019v2end001
Education and New Developments 2019

RELATIONSHIP BETWEEN ITEM DIFFICULTY LEVEL AND ITEM DISCRIMINATION IN BIOLOGY FINAL EXAMINATIONS

Marthese Azzopardi[1], & Carmel Azzopardi[2]
[1]Department of Biology, University of Malta Junior College (Malta)
[2]Department of Physics, University of Malta Junior College (Malta)

Abstract

Item analysis is a useful tool for a number of reasons, including the assessment of the quality of the test items. It indicates how difficult each item is and its ability to discriminate between the better and poorer students. The aim of the current study was to examine the quality of Biology Advanced level Paper 1 final examination and to see if there was any relationship between the item difficulty index and the item discrimination index values in these examinations. The data involved scores obtained by post-secondary students attending a public institution between 2014 and 2018. Final examination scores of a total of 1311 post-secondary students aged 16-17 years were analysed. Two different discrimination values were calculated, discrimination index and discrimination coefficient, to find which is the more appropriate to discriminate between high and low achievers. Both were appropriate, however the coefficient gave more positive results. No negative discrimination values, indicative of a 'defective' item, were recorded when using the two different formulae for discrimination. The correlation between the indices was investigated. Neither the discrimination index nor the discrimination coefficient was correlated with the difficulty index. Only the discrimination index was found to be significantly correlated with discrimination coefficient (0.563; P=0.000). The overall difficulty level was 'moderate' ($0.3 < P < 0.8$) in all years investigated and optimal (P=0.50) in 2018. In the five years investigated, 7% of the items (4/56) were 'too hard' and the rest, 93% (52/56) were of 'moderate' difficulty. Recommendations that result from this study are that tutors should design questions to include 'easy' ones, place them in order of increasing difficulty and to use item analysis to shed light on the discrimination power of the set questions. Results from this study show that a bank can be developed from which questions with the appropriate level of difficulty and discrimination may be chosen to increase the effectiveness and quality of future examinations.

Keywords: *Difficulty index, discrimination index, discrimination coefficient, Biology, post-secondary.*

1. Introduction

Examinations serve for a number of reasons, such as to ensure that students have learnt the core of a course, as well as to give feedback to students and teachers on how effective the learning and teaching were. For Maltese post-secondary students, the final examination plays an important role in the students' future as it determines whether they may proceed to their final year of studies and pursue a degree, or not. Item analysis is a valuable procedure performed after the examination that provides information regarding the reliability and validity of a test item. A plethora of research exists on item analysis of multiple choice questions (MCQs) because according to Halikar et al. (2016) they are most commonly used to assess the knowledge capabilities of undergraduate, graduate, and postgraduate students. MCQs are frequently used because they are very fast to grade, prevent the student from writing unnecessary information, are objective, eliminate assessor's bias and allow extensive coverage of the subject in a short period.

The Royal College of Physicians and Surgeon of Canada (2007) describe short answer questions (SAQs) as 'questions that can be answered in a few short words or phrases'. The same source continues to explain that such questions usually contain words such as "list" or name", suggesting that the answer consists of a series of short responses. Sam et al. (2016) argue that SAQs may provide greater validity than multiple choice questions if the aim of the assessment is to examine the student's ability to synthesise or generate rather than to recognise a correct answer. Despite the potential advantages of SAQs, their use in large-scale assessments has been limited since they cannot be marked by machines (Scalise et al., 2006). Item analysis is important to determine the quality of items in examinations and can be applied to SAQs. Item analysis of SAQs conducted by Tariq (2017) on Medical Pharmacology internal assessment exams is one of the few papers encountered on such item type.

2. Objectives

1. To find out the item difficulty level, discrimination index and discrimination coefficient of individual test items in Biology Advanced level final examination of post-secondary students.
2. To find out the relationship between: item difficulty and discrimination index; item difficulty and discrimination coefficient; discrimination index and discrimination coefficient.

3. Materials and methods

3.1. The examination and data collection

The Advanced Biology Paper 1 examination consists of 10-14 compulsory short items, carrying a total number of 100 marks and must be completed in 3 hours. A group of six tutors contribute items to set up the paper. In this study, 56 short questions taken over the period 2014-2018 were analysed. Students sit for the examination in June, at the end of the first year of teaching. A total of 1311 students sat the examination over the five-years investigated, with an average of 262 per year. The scores of the students were supplied by the Biology Department and to respect anonymity, they were handed over against an index number. Thus it was not possible to carry out analysis on gender.

3.2. Item analysis

The scores were then used to determine the difficulty index and power of discrimination using Microsoft Office Excel. Steps for item analysis were:

1. Ranking students in descending order of merit based on their test scores.
2. The top 25% were taken as high achievers (H) and the bottom 25% (L) as low achievers.
3. The calculations for the difficulty index for subjective questions, followed the formula by Nitko (2004):

$$P_i = \frac{A_i}{N_i}$$

where: P_i= Difficulty index of item i, A_i =Average score to item i, N_i = Maximum score of item i

The average difficulty index P for the entire script, can be calculated by the formula below:

$$P = \frac{1}{100}\sum_{i=1}^{N} P_i N_i$$

4. The discrimination index used in this study was calculated as:

$$\text{Discrimination index} = \left[\frac{\sum H - \sum L}{N\left(Score_{\max} - Score_{\min}\right)}\right]$$

H = total score for 25% of students in the high achievement group.
L = total score for 25% of students in the low achievement group.
N = 25% of total numbers of student tested.
$Score_{max}$ = maximum (full) marks for the question.
$Score_{min}$ = minimum marks for the question.

5. The advantage of using the discrimination coefficient instead of discrimination index was emphasized by Matlock-Hetzel (1997). Discrimination coefficients include every single person taking the test but only the upper (25%) and lower scorer (25%) are included in the discrimination index calculation process. There are over twenty discrimination indices used as indicators of the item's discrimination effectiveness. The Pearson Product Moment Correlation (r) between a specific item score and the total score of the same student was computed. Values range from -1.00 to 1.00. Higher positive values for the item-total correlation indicate that the item is discriminating well between high- and low-achievers. Negative values mean the opposite: low-achievers are more likely to get the item correct. If values are near zero, means that the item is not discriminating between high- and low- achievers. All students have similar probabilities of answering the item correctly, regardless of their total assessment score.

3.3. Interpretation

Table 1. Classification of the Difficulty Index and Discrimination Power values and suggested recommendations.

Difficulty Index	**Classification of Difficulty Level**	**Modification Results**
P < 0.3	Too hard	Modify
0.3 < P < 0.8	Moderate	Accept
P ≥ 0.8	Too easy	Modify
Discrimination Power	**Description**	**Recommendations**
D = negative	Defective Item	Rejected or improved
D between 0-0.19	Poor discrimination	Poor items to be rejected
D between 0.2-0.29	Acceptable discrimination	Marginal items usually need and subject to improvement
D between 0.3-0.39	Good discrimination	Reasonably good but subject to improvement
D = 0.4	Very good discrimination	Very good items; accept
D > 0.4	Excellent discrimination	Very good items; accept

3.4. Statistical analysis

The discrimination index and discrimination coefficient values, were determined using Pearson correlation analysis by SPSS version 24. P value of <0.05 was considered to indicate statistical significance.

4. Results and discussion

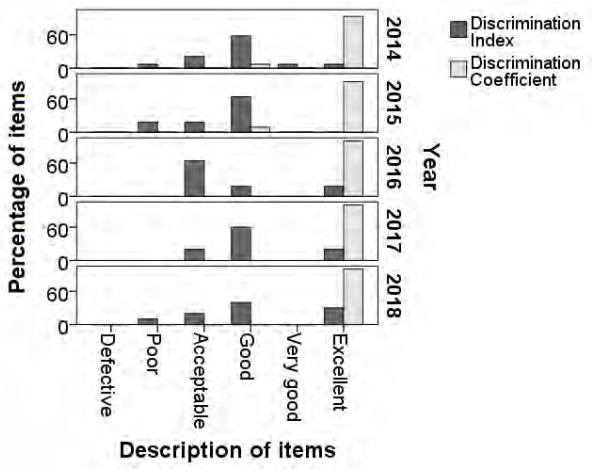
The difficulty index, discrimination coefficient and discrimination index were worked out for each item over the five-year period (total number of items = 56) and the results are shown in Table 2 and Figure 1. The same table shows that the same items were classified differently by the two different discrimination formulae. The majority (98%) of the items over the study period had a discrimination coefficient value > 0.4 that is considered as 'excellent', but the majority (93%) of the items had a discrimination index value of 0.20-0.39, classifying them as 'acceptable' to 'good'. This indicates that irrespective of the type of formula used to calculate the discrimination power, both gave positive results. However, the discrimination coefficient gave more desirable results compared with the discrimination index since the majority of items had 'excellent' discrimination.

Table 2. Discrimination Coefficient (DC), Discrimination Index (DI) and Difficulty Index (P) values for each item in Paper 1 classified by year. Values in red are less than 0.3.

Question	**2014**			**2015**			**2016**			**2017**			**2018**		
	DC	DI	P	DC	DI	P	DC	DI	P	DC	DI	P	DC	DI	P
1	0.57	0.28	0.27	0.62	0.35	0.70	0.63	0.27	0.61	0.70	0.46	0.63	0.60	0.29	0.56
2	0.57	0.35	0.45	0.56	0.31	0.47	0.77	0.44	0.41	0.60	0.35	0.53	0.68	0.43	0.47
3	0.72	0.45	0.47	0.62	0.26	0.76	0.64	0.26	0.29	0.67	0.34	0.33	0.73	0.49	0.53
4	0.69	0.37	0.38	0.67	0.32	0.57	0.79	0.20	0.54	0.57	0.27	0.45	0.72	0.45	0.45
5	0.58	0.40	0.50	0.44	0.18	0.64	0.67	0.26	0.36	0.60	0.31	0.31	0.68	0.37	0.67
6	0.64	0.36	0.39	0.63	0.32	0.61	0.51	0.20	0.73	0.66	0.31	0.37	0.64	0.33	0.40
7	0.47	0.30	0.54	0.63	0.31	0.27	0.45	0.21	0.64	0.73	0.43	0.64	0.62	0.37	0.31
8	0.50	0.15	0.42	0.36	0.16	0.48	0.70	0.41	0.50	0.61	0.29	0.44	0.57	0.19	0.65
9	0.65	0.38	0.47	0.56	0.29	0.62	0.61	0.32	0.34	0.65	0.32	0.36	0.54	0.29	0.31
10	0.58	0.24	0.34	0.55	0.33	0.36	0.58	0.39	0.43	0.55	0.31	0.74	0.53	0.31	0.56
11	0.51	0.39	0.52	0.62	0.36	0.46	0.54	0.29	0.63						
12	0.60	0.35	0.64												
13	0.33	0.32	0.27												
14	0.62	0.29	0.53												

Results from Figure 1 are encouraging since none of the items were classified as defective. This indicates that over the entire study period, tutors were able to write questions with ''good'' to ''excellent'' discrimination as calculated by the discrimination coefficient. Based on the discrimination index, only 4/56 (7%) items had a 'poor' discrimination (discrimination index = 0-0.19). Ovwigho (2014) also concluded that the discriminating power of test items could be measured by the discrimination index and discrimination coefficient. Results of the present study show that, over the five-year period investigated, the majority of the Biology Paper 1 items were able to discriminate and were valid.

Figure 1. A bar chart showing the percentage of items described according to the Discrimination Index and Discrimination Coefficient, classified by year.



An ideal examination should have a mixture of difficulty levels, however, results presented in Table 3 show that this was not the case in the examination papers investigated. 'Too easy' ($P \geq 0.8$) questions were never recorded in the five years investigated and in two consecutive years, the questions were 100% of 'moderate' difficulty. The overall difficulty level was 'moderate' ($0.3 < P < 0.8$) in all years investigated and optimal (P=0.50) in 2018. Thus paper setters consistently design examination questions of an overall 'moderate' level, irrespective of different persons involved each year and no written guidelines are given.

Table 3. Percentage of items per year having a Discrimination Coefficient and Discrimination Index of > 0.3 (good to excellent discrimination) and a Difficulty Index classifying them as 'too hard' ($P < 0.3$) and 'moderate' ($0.3 < P < 0.8$).

	Year					Discrimination power or Difficulty level & Description
	2014	**2015**	**2016**	**2017**	**2018**	
Discrimination Coefficient	100	100	100	100	100	0.3 and above [good to excellent]
Discrimination Index	71.4	63.4	36.4	80	60	
Difficulty Index	14.3	9.1	10	0	0	$P < 0.3$ [too hard]
	85.7	90.9	90.0	100	100	$0.3 < P < 0.8$ [moderate]

5. Analysis of correlation

The relationship between the difficulty index, the discrimination coefficient and index over the whole study period (2014-2018), was determined by Pearson correlation analysis and is given in Table 4. The strength of association as shown by Pearson correlation coefficient is as follows: small (r = 0.1 to 0.3), medium (r = 0.3 to 0.5) and large (r = 0.5 to 1.0). According to Suruchi et al. (2014), the difficulty indices and discrimination indices are most often reciprocally related. This was the case in this study. A small negative correlation (r = -0.082), which was not significant, was obtained. A linear relationship was observed from scatterplots between the difficulty index and discrimination coefficient as well as between the difficulty index and discrimination index, respectively. A negative correlation indicates that as the difficulty index values increase, the discrimination index decreases. This means that as the test items get easier, the discrimination index decreases, thus it fails to differentiate between high and low achievers. This finding is similar to that reported in the literature. For example, Mitra et al. (2009) obtained a negative correlation (r = -0.325) when working on multiple choice questions taken by pre-clinical students. Ahmed & Moalwi (2007) also reported (r = -0.453) for multiple choice questions taken by medical students in anatomy. Table 4 shows that a significant large positive correlation (0.563) value was obtained between the discrimination index and coefficient when all Paper 1 items were considered over the five-year study period. A linear relationship was obtained on plotting a scatterplot distribution for the discrimination coefficient and discrimination index. Table 5 shows Pearson correlation coefficient (r) between the indices calculated for each year. A significant large positive value of r was obtained for only three years 2015, 2017 and 2018 between the discrimination index and discrimination coefficient.

Table 4. Correlation between the various factors, Pearson correlation coefficient (r) and p value. (Significant at the 0.05 level).*

Variables correlated	Correlation coefficient (r) and *P*-value
Difficulty index and discrimination coefficient	-0.031 (p = 0.823)
Difficulty index and discrimination index	-0.082 (p = 0.549)
Discrimination index and discrimination coefficient	0.563 (p = 0.000)*

*Table 5. Correlation (r) between Difficulty Index and Discrimination Index, Difficulty Index and Discrimination Coefficient, and Discrimination Index and Discrimination Coefficient per year. The p-value is also given. *Correlation is significant at the 0.05 level.*

	Difficulty Index and Discrimination Index		Difficulty Index and Discrimination Coefficient		Discrimination Index and Discrimination Coefficient	
Year	r	*p*-value	r	*p*-value	r	*p*-value
2014	0.295	0.305	0.246	0.397	0.439	0.117
2015	-0.175	0.607	0.082	0.811	0.846*	0.001
2016	-0.430	0.187	-0.485	0.131	0.374	0.258
2017	0.493	0.147	0.012	0.973	0.777*	0.008
2018	-0.197	0.585	0.012	0.973	0.848*	0.002

6. Conclusion and recommendations

Results of this investigation emphasises a significant role of item analysis to educators and paper setters in determining the quality of test items. Examination of the item parameters of difficulty and discrimination will help a paper setter in detecting the defective and good individual items. Recommendations that result from this study are that tutors should design questions to include 'easy' ones, place them in order of increasing difficulty and to use item analysis to shed light on the discrimination power of the set questions. Although the discrimination index and discrimination coefficient can measure the discriminating power of test items, the discrimination coefficient gave better discrimination power than the index.

References

Ahmed, I., & Moalwi, A. A. (2017). Correlation between Difficulty and Discrimination Indices of MCQs Type A in Formative Exam in Anatomy. *Journal of Research & Method in Education* (IOSR-JRME). 7. 28-43. 10.9790/7388-0705042843).

Halikar, S.S., Godbole V., & Chaudhari S. (2016). Item Analysis to Assess Quality of MCQs. *Indian J Appl Res.*; 28;6(3);123-125.

Matlock-Hetzel, S. (1997). Basic Concepts in Item and Test Analysis. Paper Presented at the Annual Meeting of the Southwest Educational Research Association, Austin.

Nitko, A.J. (2004). Educational Assessment of Students. 4th Ed. Upper Saddle River, N.J.: Pearson/Merill Prentice Hall.

Mitra, N.K., Nagaraja, H.S., Ponnudurai, G. & Judson, J. P. (2009). The levels of difficulty and discrimination indices in type A multiple choice questions of Pre-clinical Semester 1 multidisciplinary summative tests. *IeJSME*, 3, 1, pp. 2-7.

Ovwigho, B.O. (2014). Empirical Demonstration of Techniques for Computing the Discrimination Power of a Dichotomous Item Response Test. *IOSR Journal of Research & Method in Education* (IOSR-JRME) e-ISSN: 2320–7388, p-ISSN: 2320–737X Volume 3, Issue 2 (Sep. –Oct. 2013), PP 12-17 www.iosrjournals.orgwww

Sam A.H., Hameed S., Harris, J., & Meeran K. (2016). Validity of very short answer versus single best answer questions for undergraduate assessment. *BMC Med Educ*; 16 (1):266.

Scalise, K., & Gifford, B. (2006). Computer-based assessment in e-learning: a framework for constructing 'intermediate constraint' questions and tasks for technology platforms. *J Technol Learn Assess*; 4 (6):1–44.

Suruchi & Rana, S. R. (2014). "Test Item Analysis and Relationship Between Difficulty Level and Discrimination Index of Test Items in an Achievement Test in Biology". *Paripex - Indian Journal of Research*, Vol. 3(6) pp.56-58.

Tariq, S., Tariq, S., Maqsood, S., Jawed, S., & Baig, M. (2017). Evaluation of cognitive levels and item writing flaws in medical pharmacology internal assessment examinations. *Pak J Med Sci*.;33(4): 866-870. doi: https://doi.org/10.12669/pjms.334.12887

The Royal College of Physicians and Surgeon of Canada. Editorial Revisions (2007). Short-answer questions: guidelines for their development. Educational Evaluation and Analysis Unit https://www.macpeds.com/documents/GuidelinesforDevelopmentSAQRoyalCollege.pdf [accessed 16/09/2018]