PREPRINT

Quality of Multiple Choice Questions
in a Numerical and Statistical Methods Course

J P Cruz[1,2,3]
A Freitas[2,3]
P Macedo[2,3]
D Seabra[3,4]
[2]Assistant Professors at Mathematics Department
[3]Center for Research & Development in Mathematics and Applications
[4]Assistant Professor at Polytechnical School ESTGA
University of Aveiro
Aveiro, Portugal
E-mail: {pedrocruz, adelaide, pmacedo, dfcs}@ua.pt

Conference Key Areas: Discipline-specific Teaching & Learning; Quality Assurance and Accreditation; another topic relevant to the conference but not listed above

Keywords: Item Response Theory, Multiple Choice Questions, Numerical and Statistical Methods

INTRODUCTION

The quality control of written examination is very important in the teaching and learning process of any course. In educational assessment contexts, Item Response Theory (IRT) has been applied to measure the quality of a test in areas of knowledge like medicine, psychology, and social sciences, and its interest has been growing in other topics as well. Based on statistical models for the probability of an individual answering a question correctly, IRT can be addressed to measure examiners' ability in an assessment test and to estimate difficulty and discrimination levels of each item in the test. In this work, IRT is applied to Numerical and

[1] Corresponding Author:
J P Cruz
pedrocruz@ua.pt

Statistical Methods course to measure the quality of tests based on Multiple Choice Questions (MCQ).

The present study focuses on three school years, namely 2015, 2016 and 2017, more specifically on the 1st semester of the 2nd year of the degree course. It has involved more than 300 students in each year, and it points out questions (also called items) from some chapters of the program that were evaluated through MCQ. Emphasis is given on the range of item difficulty and item discrimination parameters, estimated by IRT methodology, for each question in those exams. We show where each partial exam explores ability levels: at a passing point or at more demanding levels.

After the application of IRT to each test, which was composed of eight questions, we got 48 item difficulty and item discrimination parameters. The application of standard boxplots shows few atypical responses from students in terms of extremal values of difficulty and discrimination, which corresponds to MCQ that deserve further attention.

We have concluded that the vast majority of questions are well posed considering that they are designed to focus on the cut-off point (passing/not passing). A proposed reflection, about the learned benefits from 'good' outliers and possible causes for those 'bad' items, suggests future improvements to classes, study materials and exams.

1 GENERAL

1.1 Context

Numerical and Statistical Methods is a curricular unit with those components isolated from each other and it makes part of the curricula of several engineering courses since its creation in 2004. Each component is examined in two folds with equal weights: using Open Questions and Multiple Choice Questions (MCQ), where only one out of four choices is correct. In 2014 we started using automatic digital scan correction of MCQ answer sheets [1] and in the years that followed, that were 2015, 2016 and 2017, we used exactly the same scheme for exam moments: the first three chapters of the numerical component (Errors, Interpolation and Numerical Integration) and the first two chapters of statistics (Exploratory Analysis and Distributions) were evaluated using MCQ.

Inspired by the question 'Do you evaluate your examinations in your courses?,' posed in a seminar [2], the quality of MCQ used to individual evaluation during those three mentioned years was investigated. It must be noted that the course has had a stable teaching staff and the same curricula during the study period.

We have been collecting data from the application of Item Response Theory (IRT), since we have started using an automatic correction method of MCQ. IRT model has a long tradition in social and psychology sciences, in what the analysis of personal traits is concerned, as well as in medicine courses to evaluate the quality of exams (e.g., [3]),. It has also been applied to engineering and other sciences (e.g., [4]), so this method is a widely used instrument for the study of the exams quality (e.g., [5]).

1.2 Item Response Theory Summary

In our context, an item (each posed MCQ) has a binary value as result - if it is correct it is valued 1 and if it is incorrect or ignored it is valued 0. A powerful feature of IRT in characterizing each item (question) is the so called latent traits [6]. They are called "latent" due to the fact that they are not directly observed. In IRT, the probability of an individual with ability $z \in \mathbf{R}$ to

respond correctly to each question can be estimated using regression models with one, two or three latent traits. In order to evaluate latent qualities of each MCQ, we propose a model with two latent traits: the difficulty and discrimination parameters by MCQ. Objectively speaking, the probability of an individual with ability z to respond correctly to MCQ (*i*), with difficulty (β_{0i}) and discrimination (β_{1i}), is estimated by a logistic function defined by

$$\mathrm{p}(\mathrm{i}, z) = \frac{1}{1+\exp\left(-\beta_{1i}(z-\beta_{0i})\right)}, \quad i = \text{item}, z = \text{ability}.$$

The curve of this function is the Item Characteristic Curve (ICC) and an example is given in Fig. 1. For a given ability, we get the probability of choosing a correct answer to a given item (in this case, three curves for questions about the Poisson distribution, Normal distribution and Bayes rule).

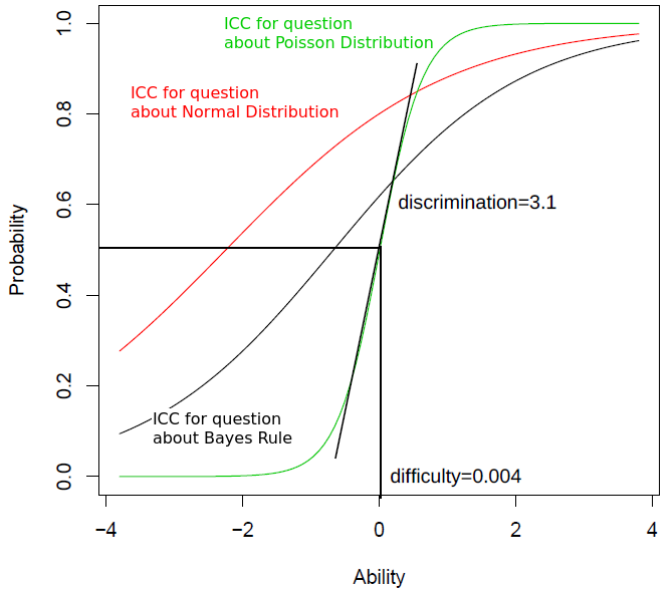

Fig. 1. Item Characteristic Curves (ICC)

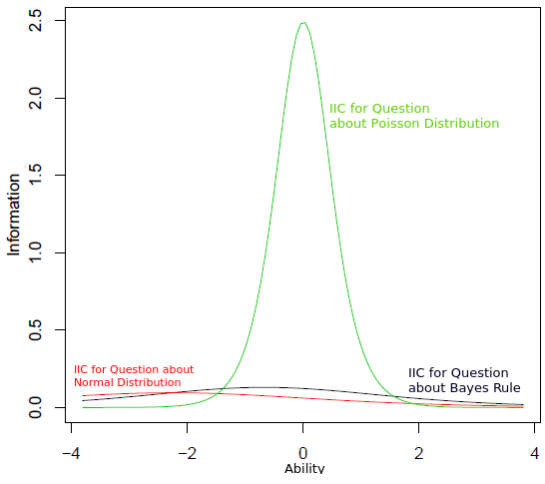

Fig. 2a. Item Information Curves (IIC)

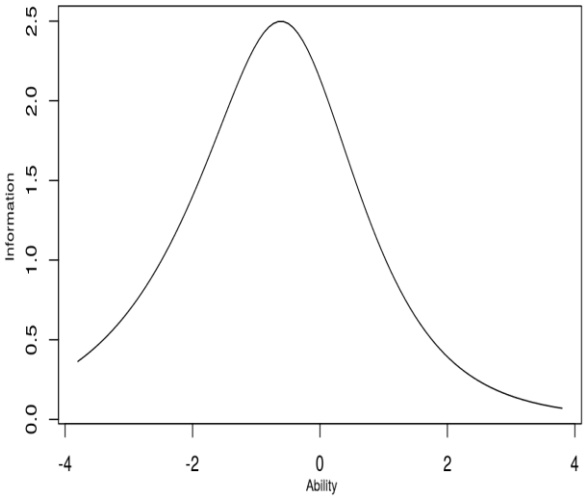

Fig. 2b. Test Information Function for 'Statistics/2015'

In the example above, the curve for the named 'Poisson' question, Fig. 1, has difficulty level $\beta_{0i} = 0.004$ and discrimination level $\beta_{1i} = 3.1$ (here *i* is the named 'Poisson' question). The difficulty parameter is described as the necessary ability to have 50% probability to answer correctly (median ability). The discrimination parameter is the slope of the curve at this point and it is intended to determine how well an item differentiates the performance of respondents. If a respondent presented lower ability to answer an item, we expect lower probability to answer correctly, and when a respondent presented higher ability to respond correctly, then we expect greater probability that his answer is correct.

Another informative curve is provided by the graphic of the Item Information Function, which is defined by a normalized version of the derivative $\frac{\partial \mathrm{p(i,z)}}{\partial z}$. This curve gives a visual perspective on where an item is more discriminative for a given ability level (z). For instance, in Fig. 2a, the question about Poisson distribution is discriminating much more, among the range of ability levels, than the other two questions.

The sum of all Item Information Functions, over all items in a test, defines the Test Information Function (TIF). From its curve (an example in Fig. 2b), one can see if a test gives more information about the requirement for the passing grade ability or if it focuses on higher ability levels. An application of this curve, which is described in the next section, shows different ranges of ability being discriminated.

1.3 Application

We will start by mentioning the oscillatory behaviour of the required ability at each exam. The study was done by plotting a Test Information Function (see above), for each of the six tests in the three years under study. *Table 1* shows the ability unitary length intervals, for each examination, in which the curve has its peak of information (see in Fig. 2b that the unitary interval occurs in [-1,0]). These six intervals show an oscillatory pattern to differentiate ability. In 2015, the first test contained questions that showed that less ability was needed to answer correctly compared to the second test of the same semester. The same phenomenon happened in 2016, but with less magnitude. In 2017, there was an effort to evaluate ability to a more central level. However, the decision to use hard questions has produced a test in which the peak of information distances 3 units of ability from the first test in 2015. Questions in the second test matched the capacity to discriminate ability of the first test in 2015. We recall that these MCQ tests are only 50% of the student grade and, yet, no study has been done for the Open Questions.

Table 1. Most Informative Ability Intervals

2015		2016		2017	
Numerics	Statistics	Numerics	Statistics	Numerics	Statistics
(-2,-1)	(-1,0)	(-1.5,-0.5)	(-1,0)	(+1,+2)	(-2,-1)

1.4 Study of the Boxplot Outliers

Next, based on the IRT method, we have studied some Multiple Choice Questions (MCQ) that called our attention and are related to the previously presented oscillatory effect in ability demanding. As described in the introduction, we have studied the difficulty and discrimination parameters. IRT has been applied to six tests, containing eight MCQ each. We have made two standard boxplots for the 48 MCQ, one for each parameter. What follows is the study of

questions characterized by values of difficulty or discrimination that are outliers in one, or both, of the boxplots.

The boxplot describing the difficulty has values ranging from -4.7 to 404.4, where the standard values are from -3 to 3. There are six outliers' cases and, when removing them, we obtain the histogram of the difficulty parameter, characterizing 42 MCQ, in Fig. 3. We can conclude, for the course under study, that MCQ have typical values for the difficulty parameter in a IRT analysis. Half of the MCQ has been rated within the difficult level from -1.5 to -0.5.

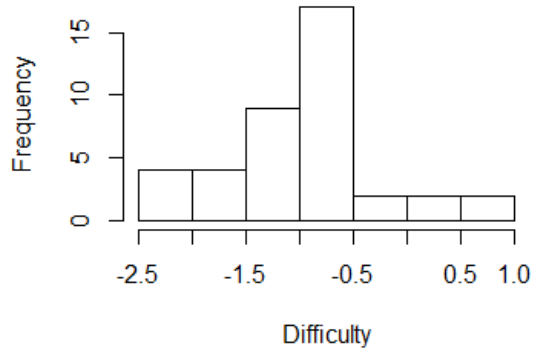

Fig. 3. Histogram of difficulty level estimates, after removing six outliers

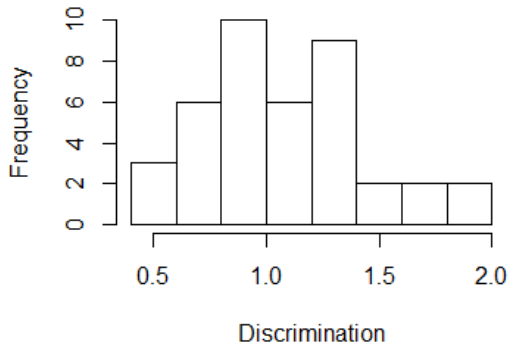

Fig. 4. Histogram of discrimination level estimates, after removing three outliers

In the case of the discrimination parameter, only three outliers have been observed for the 48 questions. After removing the outliers, the histogram in Fig. 4 shows good values for discrimination in the majority of the questions.

Next, we will present questions that cause the atypical behaviour. Although the correct answer is always identified as the first option, the questions and the options within each question were both mixed by the software (see [1]) in all the six exams.

1.5 Questions with small value of difficulty parameter

Three 'easy' questions were rated with difficulties -4.473 and -2.805, in the numerical part (years 2015 and 2017), and -2.845 in the statistical part (year 2016). The two numerical questions are about Lagrangian interpolation with three points and a standard trapezoidal integration. This simple type of questions is presented and practiced in several moments in classes. Therefore, we believe this is the reason for the required lower difficulty levels.

The next atypical question is a statistics question; see Fig. 5. A student, using only a bird-eye look and, without ability in statistics, easily and intuitively chooses the first two options as targets. At this time, it's easy to fulfil the task because the correct option is very close, in form, to the formula in the question's text. This possible reasoning, outside knowledge about statistics, combined with students that know how to solve it, turn this question into an 'easy' one.

Question [Pbb2] Let A and B be two events such that $P(A) > P(B)$ and $P(A \cap B) \neq 0$. Then

- [x] $P(A|B) > P(B|A)$.
- [ ] $P(A|B) < P(B|A)$.
- [ ] $P(A) = 0$.
- [ ] (None of the other three options is correct.)

Fig. 5. Question 'Pbb2' with difficulty= -2.845 and discrimination= 0.840

1.6 Questions with higher value of difficulty parameter

Question in Fig. 6 is almost a standard question in the Numerical and Statistical Methods course, with one exception: the problem description has the expression $f^{(4)}(x) \in [-5,2]$. Usually, an upper bound of $|f^{(4)}(x)|$ is obtained using a calculator to draw a plot of a given fourth derivative function. This slight change in the question's text caused a more demanding ability to answer this question correctly. The discrimination parameter indicates a good separation between ability levels, which means that even small changes in questions can transform a well-practiced problem into a more difficult one.

Question [Cap2Q2] Consider a function $f \in C^4([-4,4])$ and p_3 the cubic polynomial that interpolates f at the nodes $\{-4,-2,2,4\}$. Assuming that $f^{(4)}(x) \in [-5,2]$, $\forall x \in [-4,4]$, an upper bound for the error $|f(0) - p_3(0)|$ is

- [x] $\frac{40}{3}$.
- [ ] $\frac{16}{3}$.
- [ ] $\frac{8}{3}$.
- [ ] (None of the other three options is correct.)

Fig. 6. Question 'Cap2Q2' with difficulty=0.718 and discrimination=1.651

Question in Fig. 7 is about the Runge phenomenon in numerics. This theme is usually presented in expository classes only, and students rarely practice the concept. We can conjecture that the high ability required to answer correctly is due to the tendency of students to read less theoretical materials. The discrimination parameter is small, revealing that the concept is not well understood, even by students with higher abilities.

Question [Cap2Q1b] Let p_n be the interpolating polynomial of a function f on an interpolation support with $n+1$ equally spaced points in the interval $[a,b]$. Under these conditions,

- [x] there are functions f for which the distance between f and p_n increases as the number of points in the support increases.
- [ ] for any function f, the distance between f and p_n decreases as the number of points in the support decreases.
- [ ] for any function f, the distance between f and p_n decreases as the number of points in the support increases.
- [ ] (None of the other three options is correct.)

Fig. 7 Question 'Cap2Q1b' with difficulty=1.934 and discrimination=0.606

The next problematic situation arises from the definition of Lagrange interpolation: the interpolated polynomial crosses each given point in the support. We believe that almost all students know the formal definition. However, the question's text and its options lead people

to the wrong answer. The discrimination parameter is even smaller when compared to the previous question, certifying that this situation was not well understood by the vast majority of students.

Question [Cap2Q3] Consider an interpolation support for a given function f with 11 equally spaced nodes between -5 and 5, and the respective 11 nodal values. Let p_n be the Newton's divided-difference polynomial of maximum degree in this support. Under these conditions,

- [x] $f(2) - p_n(2) = 0$.
- [ ] it is not possible to use a linear spline because the number of nodes is odd.
- [ ] p_n is the polynomial that always provides the smallest interpolation error in this support.
- [ ] (None of the other three options is correct.)

Fig. 8 Question 'Cap2Q3' with difficulty=13.35 and discrimination=0.158

Our last question under study, which is linked to statistics, has difficulty of 404.4 and almost none discrimination. This has surprised the authors of the question since it had been inspired in a question presented in the course's textbook (though it had only been practiced once in classes). Possible causes for the high value of difficulty in this question are the phenomenon of repeated observations in a sample and the mixture of definitions in a same question. Also, the simplified definition of median, as the value that divides the ordered sample into two halves, does not always help thinking about the possibility of repeated observations in a sample. A study about this issue has been done in [7].

Question [ED] In a recent environmental study, 200 samples of water were collected from a swamp and the nutrient concentration was recorded. It was concluded that: percentile of order 25 = 0.4gr/cm^3 and 3º quartile = mean = 0.5gr/cm^3. Under these conditions,

- [x] the percentage of observations greater than or equal to the mean is not less than 25%.
- [ ] the median of the data will necessarily be a value greater than 0.4 and less 0.5.
- [ ] if samples with atypical levels of nutrient concentration are observed in the boxplot, they will correspond to samples with concentration levels above 0.55.
- [ ] the length of the polygon in the center of the boxplot is equal to 0.5.

Fig 9. Question 'ED' with difficulty=404.4 and discrimination=0.004

2 SUMMARY AND ACKNOWLEDGMENTS

The use of Multiple Choice Questions (MCQ) in written tests, supported by the digital scan and automatic correction, has been evaluated using the application of Item Response Theory methodology. A simple detection of outliers using boxplots of difficulty and discrimination values shows few cases from the 48 questions that required investigation. In overall, the six partial exams contained appropriate questions for the teaching/learning binomial in our Numerical and Statistical Methods course.

The following summary could be useful to future evaluations and teaching/learning formats:

- questions with lower difficult value have been practiced in several moments. This suggests a web tool to help students practicing the basic concepts when time in classes is not enough for all students;

- questions designed to evaluate theoretical concepts could demand a higher ability to be answered but they are not necessarily strong discriminators between different levels of ability. This may represent a problem to define an optimal format of the test: how to design a discriminative question? A possible solution came from one question, where a simple combination of known concepts strongly increased the difficulty value. An electronic database of this type of more demanding problems could be created for students with more ability in this course;
- knowing the properties of difficulty and discrimination can help the team to prepare exams that avoid the oscillatory effect in overall difficulty and avoid the dropout rate.

Given the good results achieved in the last three years, with a careful balance between MCQ and open questions, we plan to keep applying and improving these evaluation procedures in our Numerical and Statistical Methods course.

ACKNOWLEDGEMENTS

This work is supported in part by the Portuguese Foundation for Science and Technology (FCT - Fundação para a Ciência e a Tecnologia), through CIDMA - Center for Research and Development in Mathematics and Applications, within project UID/MAT/04106/2013.

REFERENCES

[1] Bienvenüe, A., Auto-Multiple-Choice software, version 1.3.0. 2016.

[2] Severo, M., (in portuguese) Application of mathematical models of item response in the quality control of multiple choice exams, MATEAS Seminars, https://mateas.wikidot.com, 2015.

[3] Severo, M., Tavares, M. A. F., Meta-Evaluation in Clinical Anatomy: A Practical Application of Item Response Theory in Multiple Choice Examinations, Anat Sci Educ 3:17–24 (2010).

[4] Nevin, E., Behan, A., Duffy, G., Farrell, S. Harding, R., Assessing the validity and reliability of dichotomous test results using Item Response Theory on a group of first year engineering students, The 6th Research in Engineering Education Symposium (REES 2015), Dublin, Ireland, July 13-15, 2015.

[5] Huntley, B., Engelbrecht, J., and Harding, A., Can multiple choice questions be successfully used as an assessment format in undergraduate mathematics?, Pythagoras, (69), 3-16, 2009.

[6] Baker, F. B., Item Response Theory – Parameter Estimation Techniques, Statistics Textbooks and Monographs, volume 129, Marcel Dekker, Inc. New York, 1992.

[7] Freitas, A., Cruz J.P., Silva N., (in portuguese) Median in non-grouped data, “Mediana de dados não agrupados: a questão de ser pelo menos 50%”, Revista da Associação de Professores de Matemática, http://www.apm.pt/files/_18_Mediana_de_dados_59f1fa49eb417.pdf, 2017.