DOI: 10.36315/2019v2end069
Education and New Developments 2019

EVALUATING RELIABILITY AND DISCRIMINATORY CAPABILITY OF BEMA IN TWO SPANISH ENGINEERING DEGREES

M. Amparo Gámiz-González[1], Ana Vidaurre[1], Roser Sabater I Serra[2], Isabel Tort-Ausina[1], María-Antonia Serrano[2], Jaime Riera[1], José Maria Meseguer-Dueñas[1], José Antonio Gómez-Tejedor[1], José Molina-Mateo[1], & Tania Garcia-Sanchez[2]

[1]Departament de Física Aplicada
[2]Departament de Ingenieria Electrica
E.T.S.E. Disseny, Universitat Politècnica de València (Spain)

Abstract

In this work, we analyzed the reliability and discriminatory capability of BEMA (Brief Electricity and Magnetism Assessment) for students of Electricity and Physics courses in Engineering Degrees taught at the School of Engineering Design (ETSID) from Universitat Politècnica de València (UPV). BEMA is a 30-item multiple-choice test designed to assess students understanding of basic electricity and magnetism concepts. The questions are mostly qualitative and some of them require simple calculations. The test is useful when combines validity, reliability and discriminatory capability. The validity is usually determined by expert opinions. The BEMA test is valid for the Electricity and Physics courses because the tested concepts are related to the course subject. A reliable test is consistent within itself and across time. Besides, a large fraction of the variance in scores is caused by systematic variation in the population of the test takers. The reliability of an assessment instrument is particularly important when it is going to be used to compare the performance of different groups. In this work, the reliability and discriminatory capacity of BEMA is assessed statistically. From the post-instructional data three parameters are focused on individual test items (item difficulty index, item discrimination index, item point biserial coefficient) and two parameters are focused on the test as a whole (test reliability and test Ferguson's).

Keywords: *BEMA test, reliability, discriminatory capability, Ferguson's.*

1. Introduction

Brief Electricity and Magnetism Assessment (BEMA) was developed in 1997 by Chabay and Sherwood, aided by Fred Reif, to measure students' qualitative understanding and retention of basic concepts in electricity and magnetism (Ruth Chabay & Sherwood, 1997) (Chabay & Sherwood, 2006). This standardized multiple-choice test is a useful tool to assess students' understanding about electricity and magnetism concepts. The overall performance of a group of students can be obtained through the mean and standard deviation. These parameters enable the comparison between different groups, which can be useful, for instance, to check if a teaching innovation has had a positive effect or not. The fact that the quality of the test is good enough to measure the knowledge of students is implicit in this approach. The standard measures of the quality of a test consider two parameters: validity and reliability. Validity is an estimate of how well the test measures what it intends to measure. The reliability of a test is a measure of how consistently the test will reproduce the same score under the same conditions. Reliability of a test can be stablished by the Kuder-Richardson formula (Kuder & Richardson, 1937).

There are two standard measures of the quality of items on a test: difficulty and discrimination (Maloney, O'Kuma, Hieggelke, & Van Heuvelen, 2001). Difficulty is usually measured by finding the percentage of subjects who get the item correct. Discrimination is a measure of how well an item differentiates between competent and less competent students. Classical item analysis is concerned with a number of item specific statistics such as classical item difficulty, classical item discrimination, and the item point biserial (Eaton, Johnson, Frank, & Willoughby, 2019). Ding et al. (Ding, Chabay, Sherwood, & Beichner, 2006) evaluated the BEMA test after it had been administered to a large number of students at North Carolina State University (NCSU). Their results indicate that BEMA is a reliable test with adequate discriminatory power. In this work, the BEMA test was administered to students of Electricity and Physics courses in Engineering Degrees taught at the School of Engineering Design of the Universitat Politècnica de València (UPV) as both a pre- and post-test. In a previous paper the gain was analyzed (Vidaurre et al., 2019), and the focus here is if the results obtained in NCSU can be extrapolated to our courses in Spain. In this paper, we analyze the reliability and discriminatory capacity of the BEMA test, as measured by statistical tests, focusing both on individual items and on the test as a whole.

2. Design

This paper is focused on the reliability and discriminatory capability of BEMA for students of Electricity and Physics courses in Electronic Engineering and Industrial Automation (EEIA) and Aerospace Engineering (AE) Degrees, taught at the School of Engineering Design of the UPV. The BEMA pre-test was delivered to students during the first week of the course while the post-test was delivered at the end. The pedagogical aspects of both the EEIA and AE were quite similar, and the methodology used in both cases was a combination of flip-teaching (FT) and traditional methodology where the university's e-learning platform was intensively used.

The BEMA test was administrated following the usual instructions (time limit of 45 min, the same grade for all students who completed the test regardless the score) to 116 students out of 154 in the case of EEIA, and 61 out of 78 students in AE. Using all the data obtained, we performed 5 statistical tests, 3 of them focusing on individual test items and 2 of them on the test as a whole (Ding et al., 2006):

1. The **item difficulty index** is calculated as the ratio of the number of correct answers over the total number of students who tried the question, and it is a measure of the difficulty of a single question.
2. The **item discrimination index** measures the extent to which a single test item distinguishes students who know the material well from those who do not. For a specific test item, it relates the number of correct responses in a high-level group to the low-level group.
3. The **point biserial coefficient** is a measure of the consistency of a single test item compared to the entire test. Reflects the correlation between students' scores on an individual item and their scores throughout the test.
4. The **Kuder-Richardson reliability index** is a measure of the self-consistency of a whole test, by dividing a test into its smallest components.
5. The **Ferguson's delta** measures the discriminatory power of an entire test by analizing how widely the total scores of a sample are distributed in the possible range of scores.

3. Results

Since our objective was the evaluation of BEMA, in this paper we study post-test data in order to test statistics. The data and scores expressed as mean (25-75 percentiles) corresponding to the sample of students of the two courses who participated in the study are shown in table 1.

Table 1. BEMA post-test data results.

Course	Number students	Mean (p25-p75)	Standard deviation
EEIA	116	11.0 (7.0-14.8)	4.5
AE	61	11.8 (8.0-14.0)	5.0

Considering the data of each course, we performed the five beforementioned statistical test. The results of these calculations, expressed as mean (25-75 percentiles) for test items, are shown in table 2, where the indicated desired values are close to those obtained from NCSU (Ding et al., 2006).

Table 2. Summary of BEMA statistical test results for the two courses.

Test statistics	Desired values	EEIA	AE
Difficulty index P	≥0.3	0.37 (0.21-0.55)	0.39 (0.25-0.52)
Discrimination index D	≥0.3	0.25 (0.13-0.37)	0.25 (0.09-0.37)
Point biserial coefficient r_{pbs}	≥0.2	0.33 (0.20-0.44)	0.36 (0.21-0.53)
Reliability index KR-21	≥0.7	0.75	0.79
Ferguson´s delta	≥0.9	0.96	0.94

BEMA item difficulty index values range from 0.03 to above 0.9, with about half of the questions between 0.20 and 0.5 with an average difficulty index value around 0.38 for the two courses, which is above the desired value. Questions 28&29 stand out for having the lowest P value for both courses. Regarding the discrimination index D, it has been calculated dividing the groups into two according to the median, and most items have values between 0.1 and 0.4 (22 for EEIA and 18 for AE) with an average value of 0.25 for both courses. This is not in the desired range of values, and for this reason we have recalculated D using

25%-25% calculation and the average difficulty index values obtained have risen up to 0.36 and 0.50 for EEIA and AE respectively, which are above the desired value. Question 9 and questions 28&29 show the lowest D values for both courses.

The average point biserial factor obtained in BEMA is 0.33 for EEIA and 0.36 for AE, which are greater than the desired value of 0.2. This means that we can consider that BEMA items have a good correlation with the whole test. In the two groups, a majority of questions (24) shows a r_{pbs} higher than 0.2, indicating that they are reliable and consistent. Again, it should be noted that question 9 and questions 28&29 are the ones having the lowest r_{pds} values for both courses.

To get the reliability index, the Kuder-Richardson formula (Kuder & Richardson, 1937) has been used, which is the indicated one for a multiple-choice test where each question has only 2 possible answers: correct or wrong. A widely accepted criterion (Doran 1980) is that if the reliability index of the test is higher than 0.7, the test is reliable for group measurements, which is our case for both groups. If the reliability index of the test were higher than 0.8, then the test would be reliable for individual measurements, being the AE group very close to this value. Finally we found the Ferguson´s delta for BEMA test to be around 0.95 for both groups and since it is greater than 0.9 we can consider that the test offers a good discrimination.

4. Conclusions

The results from BEMA from a post-test from students of Electricity and Physics courses in engineering degrees (EEIA and AE at the UPV-Spain) were analysed statistically, focusing on reliability and discriminatory capacity. Post-instruction mean values and standard deviation for both degrees are in good agreement with those obtained in NCSU.

The analysis of the individual test items, by means of difficulty index, discrimination index and point biserial coefficient, shows average values higher than the desired values (adopted criterion) in the introductory E&M courses in both degrees, with slightly higher values for AE for difficultly index and point biserial coefficient. However, questions 28&29 stand out with the lowest values in both courses, indicating that probably the concept related to these questions should be emphasized. In addition, considering the test as a whole, the two indexes analysed (reliability index and Ferguson's delta) also have values higher than the adopted criterion. Based on the obtained results, we can conclude that the use of BEMA as a tool to measure students' understanding in the delivered E&M courses offers adequate discrimination and reliability.

Acknowledgements

This work has been supported by the UPV through the Project of Innovation and Educational Improvement Program (Projects PIME/2018/B26 and PIME/2018/B25 *Convocatoria de Proyectos de Innovación y Convergencia de la* UPV).

References

Chabay, R., & Sherwood, B. (1997). Qualitative understanding and retention. *AAPT Announcer*, *27*, 96.

Chabay, R., & Sherwood, B. (2006). Restructuring the introductory electricity and magnetism course. *American Journal of Physics*, *74*(4), 329–336. https://doi.org/10.1119/1.2165249

Ding, L., Chabay, R., Sherwood, B., & Beichner, R. (2006). Evaluating an electricity and magnetism assessment tool: Brief electricity and magnetism assessment. *Physical Review Special Topics - Physics Education Research*, *2*(1), 010105. https://doi.org/10.1103/PhysRevSTPER.2.010105 G.

Doran, R. (1980). Basic Measurement and Evaluation of Science Instruction. NSTA, Washington, DC.

Eaton, P., Johnson, K., Frank, B., & Willoughby, S. (2019). Classical test theory and item response theory comparison of the brief electricity and magnetism assessment and the conceptual survey of electricity and magnetism. *Physical Review Physics Education Research*, *15*(1), 010102. https://doi.org/10.1103/PhysRevPhysEducRes.15.010102

Kuder, G., & Richardson M. (1937). The theory of the estimation of psychometrika test reliability. *Psychometrika 2*, 151.

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & Van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics*, *69*(S1), S12–S23. https://doi.org/10.1119/1.1371296

Vidaurre, A., Riera, J., Meseguer-Dueñas, J. M., Molina-Mateo, J., Gomez-Tejedor, J. A., Tort-Ausina, I., & Gamiz-Gonzalez, M. A. (2019). Measuring Innovation Effectiviness by Means of a Conceptual Test of Electricity and Magnetism. *INTED2019 Proceedings*, 8728–8733. https://doi.org/10.21125/inted.2019