Analysis of the Discrimination Index of Final Biology Examinations in Malta

Marthese Azzopardi, Carmel Azzopardi
marthese.azzopardi@um.edu.mt

Abstract

Item analysis is a range of statistics that helps to determine the effectiveness of each item in an examination. It plays an important role in contributing to the fairness of the examination as well as helps to identify content areas that may be problematic for students. The validity and reliability of an examination finally depend on the characteristics of the items. Item analysis permits a high reliability and validity to be built into a test in advance. Discrimination index (D) is part of item analysis that measures the difference in item difficulty between groups of students with high and low marks. The index varies between -1 and 1 where the item ideally should be between +0.3 and +1.0. A highly discriminating item is indicative of students who gained high tests scores got the item correct whereas those who had low test scores got the item incorrect. The objective of our study was to calculate the discrimination indices of Advanced level Biology final examinations at a public post-secondary institution in Malta. The final scores obtained by first-year students over a five-year period (n = 1315), 2014-2018, in Papers 1 (short-type items) and 2 (comprehension, structured and unstructured essays) were used to calculate the discrimination index for each item. Results are encouraging since negative discrimination, indicative of a defective item, were not observed in any of the items. Paper 1 is better at discriminating between high and low achievers since over the study period, 93% of the items had acceptable (D between 0.2-0.29) or good discrimination (D between 0.3-0.39) while less, 54%, in Paper 2. Also, fewer (5%) of the items in Paper 1 had poor discrimination (D < 0-0.19) but 46% in Paper 2. Results show that comprehension items are better than the essay type to discriminate between high and low achievers. This finding may be used to start a discussion at the institution to consider the validity of the essay-type of items in final Biology examinations.

Keywords: *Biology, Discrimination index, Malta, Post-secondary.*

1. Introduction

Item analysis is a process that examines learner responses to individual test items in order to assess the quality of those items and of the test as a whole.

Discrimination Index (D) is an instrument to measure the difference in item difficulty between groups of students with high and low marks. Item discrimination power has been described by MacDonald et al., (2002) to indicate the extent of an item to differentiate students with different ability levels. Analysing the discrimination indices of each item provides information regarding what the students have learned and enables teachers to determine and correct the faulty items. The calculation of the discrimination index provides a valuable tool in designing the test. Many studies encountered in the literature investigate multiple choice items as this is a widespread method of student assessment in colleges and universities. According to Khan et al., (2015), there are very scanty studies of on essay type, structured essay type and short answer type questions. These authors and Johari et al., (2012) have applied it to essay type, structured essay type, and short answer questions.

The discrimination index varies between -1 and 1, where the item should have a positive discrimination index of at least 0.2. If the item equals to 0, it means that there is no discrimination. However, items with negative indices need to be revised. Over twenty discrimination indices exist in the literature that may be applied to multiple choice or subjective items. The discrimination index for subjective items used in this study follows the calculation as used by Johari et al., (2012):

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

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 item.
$Score_{min}$ = minimum marks for the item.

If the test and an item measure the same ability or competence, it would be expected that those having a high overall test score would have a high probability of being able to answer the item. Thus, a good item should discriminate between those who score high on the test and those who score low. Table 1 shows the discrimination values and their corresponding interpretation and recommendations as described by Ebel (1972).

Table 1. The discrimination values and their corresponding interpretation and recommendations as described by Ebel (1972).

Discrimination Index	Description	Recommendations
D = negative	Defective Item	Rejected or improved
D < 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

2. Objectives of The Study

To find:

1. the mean discrimination index of Biology Papers 1 and 2 examinations taken by Maltese post-secondary students over the period 2014-2018.
2. the discrimination index of each item in both papers.
3. if the discrimination index of Paper 2 items (comprehension, structured essay and unstructured essay) differs.

3. Methodology

This is a retrospective study based on final examination scores attained by first-year Biology Advanced students attending a public post-secondary institution. The entire cohort (n = 1315), including both male and female students, attending the institution over the period 2014-2018 was investigated. The number of students attending the institution varied with the year investigated, and ranged from a maximum in 2014 (n = 315) to a minimum in 2018 (n = 215) (Table 3). Student scores provided by the Biology Department were written against an index number. Data was entered in a Microsoft Excel 2010 sheet to allow the computation of the discrimination index as explained previously. The list of scores obtained for Paper 1 was arranged in a descending order. The sum of the scores for the top 25% of the student population [Σ(H)] and that for the bottom 25% [Σ(L)] were obtained. The discrimination index was then calculated as per equation given in the introduction. The same exercise was repeated for Paper 2 results. SPSS version 24 was used to analyse the data.

3.1. The Examination

Details of the set-up of the Biology examination papers that students sat for are shown in Table 2.

Table 2. Total marks, and type of item and answer for each paper. Each paper lasts 3 hours.

Paper	Type of item	Type of answer	Total number of marks
1	Short	Compulsory	100
2: Section A	Comprehension	Compulsory	25
Section B	Structured essay	2 to choose 1	25
Section C	Unstructured essay	4 to choose 2	50

4. Results and Discussion

4.1 Research objective 1

To find the mean discrimination index of Biology Papers 1 and 2. According to Anigbo (2015), the ideal discrimination index is between +0.3 and +1.0. Table 3 shows that the mean discrimination index of Paper 1 examinations ranged between 0.29-0.35: being of acceptable discrimination in 2015 and good in the other four years investigated. However, for Paper 2, it was poor (D = 0.19) in 2016 and 2017 and acceptable (D between 0.2-0.29) in the rest. Thus the mean discrimination index for Paper 1 was ideal in four out of five years but never for Paper 2.

Table 3. The mean discrimination index and standard deviation of Papers 1 and 2 over the five-year study period.

Year	Number of students	Mean and Standard deviation of the Discrimination Index & Description	
		Paper 1	Paper 2
2014	315	0.33±0.073 (Good)	0.20±0.122 (Acceptable)
2015	276	0.29±0.062 (Acceptable)	0.21±0.136 (Acceptable)
2016	232	0.30±0.081 (Good)	0.19±0.094 (Poor)
2017	277	0.34±0.058 (Good)	0.19±0.068 (Poor)
2018	215	0.35±0.085 (Good)	0.21±0.122 (Acceptable)

4.2 Research objective 2

To find the discrimination index of each item in Papers 1 and 2. An item answered correctly by low achievers but not by the high achievers possesses negative discrimination. Results of the discrimination index for final Biology papers are encouraging as no negative values were observed in Paper 1 (Table 4) and Paper 2 (Table 5). Items with a negative discrimination index are useless and reduce the validity of the test. An item which does not discriminate between the upper and lower achievers, contributes nothing to the establishment of an order of merit.

Over the whole study period (Table 4) the percentage of Paper 1 items could be classified as follows:

- 7.1% (4 items out of 56) had poor discrimination (D < 0-0.19);
- 26.8% (15 items out of 56) had acceptable discrimination (D between 0.2-0.29);
- 50.0% (28 items out of 56) had good discrimination (D between 0.3-0.39);
- 1.8% (1 item out of 56) had very good discrimination (D = 0.4);
- 14.3% (8 items out of 56) had excellent discrimination (D > 0.4).

Thus, over the five-year period investigated, the majority of the Biology Paper 1 items were able to discriminate and were valid. Due to the nature of the question given, the probability of weak students to give answers on guessing basis was eliminated. According to Brown (1983) and Crocker & Algina (1986), if D > 0.2, the item is acceptable and able to discriminate between good and weak students. If this information is applied to Paper 1 items (Table 4), 93% (52 out of 56 items) fall in this category. Figure 1 is a plot of the percentage of items in each year classified by the description of the discrimination index. The figure indicates that over the years, especially from 2016 to 2018, there is a shift for more items to be classified as having good to excellent discrimination. The more items classified as highly or moderately discriminating, the better the test (Ebel & Frisbie, 1991). Khan et al., (2015) note that although essays, structured essays and short answer questions are frequently given in examination, they are studied to a lesser extent compared to multiple choice questions. Results obtained in this study resemble those of Khan et al., (2015) in two ways. The first is that no negative values were obtained and the second, that a comparable percentage of items having a discrimination index of 0.2 and above. In this study 93% were like so and 87.5% in the study by Khan et al., (2015).

Table 4. The discrimination index for each item (total 56) in Paper 1 classified by year. Values in bold type are greater than 0.3 (good discrimination and better).

Item	2014	2015	2016	2017	2018
1	0.28	**0.35**	0.27	**0.46**	0.29
2	**0.35**	**0.31**	**0.44**	**0.35**	**0.43**
3	**0.45**	0.26	0.26	**0.34**	**0.49**
4	**0.37**	**0.32**	0.20	0.27	**0.45**
5	**0.40**	0.18	0.26	**0.31**	**0.37**
6	**0.36**	**0.32**	0.20	**0.31**	**0.33**
7	**0.30**	**0.31**	0.21	**0.43**	**0.37**
8	0.15	0.16	**0.41**	0.29	0.19
9	**0.38**	0.29	**0.32**	**0.32**	0.29
10	0.24	**0.33**	**0.39**	**0.31**	**0.31**
11	**0.39**	**0.36**	0.29		
12	**0.35**				
13	**0.32**				
14	0.29				

Figure 1. Percentage of Paper 1 items classified by the description of the discrimination index for each year investigated.

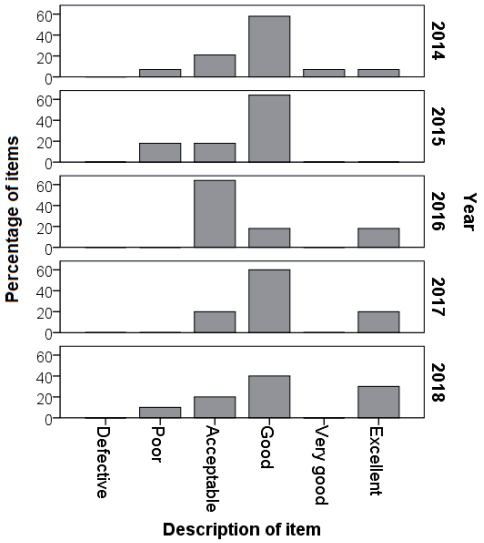
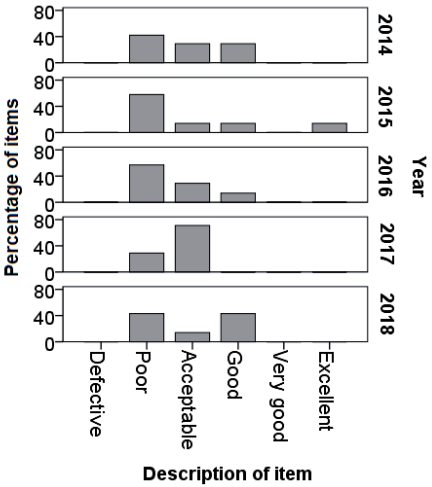


Table 5 shows that no Paper 2 items with a negative discrimination index were obtained over the period 2014-2018. This indicates that none of the items were defective. This is important to know since in many instances tutors refer to past paper items when designing new ones. As the material to be examined does not cover the whole Biology syllabus, tutors are limited to the type of items to set, especially the unstructured essay type. From Table 4, it is observed that over the five-year period of investigation, the discrimination index value ranged from 0.00 in Item 5 (2014) to 0.48 in Item 5 (2015). The percentage of items in each year was classified by the discrimination index and plotted against the description in Figure 2. This figure shows that in each year, except 2017, the highest percentage of items were classified as poor (D < 0-0.19). In fact, over the whole study period, 46% (16 items out of 35) had poor discrimination, 31% (11 items out of 35) had acceptable discrimination (D between 0.2-0.29), 20% (7 items out of 35) good discrimination (D between 0.3-0.39) and 3% excellent discrimination (D > 0.4). From Tables 4 and 5, it is observed that the majority of Paper 1 items (66%) fall within the ideal range (+0.3 and +1.0) showing that items were of good quality, but fewer (23%) in Paper 2.

Table 5. The discrimination index for each item in Paper 2, classified by year and item type. Values in bold type are greater than 0.3 (good discrimination and better).

Item number	Type of item	Discrimination Index				
		2014	2015	2016	2017	2018
1	Comprehension	**0.37**	**0.32**	**0.36**	0.28	**0.32**
2	Structured essay	0.25	0.20	0.18	0.16	0.26
3		0.07	0.15	0.15	0.24	0.12
4	Unstructured essay	**0.32**	0.10	0.10	0.20	**0.34**
5		0.00	**0.48**	0.27	0.21	0.11
6		0.19	0.05	0.21	0.21	**0.33**
7		0.23	0.15	0.06	0.05	0.01

Figure 2. Percentage of Paper 2 items classified by the description of the discrimination index for each year investigated.

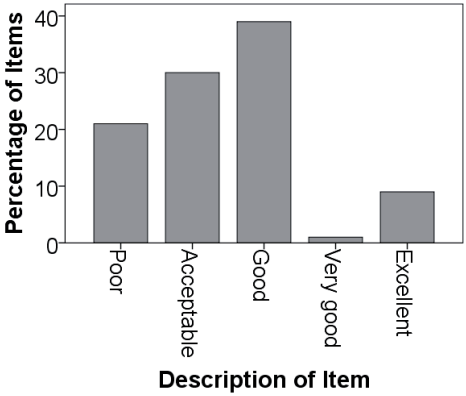


A discrimination index of zero was obtained only once in Paper 2 and never in Paper 1: the unstructured essay (Item 5 in 2014). This indicates that the item was either obtained correctly by everybody or incorrectly by everybody. In fact this was a difficult item chosen by 25% (79 out of 315) students and the mean score was 3.78 $\pm$ 3.039 (maximum: 25 marks). A discrimination index should be evaluated with reference to the difficulty level of the item, because a correlation method is used to assess the item's success in discriminating between low and high achieving students. If the items are very easy or difficult, indicating homogenous performance, there is less variation in the scores, thus resulting in a reduced potential for discrimination. Azzopardi & Azzopardi (2019) calculated the difficulty index values for Paper 2 items and results showed that Item 5 in 2014 was classified as too hard.

Figure 3 shows that, when combining both papers, 21% of the items had a poor discrimination, 30% acceptable discrimination and 39% had good discrimination. Results of this study are different from those reported by Khan et al., (2015) who investigated the discrimination index of essays, structured essays and short questions of 150 medical students. They reported 12.5% poor discrimination (D < 0-0.19), 75% of items with good discrimination (D between 0.3-0.39) and 12.5% acceptable (D between 0.2-0.29). The importance of item analysis is highlighted from this study. Clearly, an improvement of examination items with high discrimination should be implemented into future examinations to improve test scores and properly

discriminate among the students. A high discrimination index indicates a higher probability for students with a robust knowledge to answer the item correctly, while those with less knowledge will answer wrongly. The values of the discrimination index fall between -1 and 1. Ideally, an item should be able to distinguish the more able from the less able students, meaning that the high achievers get the item correct while the low achievers get it wrong. According to Anigbo (2015) the ideal discrimination index is between +0.3 and +1.0, but in this study only half of the items fall in this range.

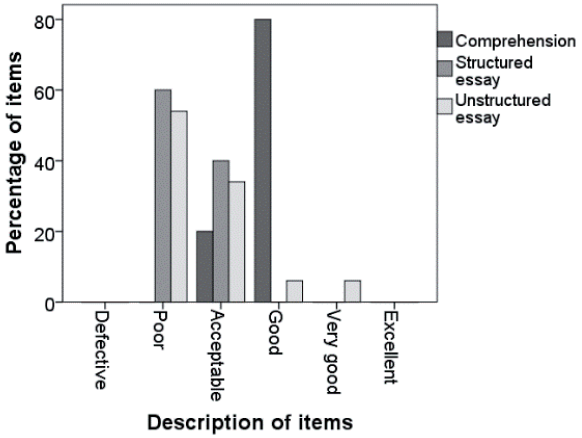
Figure 3. Percentage of Paper 1 and 2 items classified by the description of the discrimination index for all years investigated.



4.1 Research objective 3

To find out if the discrimination index of Paper 2 items (comprehension, structured essay and unstructured essay) differs. A one-way ANOVA between subjects was conducted to compare the discrimination index obtained in comprehensions, structured essays and unstructured essays. There was a significant difference in the discrimination index at the $p < 0.05$ level for the three types of question [$F(2, 32) = 4.568$, $p = 0.018$]. This difference is also illustrated in Figure 4 which shows a classification of the percentage of items according to the discrimination index for the items that appeared in Paper 2 in the five-year study. The figure shows that 80% of the comprehension items had good discrimination. However, a large percentage of items for structured essays (60%) and unstructured ones (54%), had poor discrimination. Thus it may be said that comprehension items are the best to discriminate between high and low achievers.

Figure 4. Percentage of comprehension, structured and unstructured essay type of item in Paper 2 classified by the description of the discrimination index for the five-year period investigated.



5. Conclusion and recommendation

The conclusion drawn is that Paper 1 is better at discriminating between high and low achievers than Paper 2. Structured essays and unstructured essays are not as good as comprehensions at discriminating such groups. This finding may be used to start a discussion at the institution to consider the validity of the essay-type of items in final Biology examinations. This is the first study ever to be conducted on Biology final examinations at the public institution concerned, regarding the discrimination index. Thus so far, test developers constructed novel items each year, especially for Paper 1, running the risk of reducing the reliability of the test by including, unknowingly, items with low discrimination. Constructing items for final examinations is time-consuming and requires careful selection of content that will produce the desired test results. Item analysis, that includes the discrimination index, provides valuable information for further item modification and future test development. Knowledge of the findings of this study opens a discussion at the institution to consider changing the method of setting up items. Since information is now available for items that appeared during the period 2014-2018, paper setters may choose past items having the ideal discrimination index. Extending the analysis of papers prior to 2014, amplifies the data base for future examinations.

References

Anigbo, L.C. (2015) Item Analysis of National Examination Council Senior School Certificate Examination Economics Objective Tests. *International Journal of Innovative Education Research.* [Online] 3 (4):23-3. Available from: https://www.researchgate.net/publication/309311077_Item_Difficulty_and_Discrimination_Indices_of_Multiple_Choice_Biology_Tests_Adekunle_Thomas_Olutola_doi_httpdxdoiorg107828ljherv11i1890

Azzopardi, M. & Azzopardi, C. (2019) The statistical analysis and evaluation of final advanced level biology examinations, In J. Vopava, V. Douda, R. Kratochvil, M. Konecki (Eds.), Paper presented at the Multidisciplinary Academic Conference 19-21 April, 2019 in Prague. *Proceedings of The 14th IAC 2019* (pp. 14-22).

Brown, F.G. (1983) Principles of educational and psychological testing (3rd Ed.). New York: Holt, Rinehart and Winston.

Ebel, R.L. (1972) *Essentials of Educational Measurement* (1st Ed.). New Jersey: Prentice Hall. 634 pp.

Ebel, R.L. & Frisbie, D.A. (1991) *Essentials of Educational Measurement* (5th Ed.). New Delhi: Prentice Hall of India Pvt. Ltd. 370pp.

Crocker, L. & Algina, J. (1986) *Introduction to classical and modern test theory*. New York: Holt, Rinehart and Winston. 527 pp.

Johari, J., Sahari, J., Abd Wahab, D., Abdullah, S., Abdullah, S., Omar, M. Z., & Muhamad, N. (2011) Difficulty index of examinations and their relation to the achievement of programme outcomes. In *Procedia - Social and Behavioral Sciences.* [Online] 18, 71-80. Available from: https://www.sciencedirect.com/science/article/pii/S1877042811011244 [Accessed 20 August 2018]

Khan, G.N., Ishart, N. and Khan, A.Q. (2015) Using item analysis on essay types questions given in summative examination of medical college students: facility value, discrimination index. *International Journal of Research in Medical Sciences.* [Online] Jan3(1):178-182. Available from: DOI: 10.5455/2320-6012.ijrms20150131 [Accessed 20 August 2018]

Macdonald, P. & Paunonen, S. V. (2002) A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*. [Online] 62, 921-943. Available from: https://journals.sagepub.com/doi/abs/10.1177/0013164402238082 [Accessed 20 August 2018]

Bio-notes

Dr Marthese Azzopardi is Senior Lecturer with the Department of Biology at the University of Malta, Junior College. In 1988 she graduated B.Ed. (Hons.) in Biology and Chemistry, M.Sc. in 1996, Diploma in Education (Administration & Management) in 2001 and obtained her doctorate in 2009 from the University of Malta. She has been teaching Biology for 30 years. She has prepared slide presentations at [www.slideshare.net/secbio]. She has both published and presented findings from her research at a number of conferences over the past years, the latest one being '*The Statistical Analysis and Evaluation of Final Advanced Level Biology Advanced Level topics.'*

Carmel Azzopardi is Senior Lecturer with the Department of Physics at the University of Malta, Junior College. He obtained B.Ed. (Hons.) in Physics and Maths in 1991, Diploma in IT in 2000 and M.IT in 2008. He has been teaching physics since 1991 at all levels in various schools and is the author of the book '*Mathematical Requirements for Advanced and Intermediate Level Physics*'. He also took part in the making of a TV series entitled *Physics Highway*. He created a series of videos about physics experiments and demonstrations uploaded on Youtube (www.youtube.com/carmelazzopardi). He presented a number of papers including '*The impact of teaching style and FCI gain on the performance in Mechanics test*' and '*Diagnosis of misconceptions about Force and Motion held by first-year post-secondary students.*'