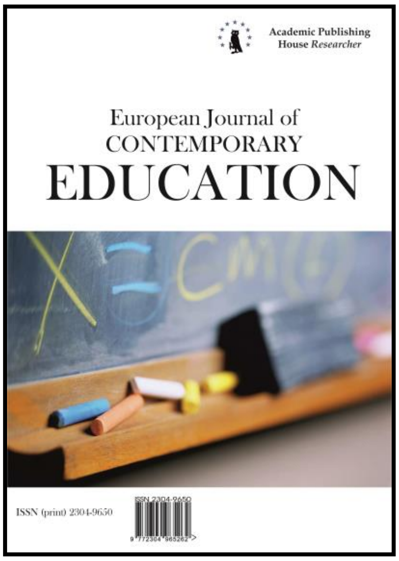
Published in the Slovak Republic
European Journal of Contemporary Education
E-ISSN 2305-6746
2019, 8(4): 864-878
DOI: 10.13187/ejced.2019.4.864
www.ejournal1.com



An Approximation of University Students' Learning Ability in the Area of Probability

Salvador Sandoval-Bravo [a], Pedro Luis Celso-Arellano [a], Victor Gualajara [a], Semei Coronado [b, *]

[a] Centro Universitario de Ciencias Económico Administrativas, Universidad de Guadalajara, Jalisco, México
[b] Independent researcher, San Diego, California, USA

Abstract

The objective of this study is to analyze the ability of students of the University Center for the Economic Administrative Sciences which forms part of the University of Guadalajara from different economic-administrative undergraduate programs, to solve distinct problems in the area of probability, applying a multiple-choice instrument aligned to the learning objectives via a qualitative-descriptive methodology. The study comprised a sample of 251 students from 14 different undergraduate degree programs who were enrolled in the same statistics course. Multivariate tests were conducted in order to identify any differences in performance related to undergraduate degree program and sex, while a RASCH model was applied to provide validity evidence for the assessment. The results show that the students do not have a good level of ability for solving conditional probability problems, which they confused with the formulation of independent event problems although they did show a satisfactory level of ability for solving other types of problems.

Keywords: probability, assessment, homogeneity tests, RASCH.

1. Introduction

Probability has been studied from distinct approaches, from a conceptual level to university curricula and the manner in which it is taught, among others. Contemporaneously, these research topics have been subject to distinct and often controversial interpretations (Batanero et al., 2005; Borovcnik, 2011; Carolyn, Kirk, 2001). These controversies could be due to the fact that probability requires an approach distinct to who we think about and apply reasoning to situations in real life (Batanero et al., 2016).

[*] Corresponding author
E-mail addresses: semei.coronado@gmail.com (S. Coronado), victor.gualajara@academicos.udg.mx (V. Gualajara), pcelso@cucea.udg.mx (P. Luis Celso-Arellano), salvsanb@cucea.udg.mx (S. Sandoval-Bravo)

The *Centro Universitario de Ciencias Económico Administrativas* (CUCEA or University Center for the Economic Administrative Sciences), which forms part of the *Universidad de Guadalajara* (U de G or University of Guadalajara), offers 14 undergraduate degrees in the areas of economics and business. All CUCEA undergraduate students take the same statistics course, the content of which includes the topics of descriptive statistics and basic probability. Since 2005 (Del Toro, Ochoa, 2010), a Departmental Exam (DE) has been applied every semester. The DEs are designed by a commission of professors from the Academy of Statistics (AS) with the aim of measuring the level of knowledge acquired during the semester by students on the course (Del Toro, Ochoa, 2010). Through the *Departamento de Métodos Cuantitativos* (DMC Department of Quantitative Methods), since 2006, CUCEA, in coordination with the AS, has promoted significant and competitive learning in the area of statistics, organizing the annual Statistics Tournament (ST), in which all students taking the statistics course are free to participate.

Descriptive and inferential analyses were conducted with the proposed instrument, with the results corroborated by a RASCH model (RM) (Rasch, 1980) to ascertain the students' ability and, thus, support the findings obtained. Moreover, the RM provided validity and reliability evidence for the proposed instrument (Aziz et al., 2013; Raudzatul, 2016).

Average grades, both for the DE and the ST, were obtained from the SA (Academia de Estadistica, 2018), revealing that the students obtained lower grades for probability topics than for descriptive statistics topics. Thus, for the year in question, the DE students' average score was 61.25 for descriptive statistics and 44.17 for probability questions, while, for the ST students, these were 58.26 and 37.81, respectively (results for previous tournaments have been very similar). With these results showing poorer student performance for probability items, it is necessary to study their learning in the area of probability via an instrument that not only measures what they have learned, but which is also aligned with the results for the statistics course, specifically the topics on probability.

Research has been conducted in the area of probability at distinct educational levels in Mexico. For example, Flores et al. (2014) studied the quality of secondary level students' answers in probability topics, both before and after undertaking technology-based learning activities. On this same research line, but with students in their sixth semester of high school, Landin and Sánchez (2010) analyzed students' probabilistic reasoning as related to the binomial distribution and based on symbolic software.

Sánchez and Inzunza (2006) organized different activities with university students to study their analysis of the meaning of different distributions, such as binomial, uniform, normal and irregular distributions, using a computer as a tool to carry out different simulations. However, a literature review did not identify any studies that determined the ability of university students in the economic-administrative area.

From this point on, the present study is divided into three sections, with Section 2 presenting the methodology and Section 3 the conclusions.

2. Methods

The research was carried out at CUCEA, which, as noted earlier, offers 14 undergraduate degrees in the economic-administrative area. As a requirement for all CUCEA degree programs, the course Statistics I covers different topics, including probability, and is often taken by students in their second or third semester.

The proposed assessment was offered to 57 Statistics I course sections during the first semester of 2019. Due to various circumstances external to the research, only seven course sections took the assessment. With each group comprising between 30 and 40 students, the total sample was 251 students.

The exam was designed in various stage. First, a commission was established by the professors teaching the course in order to decide on the topics which would be on the exam, concluding on the following: combinations; permutations; simple probability; conditional probability; Bayes' theorem; and, binomial and Poisson distributions. It was then decided during the second stage that the number and type of questions to be applied would correspond to 10 multiple choice questions. Moreover, both the correct answer option and the distractors had to be justified in order that they reflected the students' most common difficulties and errors. It was also stipulated that the questions corresponded to problems from the area of business and economics, in order to guarantee that the knowledge acquired was significant for the students.

The third stage consisted in the preparation of a test by each member of the commission, in accordance with the initial guidelines, while the fourth stage reviewed the individual questions and chose the best per topic, on the basis of which, the first version of the test was devised. An independent commission of experts evaluated the test in the fifth stage, making some important suggestions for the modification of the test, which were then implemented. Finally, the group of experts validated and approved the modified version, which was then applied. The exam is available upon request from the authors.

3. Results

Figure 1 presents the number of students participating in the study per undergraduate degree program, of which 40 % were enrolled in public accounting, business administration and marketing undergraduate degree programs, while only 1.2 % were enrolled in information technology undergraduate degree programs. A total of 149 women and 102 men participated, with Figure 2 presenting the distribution by degree and sex.

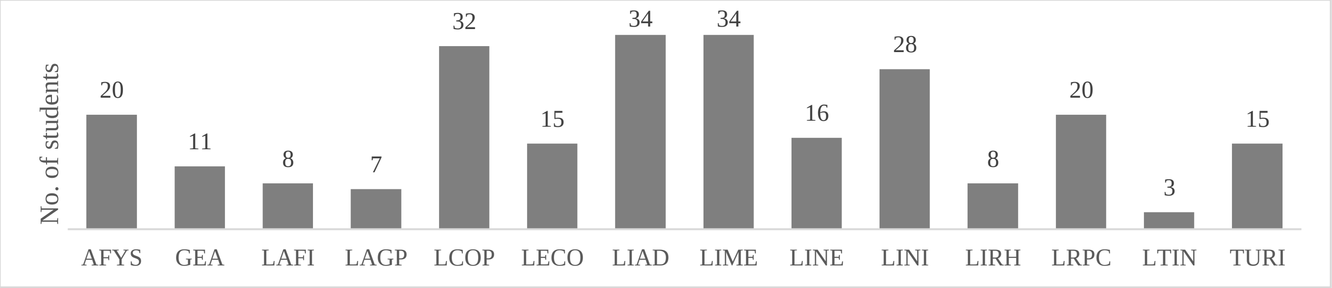


Fig. 1. Number of students per undergraduate degree program

Note: Nomenclature of the different economic-administrative undergraduate degrees offered at CUCEA:

Administración Financiera y Sistemas (AFYS or Financial Administration and Systems), *Licenciatura en Gestión y Encomia Ambiental* (GEA or Undergraduate Degree in Environmental Management and Economics), *Licenciatura en Administración Financiera* (LAFI or Undergraduate Degree in Financial Administration), *Licenciado en Administración Gubernamental y Políticas Públicas* (LAGP or Undergraduate Degree in Government Administration and Public Policy), *Licenciado en Contaduría Pública* (LCOP or Undergraduate Degree in Public Accounting), *Licenciado en Economía* (LECO or Undergraduate Degree in Economics), *Licenciado en Administración* (LIAD or Undergraduate Degree in Administration), *Licenciado en Mercadotecnia* (LIME or Undergraduate Degree in Marketing), *Licenciado en Negocios* (LINE or Undergraduate Degree in Business), *Licenciado en Negocios Internacionales* (LINI or Undergraduate Degree in International Business), *Licenciado en Recursos Humanos* (LIRH or Undergraduate Degree in Human Resources), *Licenciado en Relaciones Publicas y Comunicación* (LRPC or Undergraduate Degree in Public Relations and Communication), *Licenciado en Tecnologías de la Información* (LTIN or Undergraduate Degree in Information Technology), *Licenciado en Turismo* (TUR or Undergraduate Degree in Tourism).

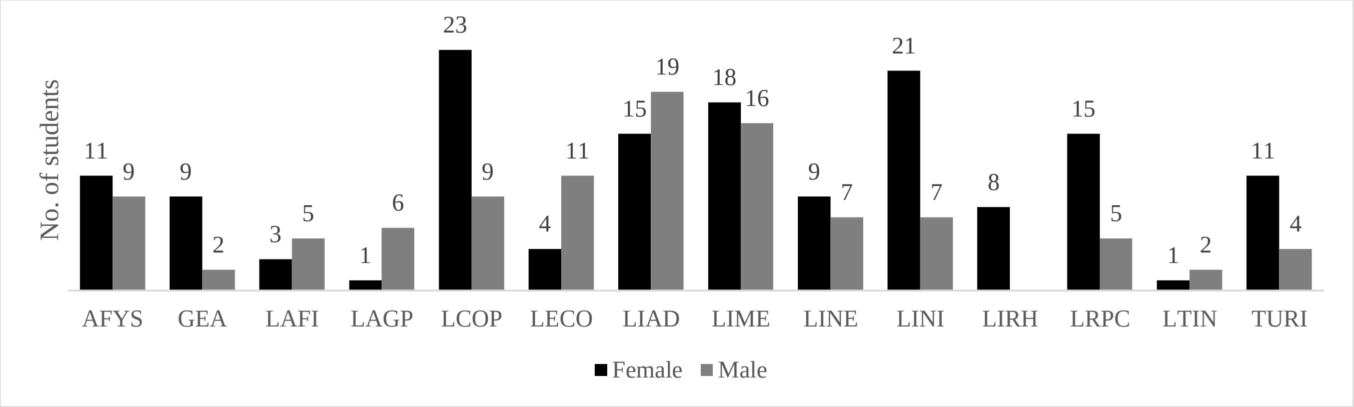


Fig. 2. Number of students by sex and undergraduate degree
Source: Prepared by the author based on the sample

The exam lasted an hour. Binary (one and zero) scoring was applied to the results, in which a correct response was scored 1 and an incorrect response scored 0. Figure 3 presents the percentage of correct and incorrect responses per item, from which it was observed that there were less than 50 % correct responses for items 2 and 6. Items 8 and 9 had 5 1 % and 53 % correct responses, respectively, with between 57 % and 78 % correct for the remaining items. We applied a one proportion z test for each item to support the result. Where the null hypothesis is the proportion of questions correct less or equal than the proportion of questions incorrect at a significance level of 5 %. The result presented is in Table 1. According to the *p*-value in Table 1, we found four items that were not significant. We could infer the student had difficulty solving the problem or that the questions did not clearly establish the problem.

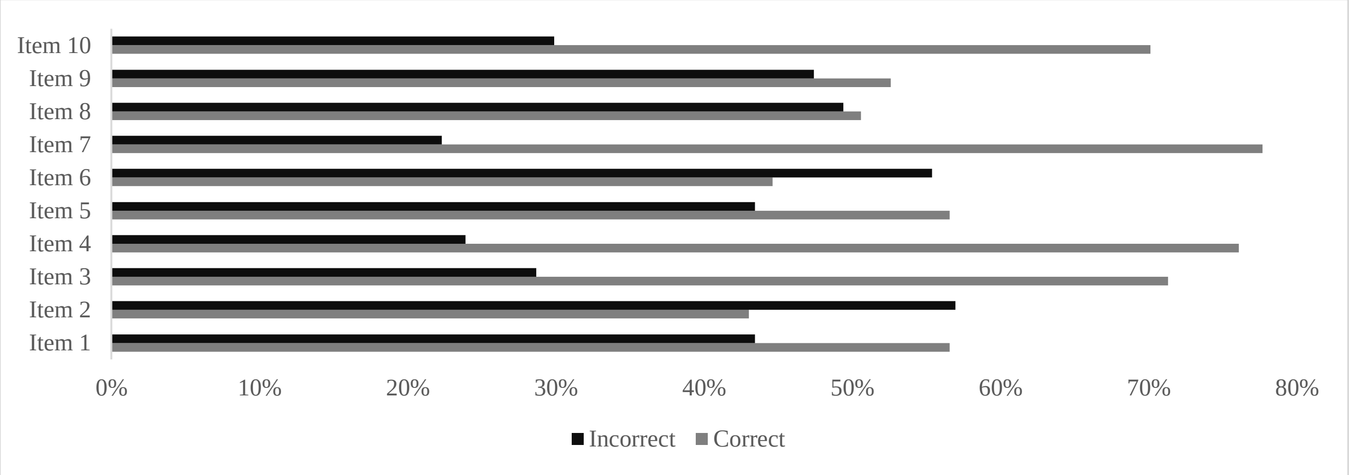


Fig. 3. Percentage of correct and incorrect responses per item
Source: Prepared by the author based on the sample

Table 1. One portion z test to compare the correct and incorrect answers

Item	portion correct	portion incorrect	z	*p*-value
1	0.57	0.43	2.22	0.0186
2	0.43	0.57	-2.22	0.9864
3	0.71	0.29	6.65	0.0000
4	0.76	0.24	8.24	0.0000
5	0.57	0.43	2.22	0.0186

6	0.44	0.56	-1.90	0.9558
7	0.78	0.22	8.87	0.0000
8	0.51	0.49	0.32	0.4249
9	0.53	0.47	0.95	0.2060
10	0.70	0.30	6.34	0.0000

Source: Prepared by the author based on the sample

Table 2 presents the percentage of correct and incorrect responses per undergraduate degree. For some undergraduate degree programs, such as LRPC and LECO, there were more than 50 % correct answers for all of the items, while, for other programs, there were more than 50 % correct answers for all but one of the items, such as LINI and LAGP. For GEA and LINE, there were only two items with more than 50 % correct answers. This difference is due to the fact that, per degree program, the course sections do not seem to be homogeneous with respect to mean performance on the assessment.

Thus, assuming homogeneous covariance across these sub-course sections and using a confidence level of 95 %, the following four tests were applied to test the hypothesis that the mean performance on the assessment is the same across all the sub-course sections: Wilks' lambda (Wilks, 1932); Pillai's trace (Pillai, 1955); Lawley-Hotelling trace (Lawley, 1938); and, Roy's largest root (Roy, 1957). The four tests reject the hypothesis of identical means across the degree programs, according to the *p*-value; therefore, it is unlikely that they have equal means. Thus, they are independent. The results are presented in Table 3.

Table 2. Percentage of correct and incorrect responses by item and degree program

AFYS										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	**60%**	30%	**85%**	**70%**	**65%**	40%	**85%**	45%	**55%**	**85%**
Incorrect	40%	70%	15%	30%	35%	60%	15%	55%	45%	15%
LIME										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	29%	29%	**79%**	**76%**	50%	41%	**62%**	41%	47%	**59%**
Incorrect	71%	71%	21%	24%	50%	59%	38%	59%	53%	41%
GEA										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	73%	45%	36%	**73%**	27%	36%	**64%**	27%	27%	45%
Incorrect	27%	55%	64%	27%	73%	64%	36%	73%	73%	55%
LINE										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	38%	25%	50%	**63%**	50%	31%	**69%**	13%	38%	81%
Incorrect	63%	75%	50%	38%	50%	69%	31%	88%	63%	19%
LAFI										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	38%	38%	**75%**	**88%**	50%	38%	50%	75%	50%	**100%**
Incorrect	63%	63%	25%	13%	50%	63%	50%	25%	50%	0%
LINI										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	**71%**	**68%**	**93%**	**75%**	**64%**	43%	**93%**	**68%**	**68%**	**82%**
Incorrect	29%	32%	7%	25%	36%	57%	7%	32%	32%	18%
LAGP										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	**86%**	**71%**	**71%**	**71%**	**71%**	43%	**71%**	57%	**71%**	**71%**
Incorrect	14%	29%	29%	29%	29%	57%	29%	43%	29%	29%
LIRH										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	25%	0%	**63%**	**63%**	50%	**63%**	**63%**	50%	38%	75%
Incorrect	75%	100%	38%	38%	50%	38%	38%	50%	63%	25%

LCOP											LRPC										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	**56%**	34%	**53%**	**69%**	**53%**	41%	**78%**	38%	**63%**	**66%**	Correct	**70%**	**55%**	**85%**	**75%**	**65%**	**60%**	**85%**	**65%**	**75%**	**75%**
Incorrect	44%	66%	47%	31%	47%	59%	22%	63%	38%	34%	Incorrect	30%	45%	15%	25%	35%	40%	15%	35%	25%	25%
LECO											LTIN										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	**67%**	**67%**	**73%**	**73%**	**67%**	**53%**	**100%**	**60%**	**73%**	**73%**	Correct	**100%**	**100%**	**67%**	**100%**	0%	33%	**100%**	33%	0%	33%
Incorrect	33%	33%	27%	27%	33%	47%	0%	40%	27%	27%	Incorrect	0%	0%	33%	0%	100%	67%	0%	67%	100%	67%
LIAD											TURI										
	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10		Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
Correct	**65%**	44%	**74%**	**97%**	**59%**	**62%**	**76%**	**71%**	32%	**74%**	Correct	**53%**	40%	**60%**	**73%**	**67%**	20%	**87%**	47%	**53%**	40%
Incorrect	35%	56%	26%	3%	41%	38%	24%	29%	68%	26%	Incorrect	47%	60%	40%	27%	33%	80%	13%	53%	47%	60%

Note: The percentages in bold correspond to those items that received 50% or more correct answers.
Source: Prepared by the authors based on the sample

Table 3. Homogeneity tests for the means for the degree programs

Wilks' lambda	0.8442	13.0 237.0	3.37	0.001
Pillai's trace	0.1558	13.0 237.0	3.37	0.001
Lawley-Hotelling trace	0.1846	13.0 237.0	3.37	0.001
Roy's largest root	0.1846	13.0 237.0	3.37	0.001

Source: Prepared by the authors based on the sample in Stata.

This finding may influence the analysis of the items that comprise the test, due to the presence of heterogeneous course sections, which could be a result of the different numbers of students in the sample per degree program and may influence the efficiency of the item. Sex is another variable that could support this hypothesis. The same tests were applied, with the results indicating that the variable of sex is homogeneous, while, according to the *p*-values obtained for the four tests, the hypothesis with identical means cannot be rejected (see Table 4).

Table 4. Homogeneity tests conducted on the means by sex

	Statistic	F(df1,	df2)	F	*p*-value
Wilks' lambda	0.9983	1	249	0.44	0.5096
Pillai's trace	0.0017	1	249	0.44	0.5096
Lawley-Hotelling trace	0.0018	1	249	0.44	0.5096
Roy's largest root	0.0018	1	249	0.44	0.5096

Source: Prepared by the authors based on the sample in Stata.

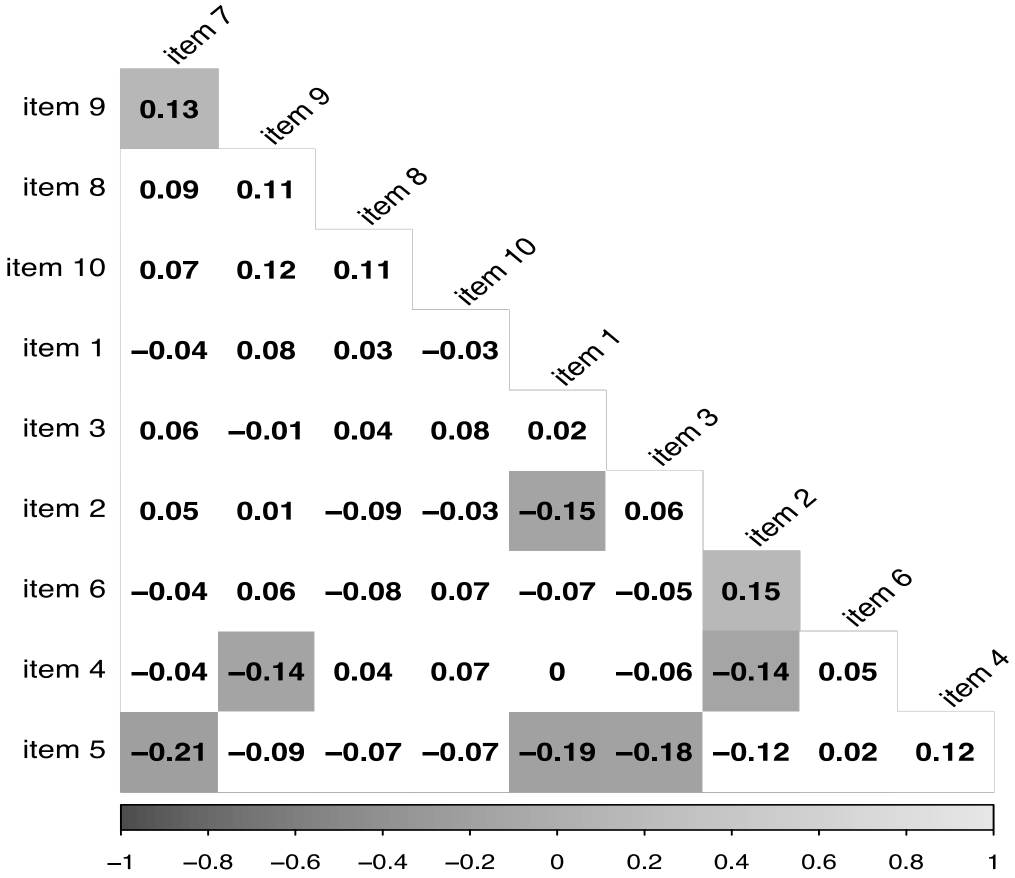
An analysis conducted per item from the test is presented in Table 5, with the aim of determining both how the students interpret each of the items and, in descriptive terms, their deficiencies. Table 5 presents a description of the objective of the item and each of the distractors. The first column is the item number, while the second is the learning objective the test is seeking to identify in the student, and columns three to six describe the purpose of each of the distractors, as well as the correct response. Each response option was identified by capital letters, from A to D.

Table 5. Objective of the items and distractors

No.	Objective	A	B	C	D
1	Ability to identify a combination-type arrangement	Confuse the arrangement with a permutation	Confuse the arrangement with a factorial	**Correct Response**	Apply the multiplier effect
2	Identification and resolution of a permutation problem	**Correct Response**	Carry out the calculation of multiplication rules	Apply the multiplier effect	Apply combinations instead of a permutation
3	Identify classical probability from the events of interest and the total events from the sample space	**Correct Response**	Confusion between events of interest and sample space	Is not aware of events of interest and sample space	Confuses the number of events of interest
4	Ability to determine the calculation of a simple probability	Calculate the probability that the suit of the card is clubs	Calculate the probability that the card is an ace	Add the club cards to the aces, repeating the ace of clubs	**Correct Response**
5	Ability to determine the complementary probability calculation and additive principle	Calculate the probability that the sum is two or twelve	Count the number of cases where the sum is two or twelve	Count the number of cases where the sum is neither two nor twelve	**Correct Response**
6	Identify and resolve a conditional probability problem	Approach the problem backwards	Confuse the formula for independent events	Does not understand the context (multiplying instead of dividing)	**Correct Response**
7	Identify and resolver an independent events problem	**Correct Response**	Marginal probability for Event A	Joint probability for events A and B	Cannot identify independent events
8	Identify the use of conditional probabilities in statements in which Bayes' theorem can be applied	Confuse it with the other event	**Correct Response**	Error in the calculation of the event	Error in both the calculation of the event and the correct event
9	Identify the data and know how to resolve a problem for a binomial distribution	Is confused, only setting for x=2	**Correct Response**	Is confused and makes the calculation for x>=2	Incorrectly makes the calculation when configuring and executes x<2
10	Identify the data and know how to resolve a problem for a Poisson distribution	Confuses the median of events with the random variable	**Correct Response**	Does not understand the context	On executing the calculation, makes an error and confuses the placing of the digits

Source: Prepared by the author.

As shown above, the means for the different undergraduate degrees are distinct, or, in other words, heterogeneous. The opposite is the case with regard to the variable of sex, where the means are equal, or, in other words, homogeneous. The correlations among the items were determined for the entire sample, with Figure 4 presenting the correlation matrix, using a 5 % significance level.



	item 7	item 9	item 8	item 10	item 1	item 3	item 2	item 6	item 4
item 9	0.13								
item 8	0.09	0.11							
item 10	0.07	0.12	0.11						
item 1	−0.04	0.08	0.03	−0.03					
item 3	0.06	−0.01	0.04	0.08	0.02				
item 2	0.05	0.01	−0.09	−0.03	−0.15	0.06			
item 6	−0.04	0.06	−0.08	0.07	−0.07	−0.05	0.15		
item 4	−0.04	−0.14	0.04	0.07	0	−0.06	−0.14	0.05	
item 5	−0.21	−0.09	−0.07	−0.07	−0.19	−0.18	−0.12	0.02	0.12

Fig. 4. Correlation Matrix
Note: The shaded part indicates that the correlation is statistically significant at 5 %
Source: Prepared by the author in R

The results presented in Figure 4 are observed to show a significant positive correlation in two pairs of items, the item pairs 6-2 and 7-9, with a correlation of 0.15 and 0.13, respectively. However, the two pairs of items correspond to distinct topics. According to Table 5, Item 2 is a permutation exercise and Item 6 is a conditional probability exercise, while Item 7 corresponds to independent events and Item 9 to binomial distribution. On the other hand, seven pairs of items are found with a significant negative correlation: 1-2; 1-5; 2-4; 3-5; 4-9; and, 5-7. Item pair 1-2 has a topic in common, namely combinations and permutations, which may not adequately identify the type of calculation that should be applied in both items. Although they are significantly correlated, pairs 1-5, 2-4, 3-5, 4-9 and 5-7 are not very clear in conceptual terms. The students, therefore, had problems identifying simple probability, combinations, permutations, conditional probability and binomial distribution calculations. The remaining pairs are not significantly correlated. It is necessary to analyze these results item by item in order to better establish the students' behavior.

Figure 5 presents the percentage of responses per option in each item, with dark grey indicating the correct option. Eight of the ten items had more than 50 % correct responses, with items 2 and 6 below the average. Item 7 obtained the highest percentage of correct responses in the sample, almost 80 %, which shows that the students had a good understanding of the concept of independent events. Moreover, it was found that, for seven items, between 1 % and 4 % the students did not answer the question.

With regard to Item 1, 57 % of the students had the ability to identify a combination problem, which 19 % incorrectly recognized as a problem to be resolved by means of the multiplier effect and which the remaining students confused with a factorial arrangement or permutation. Item 2 obtained 43 % correct responses, which shows that less than half the students had the ability to identify a permutation problem, while 36 % confused it with a combination problem. The incorrect global percentages for these two pairs of items are close to the correct percentages, which corroborates the negative correlation presented in Figure 4. However, according to the results for Item 3, 71 % of the students understood the concept of probability and identified the concept of sample space, which is corroborated by the 76 % correct responses for Item 4, which is an item related to classical, or simple, probability.

With regard to Item 5, 19 % of the students had problems representing both the additive principle and complementary probability, while 57 % knew how to respond correctly. However, 1 % did not answer the problem and, in fact, 12 % of the total number of students left at least one item unanswered from Item 4 onwards. Item 6, which is related to conditional probability, is one of the items with less than 50 % correct answers. Distractor B in this item was configured in such a way that the result arrived at by the student confuses the formula for independent events, and which 23 % of the students configured in this way.

However, for Item 7, which corresponds to independent events, nearly 80 % responded correctly, with only 5 % unable to identify the type of problem established. In addition, Item 8 established a problem to be resolved by means of Bayes' theorem, with the percentage of correct responses falling to 51 %, which was contrary to what was expected, due to the fact that, were Item 7 answered correctly, the student would have had to correlate the result to that obtained for Item 8, which did not happen. From these results, it is possible to infer that the students had different abilities in relating these concepts of probability to each other.

A binomial distribution problem was set in Item 9, which 53 % of the students answered correctly, while 21 % responded to the problem by attempting to find the correct value via probability mass function (PMF) and 19 % via accumulative probability. Therefore, 40 % had difficulty with binomial distribution problems. Strictly speaking, the basis of binomial distribution is the combinations. A very similar percentage is obtained to that for Item 1, which set a combination problem; however, a correlation was not found for this pair of items. However, for Item 10, which is a Poisson distribution problem that, like, Item 9, corresponds to discrete distribution, 70 % of the students responded correctly. One possible explanation for the variation of correct responses between questions 9 and 10 is that the students had greater difficulties calculating accumulated probabilities than PMF, as the former involve more operations.

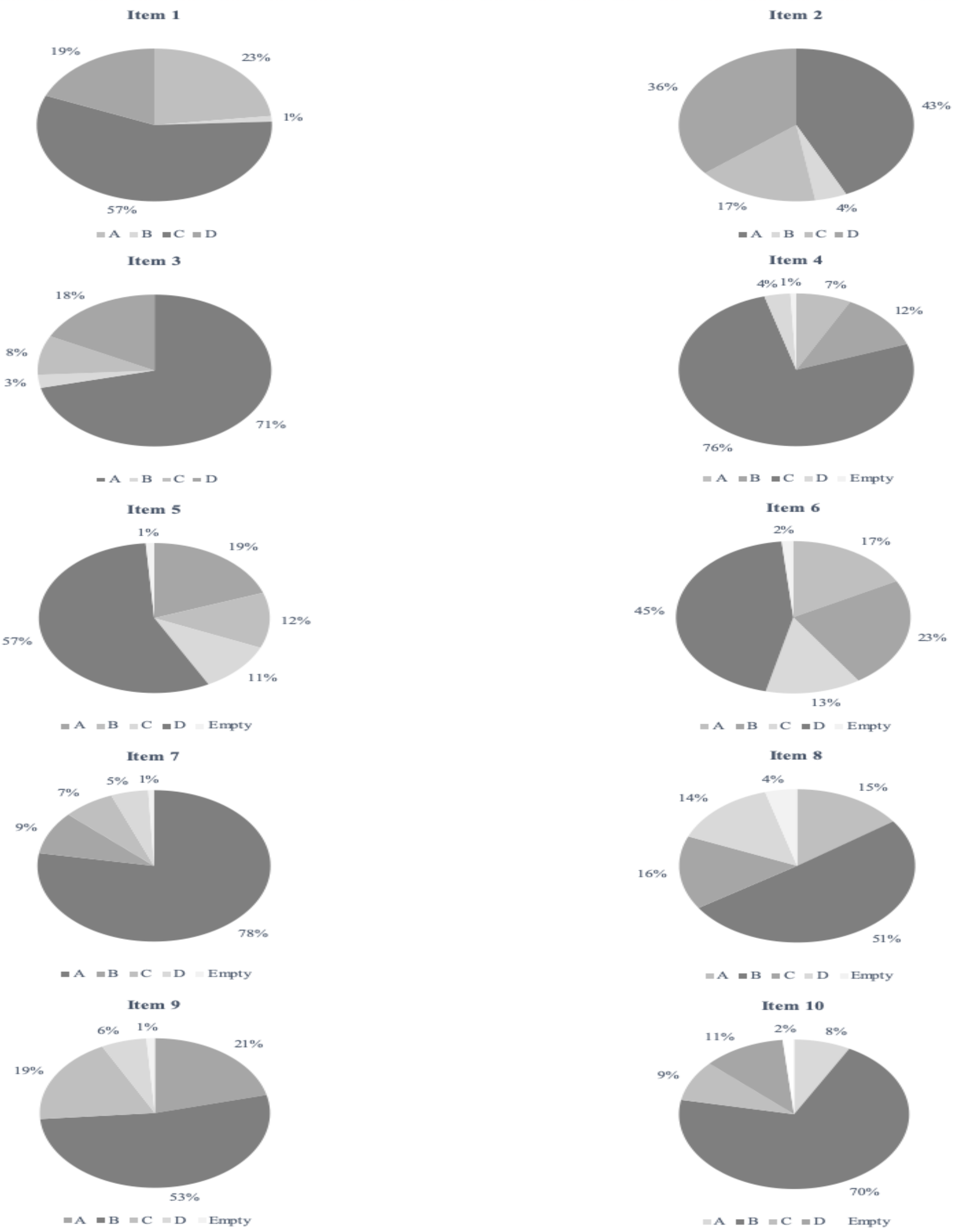


Figure 5. Percentage of responses per item and option
Resources: Prepared by the author

The above results show that, on average, 60 % of the students answered the items on the exam correctly. To corroborate these results, an RM was applied (Rasch, 1980), with binary (one and zero) coding applied, wherein a correct response is the equivalent to 1 and an incorrect response is the equivalent to 0. Applied to J students for a test consisting of I items, x_{ij} can be defined as the score obtained by the jth student on the ith item, which can be established as a logistic model with one parameter (1PLM) (Rasch, 1980; Sinharay, 2003; Thissen, Wainer, 2001):

$$P\left(x_{ij}=1\middle|\theta_j,a_i,b_i\right)=\frac{e^{a_i\left(\theta_j-b_i\right)}}{1+e^{a_i\left(\theta_j-b_i\right)}}, \qquad 1)$$

where $P(x_{ij} = 1|\theta_j, a_i, b_i)$ is the probability of student j with a score of 1 versus 0 in item i, a_i is the slope of the curve of the model, b_i is the difficulty of the item, and θ_j is the parameter of ability for student j. The calculation was carried out via Marginal Maximum Likelihood (MML) under the supposition of normal standard distribution, using the eRm package (Mair et al., 2019) included in the R software.

Prior to modeling the data, the RM was compared to the two and three-parameter logistic models, although the latter parameter model was not ultimately considered due to the fact that its Hessian matrix did not converge on a stable solution. Thus, the RM was compared to the two-parameter logistic model (LM2P) via the Bayesian criteria (BIC), obtaining results of 3129.79 and 3157.48, respectively. As the two values are very similar, the correlation between the models was calculated in order to ascertain whether some difference existed. The correlation of 0.97 indicates no difference between the models; therefore, the better model was chosen according to the BIC. Once the best model – an RM – had been chosen, the statistics of fit were determined and then used to ascertain whether the data covered the RM requirements.

There are two statistics, known in the literature as outfit and infit (Boone, Noltemeyer, 2017; Luo et al., 2009), both of which compare observed and expected values (Wright, Masters, 1982). Outfit is based on the unexpected responses positioned some distance from the measure of difficulty for the item; however, this statistic has the power to eliminate an item solely due to some unexpected responses from subjects for whom the item may be very difficult or vice versa (Bond, Fox, 2015). Infit corresponds to those subjects who did not respond in the expected manner to those items whose difficulty level corresponds to their ability levels (Susac et al., 2018). Therefore, the results will be based on the infit statistics from our analysis.

These statistics should be standardized via the standard distribution Z or the Student's t-test, for which the residuals should be obtained and used to calculate the mean squared error. See (Bond, Fox, 2015) for more information on this calculation. Table 6 presents the results for the parameter b_i, the standardized outfit and infit statistics, as well as their respective statistics and p values.

Table 6. Difficulty parameters and outfit and infit statistics

Item	b_i	Outfit	Outfit_t	Outfit_p	Infit	Infit_t	Infit_p
7	-1.479	0.884	-1.408	0.159	0.981	-0.204	0.838
4	-1.376	1.042	0.517	0.605	1.031	0.395	0.693
3	-1.089	1.020	0.283	0.778	1.032	0.480	0.631
10	-1.022	0.903	-1.517	0.129	0.967	-0.495	0.621
1	-0.345	1.025	0.488	0.626	1.043	0.839	0.401
5	-0.325	0.910	-0.183	0.608	0.950	-0.997	0.319
9	-0.133	0.946	-1.116	0.264	0.972	-0.579	0.562
8	-0.037	0.961	-0.815	0.415	0.987	-0.262	0.794
6	0.251	1.173	3.208	0.001	1.113	2.461	0.014
2	0.328	0.963	-0.726	0.486	0.968	-0.611	0.541

Source: Prepared by the authors own elaboration based on R results
Note: The calculations were undertaken considering a 5 % significance level.

The b_i results were ordered from the easiest item (Item 7) to the most difficult (Item 2) and can be found in Column 2 in Table 6. This index may take either positive or negative values, with values close to zero expressing average difficulty, negative values expressing below average difficulty (low difficulty), and positive values expressing above average difficulty (high difficulty). Thus, there are two items close to zero, items 6 and 2, which are of medium difficulty, coinciding with the percentage of questions answered correctly but below the average, as shown in Figure 5. Finally, Item 7 is the easiest, which coincides with the highest percentage of correct answers of all the items in Figure 5.

Item 6 in Table 6 does not seem to offer a good fit, due to the fact that the interval for standardized outfit and infit must be between -1.96 and 1.96 in terms of the *t* value, to a 95 % confidence level. However, rather than analyzing the probability exam itself, the present study sought to analyze the students' ability to understand different topics relating to probability, for which reason, Item 6 was not eliminated from the RM in order to then adjust the model. Figure 6 represents the behavior of the items according to infit_t.

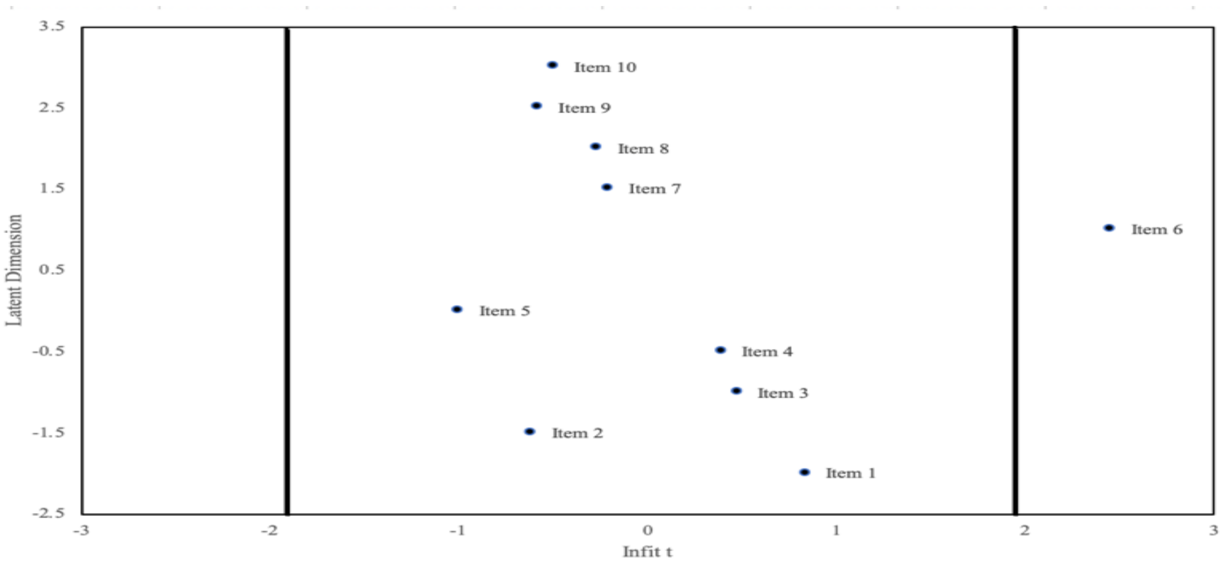


Fig. 6. Confidence interval for standardized infit statistics
Source: Prepared by the authors based on R results

Finally, Figure 7 presents an illustration of the characteristic curves for each of the items (ICC). Axis *x* represents the students' ability, which is found to be between -3 and 3, while Axis *y* represents the probability of the student answering the item correctly (score). For example, Item 7 was considered the easiest, where a medium level-ability student has an approximately 80 % probability of correctly answering the item, while a student with a very low level of ability, for example -3, would have a 20 % probability of answering correctly. Analyzing Item 2, which was the most difficult according to parameter *b*, a medium level-ability student has an approximate probability of 40 % of answering correctly. On characteristic curves, these behaviors are similar to the percentage graphs shown in Figure 5.

Finally, the total information curve is calculated, as presented in Figure 8. The test used in the present research is calculated at an interval of (-10, 10) (Rizopoulos, 2017), while applying the test at an interval of (-4,0.5) obtained a total information result of 67.63 %, which indicates that there was more information for students with low ability levels than for students with a high ability. From this result, it can be concluded that the test applied is a test for students with a low level of ability; therefore, there are various factors that should be analyzed in the future, such as the students' performance on the degree program and the type of teaching they have been receiving.

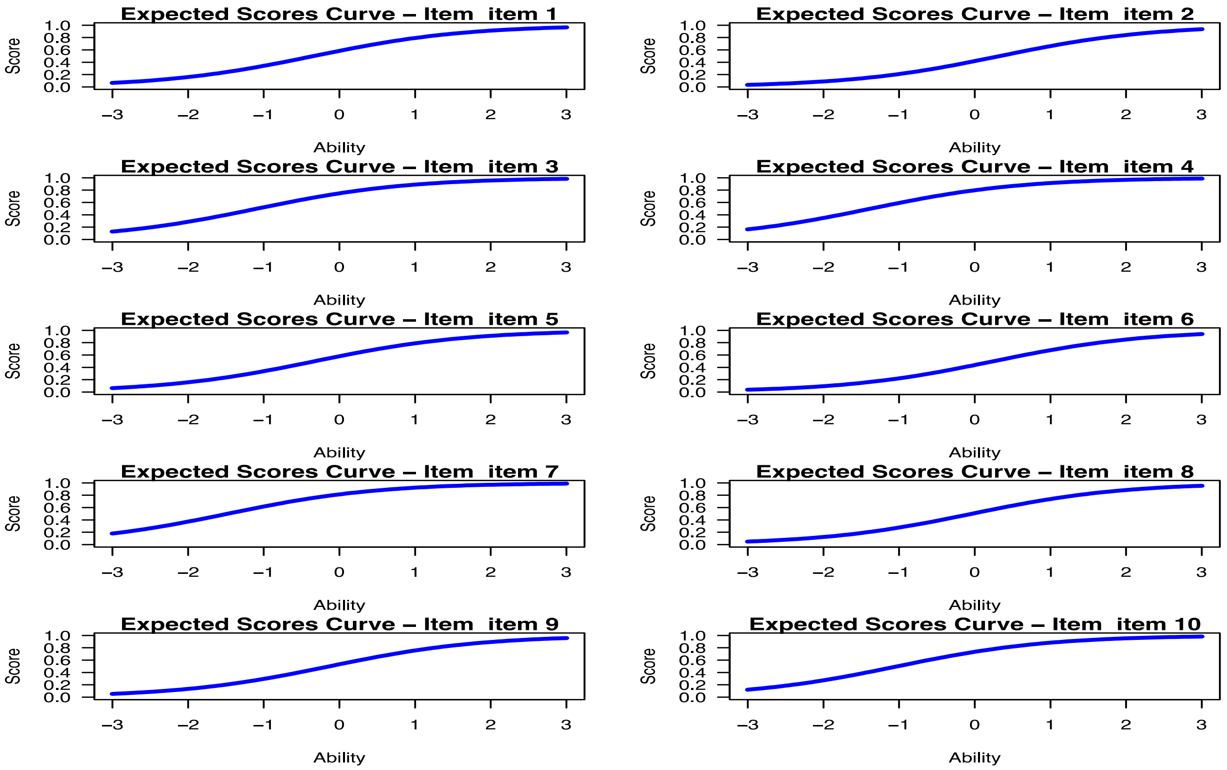


Fig. 7. ICCs for each of the items
Source: Prepared by the authors based on R results

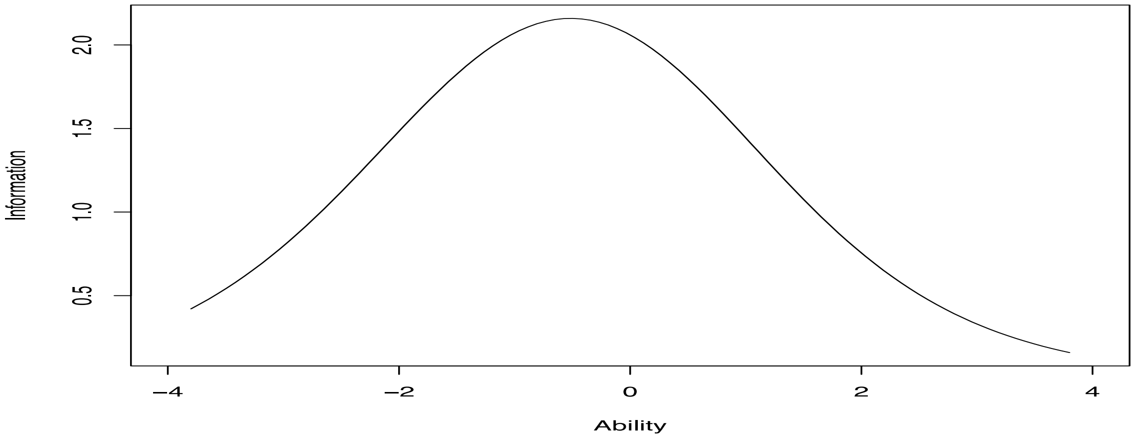


Fig. 8. Test Information Function
Source: Prepared by the authors based on R results

4. Conclusion

This article presents the design of a 10-question exam for measuring students' ability in topics of probability and which was aligned to CUCEA's learning objectives. The evidence obtained shows that CUCEA students have more difficulty – low grades – in the area of probability than in the area of descriptive statistics, both in the DEs and the annual statistics tournaments.

The results indicate, descriptively, that an average of more than 60 % of the students answered the items on the test correctly. Two questions were answered incorrectly by seven and five percentage points below the average, while another two questions were answered correctly, above the average by close to one and three percentage points, respectively. Homogeneity tests were undertaken between the means, showing that they were homogeneous by sex, but heterogeneous by undergraduate degree program.

Subsequently, the objectives of the items were analyzed along with their respective correct responses and distractors. While both negative and positive correlations between items were obtained, those pairs that were positively correlated did not have any qualitative relationship, as they corresponded to distinct topics. A correlation was found between items 6 and 2, which are item pairs and, for which, 50 % correct answers or above were obtained.

These results enable the application of an RM to determine the students' ability, revealing two difficult items, items 2 and 6. The results for Item 6 are not statistically significant, while Item 2 presented a percentage of correct answers below the average, which is statistically significant.

Therefore, the descriptive results prove that the students did have the ability to correctly answer probability questions in the economic-administrative sciences, although they did face certain difficulties with conditional probability and permutation exercises. At the same time, the RM results confirm that the students faced difficulties in solving conditional probability problems, meaning that this test could be used as a base by the SA for the construction of a reliable bank of questions for the topic of probability that would help to improve both the DEs and the STs.

Moreover, analysis of the items on the probability exam enabled the detection of the students' difficulties and their most frequent errors, which can be used to assist professors in reinforcing learning in these topics. Moreover, it was noted that the items on the probability exam coincided with the study programs currently in force and were aligned with the results obtained for institutional learning.

5. Acknowledgments

We would like to thank Professor Robert delMas of The University of Minnesota for his comments on the paper.

References

Academia de Estadistica, 2018 – Academia de Estadistica (2018). Actas de academia del departamento de métodos cuantitativos del CUCEA de la Universidad de Guadalajara.

Aziz et al., 2013 – *Aziz, A., Salleh, T.S.A., Khatimin, N., Zaharim, A.* (2013). Evaluating multiple choice items in determining quality of test. TALE2013 – IEEE International Conference on Teaching, Assessment and Learning for Engineering: 565–569. DOI: https://doi.org/10.1109/TALE. 2013.6654501

Batanero et al., 2016 – *Batanero, C., Chernoff, E. J., Engel, J., Lee, H. S., Sánchez, E.* (2016). Research on teaching and learning probability (G. Kaiser, Ed.). DOI: https://doi.org/10.1007/978-3-319-31625-3_1

Batanero et al., 2005 – *Batanero, C., Henry, M., Parzysz, B.* (2005). The nature of chance and probability. In G. A. Jones (Ed.), Exploring Probability in School: Challenges for Teaching and Learning: 15-37. DOI: https://doi.org/10.1007/0-387-24530-8_2

Bond, Fox, 2015 – *Bond, T.G., Fox, C.M.B.* (2015). Applying the Rasch: fundamental measurement in the human sciences (3rd ed.). New York: Routledge.

Boone, Noltemeyer, 2017 – *Boone, W.J., Noltemeyer, A.* (2017). Rasch analysis: a primer for school psychology researchers and practitioners. *Cogent Education*, 4(1): 1-13. DOI: https://doi.org/10.1080/ 2331186X.2017.1416898

Borovcnik, 2011 – *Borovcnik, M.* (2011). Strengthening the role of probability within statistics curricula. In C. Batanero, G. Burrill, & C. Reading (Eds.), Teaching Statistics in School Mathematics-Challenges for Teaching and Teacher Education: A Joint ICMI/IASE Study, Vol. 14: 71-83. DOI: https://doi.org/10.1007/978-94-007-1131-0

Carolyn, Kirk, 2001 – *Carolyn, K., Kirk, S.* (2001). A new approach to learning probability in the first statistics course. *Journal of Statistics Education*, 9(3): 1-25. DOI: https://doi.org/10.1080/ 10691898.2001.11910539

Del Toro, Ochoa, 2010 – *Del Toro, H.L., Ochoa, M.B.* (2010). Exámenes departamentales en el CUCEA: el caso del departamento de métodos cuantitativos. XIV Congreso Internacional de Investigación en Ciencias Administrativas [Electronic resource]. URL: http://acacia.org.mx/busqueda /pdf/EXAMENES_DEPARTAMENTALES.pdf

Flores et al., 2014 – *Flores, B., García, J.I., Sánchez, E.* (2014). Avances en la calidad de las respuestas a preguntas de probabilidad después de una actividad de aprendizaje con tecnología. In M.T. González, M. Codes, D. Arnau, T. Ortega (Eds.), Investigación en Educación Matemática XVIII: 307-316. Salamanca: SEIEM.

Landin, Sánchez, 2010 – *Landin, P., Sánchez, E.* (2010). Niveles de razonamiento probabilístico de estudiantes de bachillerato frente a tareas de distribución binomial. *Educação Matematica Pesquisa*, 12(3): 598-618.

Lawley, 1938 – *Lawley, D.N.* (1938). A generalization of Fisher's z test. *Biometrika*, 30(1/2): 180-187. DOI: https://doi.org/10.2307/2332659

Luo et al., 2009 – *Luo, X., Cappelleri, J.C., Cella, D., Li, J.Z., Charbonneau, C., Kim, S.T., Motzer, R.J.* (2009). Using the rasch model to validate and enhance the interpretation of the functional assessment of cancer therapy-kidney symptom index - disease-related symptoms scale. *Value in Health*, 12(4): 580-586. DOI: https://doi.org/10.1111/j.1524-4733.2008.00473.x

Mair et al., 2019 – *Mair, P., Hatzinger, R., Maier, M. J., Rusch, T., Debelak, R.* (2019). Package 'eRm' [Electronic resource]. URL: https://cran.r-project.org/web/packages/eRm/eRm.pdf

Pillai, 1955 – *Pillai, K.C.S. (1955)*. Some new test criteria in multivariate analysis. *The Annals of Mathematical Statistics*, 26(1): 117-121. DOI: https://doi.org/10.1214/aoms/1177728599

Rasch, 1980 – *Rasch, G.* (1980). Probabilistic models for some intelligence and attainment tests. Chicago, US: The University of Chicago Press.

Raudzatul, 2016 – *Raudzatul, M.S.* (2016). Application of Rasch measurement model in evaluating student performance for foundation of computing II. In C.Y. Fook, S. Narasuman, S.B. Abdul Rahman, G. K. Sidhu, & L. L. Fong (Eds.), 7th International Conference on University Learning and Teaching (InCULT 2014) Proceedings. DOI: https://doi.org/10.1007/978-981-287-664-5

Rizopoulos, 2017 – *Rizopoulos, D.* (2017). Package 'ltm' [Electronic resource]. URL: https://github.com/drizopoulos/ltm

Roy, 1957 – *Roy, S.* (1957). Some aspects of multivariate analysis. New York: Wiley.

Sánchez, Inzunza, 2006 – *Sánchez, E., Inzunza, S.* (2006). Meanings' construction about sampling distributions in a dynamic statistics environment. Proceedings of the Seventh International Conference on Teaching Statistics: 1-6. Salvador (Bahia), Brazil.

Sinharay, 2003 – *Sinharay, S.* (2003). Bayesian item fit analysis for dichotomous item response theory models [Electronic resource]. URL: http://www.ets.org/Media/Research/pdf/RR-03-34.pdf

Susac et al., 2018 – *Susac, A., Planinic, M., Klemencic, D., Milin Sipus, Z.* (2018). Using the Rasch model to analyze the test of understanding of vectors. *Physical Review Physics Education Research*, 14(2): 1-6. DOI: https://doi.org/10.1103/PhysRevPhysEducRes.14.023101

Thissen, Wainer, 2001 – *Thissen, D., Wainer, H.* (2001). Test scoring. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Wilks, 1932 – *Wilks, S.S.* (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4): 471–494. DOI: https://doi.org/10.2307/2331979

Wright, Masters, 1982 – *Wright, B., Masters, G.* (1982). Rating scale analysis: Rasch measurement. (1st ed.). Chicago: Mesa Press.