AIMS Mathematics, 11(7): 20170–20194.
DOI:10.3934/math.2026819
Received: 18 May 2026
Revised: 22 June 2026
Accepted: 29 June 2026
Published: 09 July 2026

https://www.aimspress.com/journal/Math

*Research article*

# A conjugate bivariate model suitable for detecting heterogeneity and getting evaluation marks via multiple-choice tests

**Emilio Gómez-Déniz**[1], **Nancy Dávila-Cárdenes**[1] **and José María Pérez-Sánchez**[2,*]

[1] Department of Quantitative Methods in Economics and TiDES Institute, University of Las Palmas de Gran Canaria, Spain

[2] Department of Applied Economic Analysis and TiDES Institute, University of Las Palmas de Gran Canaria, Spain

* **Correspondence:** Email: josemaria.perez@ulpgc.es.

**Abstract:** It is well known that academic performance varies in heterogeneous and homogeneous groups. In the latter, all students possess approximately the same level of instruction and skills. The results are usually measured by the marks achieved on exams. In this work, the marks calculated through this methodology could be considered a Bayes rule within the framework of statistical decision theory. With this methodology, we estimated the marks obtained on a multiple-choice test, considering individual student performance with the accumulated experience of the group. As a result, we addressed two major objectives: First, a mechanism was developed to detect the heterogeneity or homogeneity of the group based on the accumulated experience on past exams; and second, an evaluation system was proposed that rewarded or penalized students depending on their academic performance. The proposed framework may have implications for supporting students at risk, although such effects were not directly evaluated in this study.

**Keywords:** Bayesian; classroom; conjugate distribution; heterogeneity; loss function; marks; multiple-choice test
**Mathematics Subject Classification:** 97Mxx, 97Rxx

## 1. Introduction

Increasing student heterogeneity has become a central issue in educational research and practice. Contemporary classrooms, particularly in higher education, are characterized by substantial diversity in cultural background, prior knowledge, learning styles, and socio-emotional conditions. While this diversity reflects broader societal dynamics, it also poses significant challenges for instructional design and student assessment. Furthermore, it creates opportunities to foster inclusive learning environments

and to enhance educational outcomes through the interaction of diverse perspectives.

A key dimension of heterogeneity arises from differences in cultural and linguistic backgrounds, which have intensified due to globalization and mobility. Although such diversity can enrich learning processes, it may also generate barriers to communication and engagement, requiring pedagogical strategies that ensure equity in access and participation. Beyond cultural differences, variability in academic skills and learning trajectories is pervasive. Students often display markedly different levels of prior knowledge and abilities, which calls for differentiated instruction, flexible grouping strategies, and adaptive evaluation systems. In addition, social and emotional factors, such as socioeconomic conditions, mental health, and family support, further contribute to disparities in student performance.

From an institutional perspective, heterogeneity is shaped by administrative mechanisms used to allocate students to groups. In higher education, classroom composition frequently depends on admission scores, enrollment timing, or student preferences, leading to groups that range from highly homogeneous to strongly heterogeneous. This variability is particularly evident across degree programs: While selective programs often generate relatively homogeneous groups, others exhibit broad distributions of prior academic achievement.

The implications of homogeneous versus heterogeneous grouping have been widely examined in the literature. Grouping students according to abilities and preferences enables tailored instructional approaches and smaller-scale learning environments [28–30]. Empirical evidence suggests that interaction and achievement are positively related, particularly in heterogeneous groups, where high- and low-ability students benefit from peer interaction [12]. Heterogeneity has been shown to enhance creativity and the quality of problem-solving, although it may also introduce conflicts stemming from cultural or cognitive differences [35].

At the classroom level, [22] the researchers in conceptualize educational productivity as a function of diversification, emphasizing the role of peer interactions and the alignment between student needs and instructional capacity. Given the increasing importance of teamwork in modern labor markets, collaborative learning has become a central component of educational practice. Studies such as [36] further highlight how heterogeneity interacts with teacher quality, technological tools, and socioeconomic factors in shaping academic outcomes.

Most of the empirical literature on grouping entails primary and secondary education [8, 15, 24, 27, 34], whereas evidence for higher education remains comparatively limited. In this context, grouping policies are often controversial. For example, the researchers in [3] discuss the implications of mixing high- and low-achieving students, while the researchers in [37] highlight the hidden heterogeneity associated with students' social backgrounds and educational strategies. Researchers have also identified multiple sources of heterogeneity, including inter-individual differences, intra-individual variability, and interactions between student characteristics and instructional features [17].

Against this background, Bayesian methods provide a natural and flexible framework for modeling heterogeneity in educational data. These approaches explicitly account for uncertainty and enable the incorporation of prior information, making them particularly well suited to settings with complex and imbalanced data structures. Early contributions emphasize the importance of informative priors in improving inference on student learning [16, 19], while subsequent studies entail Bayesian techniques to analyze variance heterogeneity and performance dispersion across schools [18]. Other works highlight their usefulness for forecasting student outcomes [11] and for modeling uncertainty in educational processes such as student feedback and roles [2]. Multilevel Bayesian approaches have

also been shown to capture a substantial proportion of variation in academic achievement [13], and have been applied to compare student performance and evaluate educational interventions [5, 23].

Developments in learning analytics and educational data science further reinforce the relevance of Bayesian hierarchical models. These models enable the simultaneous analysis of multiple sources of variability-such as individual, classroom, and institutional effects-while preserving interpretability and coherence. Contemporary studies demonstrate their effectiveness in modeling academic performance and retention in higher education settings, e.g., [25, 26], as well as in integrating learning analytics data to improve predictive accuracy and decision-making [5, 10]. In particular, hierarchical Bayesian approaches enable the joint modelling of correlated outcomes, the incorporation of latent factors, and the propagation of uncertainty across levels, which are essential features when dealing with heterogeneous student populations.

Despite these advances, there remains a need for methodological frameworks that explicitly combine individual-level experience with group-level information in the assessment of student performance. In particular, approaches often treat evaluation processes as either purely individual or purely aggregate, without fully exploiting the interaction between both sources of information.

This paper contributes to this literature by proposing a Bayesian decision-theoretic framework for assessing student performance in multiple-choice tests within heterogeneous university classrooms. Building on statistical decision theory [4, 21, 32], we derive scoring rules as Bayes estimators under suitable loss functions. The proposed approach integrates individual experience (based on past responses) with collective information (derived from the student group), thereby providing a coherent mechanism to account for heterogeneity in both dimensions.

Empirically, we apply the proposed methodology to data from students enrolled in a Mathematics course in a Business Administration degree at a Spanish university. The results indicate that, while differences with respect to classical evaluation methods are moderate, the Bayesian approach provides a principled and internally coherent framework for assessment. More broadly, the methodology offers a flexible tool for modelling heterogeneous educational environments and for improving the consistency and informational content of evaluation practices in higher education.

The remainder of the paper is organized as follows: In Section 2, we introduce the statistical decision framework underlying the proposed model. In Section 3, we present the probabilistic assumptions for the response process and develop the Bayesian formulation. In Section 4, we provide the corresponding scoring rules and discuss the role of heterogeneity and its measurement. In Section 5, we provide an empirical applicationand conclude the study in Section 6.

## 2. The statistical decision framework

It is assumed that the mark achieved by a student on an essay exam or a multiple-choice test (the state considered here) is determined by the combined effect of two stochastic processes: Whether or not to answer a test question, and, if answered, whether the answer is correct or incorrect. The student's performance can be due to many causes. However, fundamentally, it stems from the student's skills in the subject of the exam, which surely reflect the student's provenance and prior knowledge in the subject we are considering. These features are assumed to vary randomly among students, reflecting some heterogeneity within the classroom. In summary, there could be enough heterogeneity in the class related to the students' provenance and their different skills in the subject we are considering.

For this reason, we will develop a tool to describe heterogeneous groups (collectives) and to address how individual and collective experience should be combined in a school or faculty to estimate the final mark in a subject. We will develop an experience-based system for assigning a score to each student based on their experience and the prior information about the collective of all students. Table 1 illustrates the comments above with the observed test outcomes, given by $x_{ij}$, for a school or faculty with $k$ students and $t$ test questions. Thus, we are going to assume that the number of total answers and the corresponding number of correct answers of one student for one exam consisting of a multiple-choice test is specified by two random variables, say $X$ and $Z$, respectively, following a joint probability function $f(x, z|\Theta)$ depending on an unknown (possibly vector) risk parameter $\Theta$. In Table 1, $\Theta_i$ corresponds to the particular realization of the random variable $\Theta$. Observe that each student has a particular value of this random variable.

**Table 1.** Illustration of the heterogeneity of students in the classroom and the results in question tests (exams).

| Exams \ Students | 1 | 2 | ... | $k$ |
|---|---|---|---|---|
| 1 | $x_{11}$ | $x_{12}$ | ... | $x_{1k}$ |
| 2 | $x_{21}$ | $x_{22}$ | ... | $x_{2k}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $t$ | $x_{t1}$ | $x_{t2}$ | ... | $x_{tk}$ |
| | $\Theta_1$ | $\Theta_2$ | ... | $\Theta_k$ |

Let $(\tilde{x}, \tilde{z}) = (X_i, Z_i)$, $i = 1, 2, \ldots, t$, be independent and identically distributed random variables representing the number of responses and correct responses during the last $t$ periods (years, months, academic years, ...). We can assign to each risk parameter $\Theta$ a fair mark $\mathcal{M}(\Theta)$ within the set of admissible marks. The mechanism for assigning a mark to a student who has answered a question on the test can be viewed as a procedure for obtaining an estimate of an unknown risk parameter $\Theta$, from observations $(\tilde{x}, \tilde{z})$ obtained according to a likelihood $f((\tilde{x}, \tilde{z})|\Theta)$, in the presence of a prior distribution $\pi(\Theta)$ for a given loss function $L(\Theta, \widehat{\Theta})$. Here, $\widehat{\Theta}$ is an estimator of $\Theta$.

The estimated mark can be obtained by minimizing, with respect to $\mathcal{M}$, the expected loss $E_f\left[L(g(x, z), \mathcal{M})\right]$, where $L$ is taken as the squared-error loss function $L(m, n) = (m - n)^2$. In this regard, decision-theoretic statisticians and economists have traditionally used the squared-error loss function for many years. Therefore, we choose $\mathcal{M}$ according to

$$\arg\min_{\mathcal{M}} \sum_{x,z} [g(x, z) - \mathcal{M}]^2 f(x, z|\Theta). \tag{2.1}$$

In this setting, $g(x, z)$ is an appropriate function of the number of responses and the correct answers made by the student, which a priori provides the mark obtained by the student.

In the case of using the squared-error loss function, we have, from (2.1), that $\mathcal{M} : \Theta \longrightarrow \mathbb{R}$ is given by

$$\mathcal{M}(\Theta) = E_f\left[g(x, z)|\Theta\right], \tag{2.2}$$

and will be called the risk mark, which is a random variable.

The mark computed above can be applied only practice if the probability function of $(X, Z)$ is wholly known. For this reason, expression (2.2) is referred to as the risk mark. Assume now that this probability function is specified up to an unknown vector of parameters. In this case, we can proceed as follows.

First, if no experience is available, we can compute the mark by minimizing the risk function, i.e., minimizing $E_\pi [L(\mathcal{M}(\Theta), \mathcal{M})]$, where $\pi(\Theta)$ is the prior distribution on the unknown parameter $\Theta$. Thus, we consider

$$\arg\min_{\mathcal{M}} \int_\Theta [\mathcal{M}(\Theta) - \mathcal{M}]^2 \pi(\Theta)\, d\theta.$$

In this case, the same procedure as above is carried out for the prior mark given by

$$\mathcal{M} = E_\pi [\mathcal{M}(\Theta)], \tag{2.3}$$

where the expectation is taken with respect to the prior distribution $\pi$.

On the other hand, if experience is available, we can take a sample $(\tilde{x}, \tilde{z})$ from the random variables $(X, Z)$, assuming $(X_i, Z_i)$ are independent and identically distributed, and then use this information to estimate the unknown mark $\mathcal{M}(\Theta)$ through the Bayes mark $\mathcal{M}^*$ obtained by minimizing the Bayes risk, i.e., minimizing $E_{\pi^*} [L(\mathcal{M}(\Theta), \mathcal{M}^*)]$. Here, $\pi^*$ is the posterior distribution of the risk parameter $\Theta$ given the sample information $(\tilde{x}, \tilde{z})$. Thus, the Bayesian mark is computed as in (2.3) by replacing $\pi$ with $\pi^*$. Later, $\pi$ will be chosen to be conjugate with respect to the likelihood $f$, in which case the computation of $\mathcal{M}^*$ from $\mathcal{M}$ will be immediate.

It should be noted that, *a priori*, all students are assumed to possess equivalent abilities; however, it is recognized that, in practice, some students perform better than others within the classroom setting. Therefore, it is neither feasible nor appropriate to determine *a priori* which students will demonstrate superior abilities and which will not. Rather, only *a posteriori*, following careful observation of individual student performance, can such conclusions be drawn. This observation formalizes our intuition of the cohort as a collective composed of individuals with heterogeneous skill sets yet comparable risk profiles.

Then, for estimating the risk mark $\mathcal{M}(\Theta)$, it is reasonable to consider the two sources of information available: prior knowledge and individual observations. The prior knowledge contains information about all the students in the classroom, and the best estimator that we can derive from this information is the prior mark $\mathcal{M}$. On the other hand, the observations contain information about the individual student. It is reasonable to consider linear estimators that are individual, i.e., conditionally unbiased, and to choose from these the one with the greatest precision, i.e., the smallest variance. Finally, we have to weigh the estimators derived from each information source. It is intuitively reasonable to weigh them according to their precision, that is, according to the inverse of the quadratic loss. Thus, we suggest a mark that satisfies the expression given in the following definition.

**Definition 1.** *The correct (fair) individual mark of a student in a class with risk profile* $\Theta$ *is*

$$\mathcal{M}_i^* = \gamma h(\tilde{x}_i, \tilde{z}_i) + (1 - \gamma)\mathcal{M}, \quad i = 1, 2, \ldots, k, \tag{2.4}$$

*where* $h(\tilde{x}, \tilde{z})$ *is a function of the sample,* $\gamma \in [0, 1]$*, and* $\mathcal{M}$ *is given in* (2.3)*.*

The expression given in (2.4) is similar to the one suggested by actuaries to compute premiums in a portfolio of insurance contracts in the setting known as credibility theory (see, for instance, [6, 9, 14]). In the following, we will consider distributions for $X$, $Z$, and $\Theta$, leading to individual marks of the form given in (2.4).

Then, to estimate theoretical mark $\mathcal{M}(\Theta)$, it is reasonable to consider the two sources of information available: prior knowledge and individual observations. The *a priori* knowledge contains information about all the students in the classroom, and the best estimator that we can derive from this information is prior mark, $\mathcal{M}$. On the other hand, the observations contain information about the individual student. It is reasonable to consider linear estimators that are individual, i.e., conditionally unbiased, and to choose from these the one with the greatest precision, i.e., the smallest variance. Finally, we must weigh the estimators derived from each information source. It is intuitively reasonable to weigh them according to their precision, that is, according to the inverse of the quadratic loss.

## 3. Basic distributional assumptions

This work is developed in the context of multiple-choice tests in which the student faces a fixed number of questions with three answer options, only one of which is correct. A correct answer earns a certain number of points, while an incorrect answer results in a penalty. This type of test is common, at least in the Spanish university system, when professors wish to administer partial tests before the final exam. These tests serve two major purposes: On the one hand, they can measure students' understanding of the material, providing instructors with information about their level of proficiency. On the other hand, they are easy to create and grade, especially when the number of students is relatively large. This assessment system is not only used in university settings but has also been implemented in other contexts, such as the Spanish healthcare system for selecting future resident physicians. Given the high number of applicants for positions within the national health system, the authorities presumably opt to simplify the grading process and enable evaluation across all topics within the extensive syllabus.

Let us consider a scenario in which the student is presented with a multiple-choice questionnaire consisting of several answer options or a true-or-false format containing questions that the student may or may not be able to answer. Most multiple-choice tests award points for correct answers and deduct points for incorrect ones; therefore, students can skip some questions. This format will be used here. Suppose that the student's selection for each question follows a Bernoulli distribution with probability of choosing to answer equal to $p_1$. If the questionnaire consists of $n$ questions, and we assume independence between questions, the likelihood of responding exactly $x \leq n$ items in $n$ questions follows a binomial distribution with probability function given by

$$f(x|p_1) = \binom{n}{x} p_1^x (1-p_1)^{n-x}, \quad x = 0, 1, \ldots n. \tag{3.1}$$

On the one hand, when a student responds to a question, this response can be correct or incorrect. We will introduce a second random variable that enables us to separate this process into two sub-events (responding correctly and responding incorrectly) as follows: Let $Z_i$ be a Bernoulli variable taking the following values,

$$Z_i = \begin{cases} 1, & \text{if the } i\text{th response is correct,} \\ 0, & \text{otherwise.} \end{cases}$$

On the other hand, let $Z = \sum_{i=1}^{x} Z_i$ be the total number of correct questions answered. Thus, if the $Z_i$ $(i = 1, \ldots, x)$ are assumed to be mutually independent, then the conditional probability function of $Z$ given that $X = x$ is binomial with parameters $x$ and $p_2$. That is,

$$f(z|x, p_2) = \binom{x}{z} p_2^z (1 - p_2)^{x-z}, \quad z = 0, 1, \ldots, x. \tag{3.2}$$

It is more realistic to assume that students first attempt to solve the question and, if they are unable to do so, decide whether to guess or skip it. In practice, the instructor does not observe or control this process; therefore, the proposed model assumes that students first decide whether to answer and, conditional on answering, their responses may be correct or incorrect. Experience also suggests that some topics may be inaccessible to certain students, leading them to skip the corresponding questions.

Thus, we obtain that the conditional mean is given by $E(Z|X = x) = xp_2$, which increases with $x$ and $p_2$. Furthermore, the conditional variance is $Var(Z|X = x) = xp_2(1 - p_2)$. Thus, the regression of $Z$ on $X$ and the conditional variance of $Z$ given $X$ are linear.

Therefore, the joint distribution of the number of total answers $X$ and the corresponding number of correct answers $Z$ has a probability function

$$f(x, z|p_1, p_2) = f(z|x, p_2) f(x|p_1) = \frac{n!(1 - p_1)^n}{(x - z)!(n - x)!z!} \left(\frac{p_2}{1 - p_2}\right)^z \left(\frac{1 - p_2}{1 - p_1} p_1\right)^x, \tag{3.3}$$

for $x = 0, 1, \ldots, n$, $z = 0, 1, \ldots, x$. This distribution corresponds to a hierarchical compound distribution obtained by combining two binomial processes: One governing the number of attempted questions and another governing the number of correct answers conditional on the number of attempts.

Some computations provide the cross moment, which is

$$E(XZ|p_1, p_2) = np_1p_2 \left[1 + (n - 1)p_1\right].$$

Since the marginal moments are $E(X|p_1) = np_1$ and $E(Z|p_1, p_2) = np_1p_2$, we get the covariance given by $cov(X, Z|p_1, p_2) = np_1p_2(1 - p_1)$, which is always positive, and the correlation coefficient is

$$\rho(X, Z|p_1, p_2) = \sqrt{\frac{(1 - p_1)p_2}{1 - p_1p_2}}.$$

### *3.1. The Bayesian model*

Implicitly, (3.1) assumes that all students have the same probability of responding to a question, and (3.2) that all the students who have answered $x$ questions have the same probability of obtaining a correct answer. Note that $p_1$ describes the probability for a student to answer a question in a test with $n$ questions, and $p_2$ describes the probability of responding correctly to a question when the student has answered $x$ questions among the $n$ questions of the test. A more general model enables parameters $p_1$ and $p_2$ in (3.1) and (3.2) to vary among students. We often do not know a priori the values of $p_i$, $i = 1, 2$. The most intuitive choice to reflect such lack of knowledge is to assign a uniform

prior on $p_i$, i.e., $\pi(p_i) = 1$ for $0 < p_i < 1$; this means that *a priori*, $p_i$ could be anything between 0 and 1 with equal probability. Nevertheless, we will consider the beta distribution, which includes the uniform distribution as a particular case. The beta distribution is the most widely used continuous prior distribution in the Bayesian approach to the binomial likelihood. Thus, we suppose that the $p_i$ $(i = 1, 2)$ parameters follow a beta prior distribution with parameters $\alpha_i > 0$ and $\beta_i > 0$, respectively. That is, the probability density function of $p_i$ is given by

$$\pi_i(p_i) = \frac{1}{B(\alpha_i, \beta_i)} p_i^{\alpha_i - 1}(1 - p_i)^{\beta_i - 1}, \quad i = 1, 2, \tag{3.4}$$

where $B(a, b)$ is the Beta function given by $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$. The mean and variance of the distribution (3.4) are given by

$$E(p_i) = \frac{\alpha_i}{\alpha_i + \beta_i}, \quad i = 1, 2, \tag{3.5}$$

$$var(p_i) = \frac{\alpha_i \beta_i}{(\alpha_i + \beta_i)^2(\alpha_i + \beta_i + 1)}, \quad i = 1, 2. \tag{3.6}$$

The beta distribution, Beta$(\alpha, \beta)$, can adopt a wide variety of shapes depending on the values of its two parameters. If $\alpha = \beta = 1$, we get the uniform distribution. For $\alpha = \beta = k$, the distribution is symmetric with a mode at $p = 1/2$ if $k > 1$, and has an anti-mode (minimum) at $p = 1/2$ and modes at $p = 0$ and $p = 1$ if $k < 1/2$. Figure 1 displays several plots of the beta distribution for selected values of the parameters.

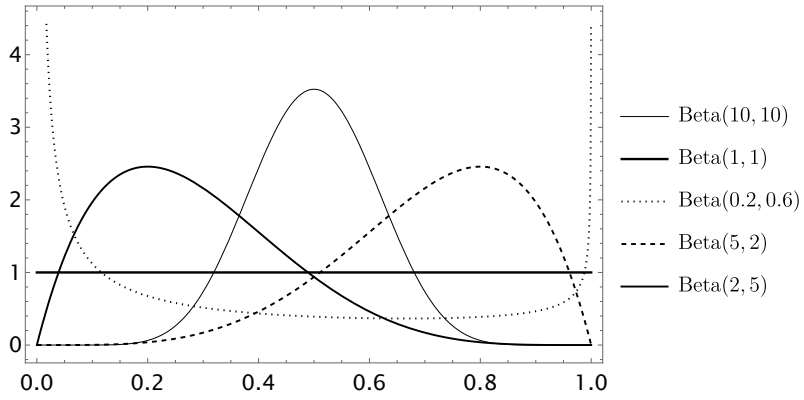


**Figure 1.** Beta distribution for different values of parameters.

Assuming now independence between the two random variables $p_1$ and $p_2$, the joint prior distribution for these two random variables is $\pi(p_1, p_2) = \pi_1(p_1)\pi_2(p_2)$.

Given a sample $(\tilde{x}, \tilde{z}) = \{(x_1, z_1), \ldots, (x_t, z_t)\}$, where $t$ is the sample size, the posterior distribution of $(p_1, p_2)$ given the sample information is computed according to Bayes' theorem and is proportional to the product of the likelihood and the prior distribution. To do so, we obtain that the likelihood function is proportional to

$$\ell((\tilde{x}, \tilde{z})|p_1, p_2) \propto p_1^{t\bar{x}} p_2^{t\bar{z}}(1 - p_1)^{t(n-\bar{x})}(1 - p_2)^{t(\bar{x}-\bar{z})}, \tag{3.7}$$

and the prior distribution is proportional to

$$\pi(p_1, p_2) \propto p_1^{\alpha_1-1} p_2^{\alpha_2-1}(1-p_1)^{\beta_1-1}(1-p_2)^{\beta_2-1}.$$

Thus, the posterior distribution is conjugate with respect to the likelihood (3.7) and remains of the form

$$\pi^*(p_1, p_2|(\tilde{x}, \tilde{z})) \propto p_1^{\alpha_1+t\bar{x}-1} p_2^{\alpha_2+t\bar{z}-1}(1-p_1)^{\beta_1+t(n-\bar{x})-1}(1-p_2)^{\beta_2+t(\bar{x}-\bar{z})-1}, \tag{3.8}$$

where the constant of proportionality does not depend on $p_1$ and $p_2$. Here, $\bar{x} = (1/t)\sum_{i=1}^{t} x_i$ and $\bar{z} = (1/t)\sum_{i=1}^{t} z_i$ are the sample means of $X$ and $Z$, respectively. Observe that we have assumed $n$ to be known, i.e., the number of questions in the multiple-choice test. On the other hand, $x_i$ denotes the number of questions answered, and $z_i$ denotes the number of correctly answered questions. Of course, $x_i - z_i$ represents the number of incorrectly answered questions.

Additionally, the posterior distribution (3.8) is the product of two beta distributions with parameters $\alpha_i^*$ and $\beta_i^*$ $(i = 1, 2)$, where the updated parameters are given by

$$\alpha_1^* = \alpha_1 + t\bar{x}, \tag{3.9}$$
$$\alpha_2^* = \alpha_2 + t\bar{z}, \tag{3.10}$$
$$\beta_1^* = \beta_1 + t(n - \bar{x}), \tag{3.11}$$
$$\beta_2^* = \beta_2 + t(\bar{x} - \bar{z}). \tag{3.12}$$

## 4. Marks and detecting heterogeneity of the classroom

Under the traditional procedure, if the teacher considers that a correct question is scored with $c > 0$ points and a wrong question with $w > 0$ points, the mark that corresponds to a student who answers $x$ questions, of which $z$ are correct, is given by

$$g(x, z) = cz - w(x - z).$$

From now on, we will take $g(x, z) = 0$ for the case where the difference is negative. The main idea is to encourage students with negative scores, which are set to zero, to achieve positive results on subsequent tests.

Now, using (2.2), we obtain, after some algebra, the risk mark as

$$\mathcal{M}(\Theta) = \sum_{x=0}^{n}\sum_{z=0}^{x} g(x, z)f(x, z) = np_1(cp_2 - wq_2),$$

where $q_2 = 1 - p_2$. Such a mark would be correct for a student if parameters $p_1$ and $p_2$ were known. It is important to observe that we assume these two parameters are random and unknown, so this mark is unknown. Because these parameters are unobserved in practice, the risk mark is a theoretical mark that cannot be exactly known and must be estimated from the data. On the other hand, we have the group (prior) mark, which is assigned to a student when nothing is known about them. It is the average mark over all possible risk marks.

Observe that, in our case, unknown parameter $\Theta$ is a vector of unknown parameters; that is, $\Theta = (p_1, p_2)$. Using now (2.3), we get the group mark, which is given by

$$\mathcal{M} = \int_0^1 \int_0^1 \mathcal{M}(p_1, p_2)\pi(p_1, p_2)\, dp_1\, dp_2 = \frac{n\alpha_1(\alpha_2 c - \beta_2 w)}{(\alpha_1 + \beta_1)(\alpha_2 + \beta_2)}. \tag{4.1}$$

Since the prior distributions on $\Theta$ are conjugate, we can obtain the Bayesian marks from (4.1) in a straightforward way by interchanging parameters $\alpha_i$ and $\beta_i$ ($i = 1, 2$) with the updated parameters using (3.9)–(3.12). The result is

$$\begin{aligned} \mathcal{M}^*(t, x, z) &= \frac{n\alpha_1^*(c\alpha_2^* - w\beta_2^*)}{(\alpha_1^* + \beta_1^*)(\alpha_2^* + \beta_2^*)} \\ &= \frac{n(\alpha_1 + t\bar{x})\left[c(\alpha_2 + t\bar{z}) - w(\beta_2 + t(\bar{x} - \bar{z}))\right]}{(\alpha_1 + \beta_1 + tn)(\alpha_2 + \beta_2 + t\bar{x})} \\ &= \frac{n(\alpha_1 + x)\left[c(\alpha_2 + z) - w(\beta_2 + x - z)\right]}{(\alpha_1 + \beta_1 + tn)(\alpha_2 + \beta_2 + x)}. \end{aligned} \tag{4.2}$$

Here, $x = t\bar{x}$ is the total number of responses and $z = t\bar{z}$ is the total number of correct answers, where $\bar{x}$ and $\bar{z}$ are the sample means of $X$ and $Z$, respectively.

Note that $\mathcal{M}^*(0, 0, 0) = \mathcal{M}$. The Bayesian mark coincides with the prior mark when no information is available. Furthermore, expression (4.2) can be rewritten as

$$\mathcal{M}^*(t, x, z) = \gamma(t, x)\mathcal{M} + [1 - \gamma(t, x)]h(t, x, z), \tag{4.3}$$

where

$$\begin{aligned} \gamma(t, x) &= \frac{(\alpha_1 + \beta_1)(\alpha_2 + \beta_2)}{(\alpha_1 + \beta_1 + tn)(\alpha_2 + \beta_2 + x)}, \\ h(t, x, z) &= \frac{(\alpha_1 + x)[cz - w(x - z)] + x(c\alpha_2 - w\beta_2)}{(\alpha_1 + \beta_1)x/n + t(\alpha_2 + \beta_2 + x)}. \end{aligned} \tag{4.4}$$

Therefore, the Bayesian mark $\mathcal{M}^*(t, x, z)$ is written as a convex sum of the group mark $\mathcal{M}$ and a function depending on the total responses $x$, the correct responses $z$, and $t$. The Bayesian mark satisfies the expression suggested in (2.4).

Observe that $\gamma(t, x)$ takes values in $[0, 1]$ with $\gamma(0, x) = 1$ (when $t = 0$, i.e., no test is carried out, then $x = 0$) and $\gamma(\infty, x) \to 0$. Therefore, if we consider that $t$ is the number of exams taken by a student in an academic year, for example, when no exam has been taken (at the beginning of the course), the mark corresponding to a student is given by $\mathcal{M}$, and after $t$ exams with total $x$ responses and $z$ correct answers, the mark achieved by the student is given by $\mathcal{M}^*(t, x, z)$. The case $\gamma(t, x) = 0$ means that only experience is used to determine the final mark. The case $\gamma(t, x) = 1$ corresponds to a situation in which experience is ignored and external information is used to evaluate. Note that the greater the weight $\gamma(t, x)$ attributed to the group grade, the smaller the effect of the student's experience. An illustration of expression (4.3) is shown in Figure 2, in which the position of the Bayesian mark with respect to $\mathcal{M}$ and $h(t, x, z)$, depending on weight $\gamma(t, x)$ is illustrated. When $t$ is assumed to be extremely large, the mark achieved by the student will be $\lim_{t\to\infty} h(t, x, z)$. That is, $\mathcal{M}^*(\infty, x, z) = c\bar{z} - w(\bar{x} - \bar{z})$, which can be obtained easily from (4.2).

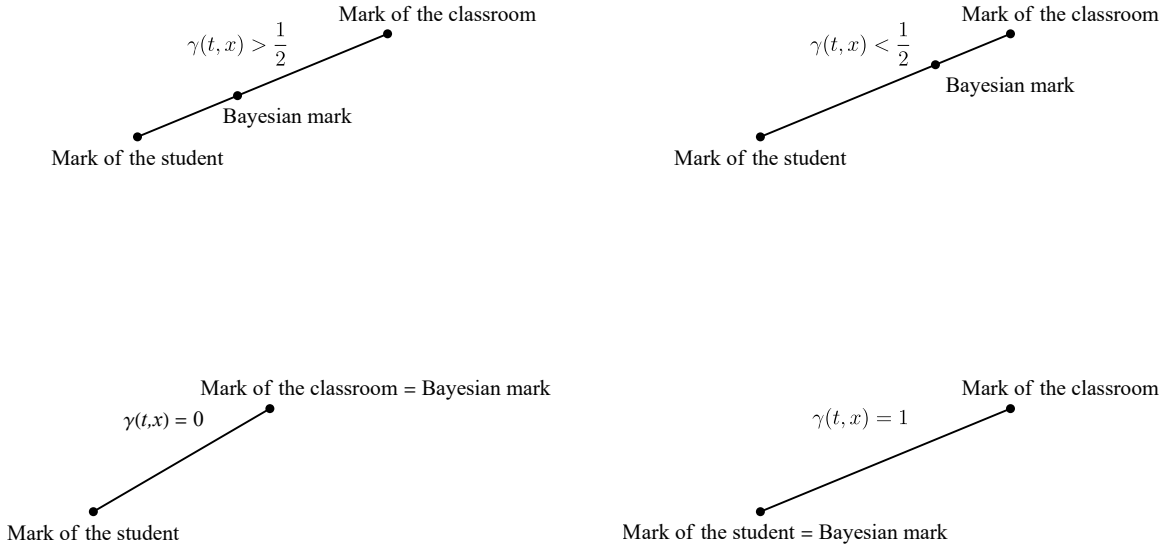


**Figure 2.** Different positions of the Bayesian mark depending on weight $\gamma(t, x)$.

### *4.1. Testing the heterogeneity of the classroom*

Considering that $E(X|p_1) = np_1$, $var(X|p_1) = np_1(1 - p_1)$, $E(Z|p_2) = xp_2$, and $var(Z|p_2) = xp_2(1 - p_2)$, and using the mean and variance of the beta distribution given in (3.5) and (3.6), it is straightforward to show that

$$var[E(X|p_1)] = \frac{n^2\alpha_1\beta_1}{(\alpha_1+\beta_1)^2(\alpha_1+\beta_1+1)}, \tag{4.5}$$

$$E[var(X|p_1)] = \frac{n\alpha_1\beta_1}{(\alpha_1+\beta_1)(\alpha_1+\beta_1+1)}, \tag{4.6}$$

$$var[E(Z|p_2)] = \frac{x^2\alpha_2\beta_2}{(\alpha_2+\beta_2)^2(\alpha_2+\beta_2+1)}, \tag{4.7}$$

$$E[var(Z|p_2)] = \frac{x\alpha_2\beta_2}{(\alpha_2+\beta_2)(\alpha_2+\beta_2+1)}, \tag{4.8}$$

and therefore expression (4.4) is equivalent to

$$\gamma(t,x) = \frac{1}{(1+\kappa_1)(1+\kappa_2)},$$

where $\kappa_1 = t\, var[E(X|p_1)]/E[var(X|p_1)]$ and $\kappa_2 = var[E(Z|p_2)]/E[var(Z|p_2)]$. Here, $v_1 = var[E(X|p_1)]$ and $v_2 = var[E(Z|p_2)]$ are the variances of the total and the correct responses and, therefore, measure the heterogeneity of the classroom. In contrast, $s_1 = E[var(X|p_1)]$ and $s_2 = E[var(Z|p_2)]$ are global measures of the dispersion of the responses and the correct answers, respectively. The $\gamma(t, x)$ increases

as classroom heterogeneity increases or within-risk variability decreases (which intuitively makes sense). In summary, as $t$ increases, the weight $1-\gamma(t,x)$ also increases, reflecting increased reliability of the sample. Conversely, when $s_i$, $i=1,2$, increases, the weight $\gamma(t,x)$ also increases, indicating greater variability within observations and suggesting lower reliability of the sample mean. As $v_i$, $i=1,2$, increase, this weight decreases, reflecting the higher variability across group means. These comments are summarized in Table 2.

**Table 2.** Marks depending on whether the class is homogeneous or heterogeneous.

| $v_1, v_2$ | $s_1, s_2$ | $\gamma(x,t)$ | Classroom | $\mathcal{M}^*(t,x,z)$ |
|---|---|---|---|---|
| Large | Small | 0 | Heterogeneous | $h(t,x,z)$ |
| Small | Large | 1 | Homogeneous | $\mathcal{M}$ |

### *4.2. Rewarding or penalizing to the student*

Now, from (4.2), we can estimate the marks achieved by a student by normalizing (4.3) in the following way: For an exam with $x$ questions where the student answers them all correctly, $x=z$, and in most academic settings, the student would achieve the maximum grade, usually 10. Therefore, if we define $x^*$ as the largest possible value for $x$ and $z^*$ as the largest possible value for $z$, the expression given by

$$\mathcal{M}^{**}(t,x,z)=\frac{10\mathcal{M}^*(t,x,z)}{\mathcal{M}^*(1,x^*,z^*)} \tag{4.9}$$

is simply a normalization of (4.3) and assigns the highest mark, 10, to a student who answers all $x$ questions correctly on the first exam, i.e., $t=1$.

The mark satisfies the following rules:

$$\frac{\partial\mathcal{M}^{**}(t,x,z)}{\partial z} = \frac{10}{\mathcal{M}^*(1,x^*,z^*)}\frac{n(c+w)(\alpha_1+x)}{(\alpha_1+\beta_1+nt)(\alpha_2+\beta_2+x)}\geq 0, \tag{4.10}$$

$$\frac{\partial\mathcal{M}^{**}(t,x,z)}{\partial t\partial z} = -\frac{10}{\mathcal{M}^*(1,x^*,z^*)}\frac{n^2(c+w)(\alpha_1+x)}{(\alpha_1+\beta_1+nt)^2(\alpha_2+\beta_2+x)}\leq 0, \tag{4.11}$$

$$\frac{\partial\mathcal{M}^{**}(t,x,z)}{\partial t} = -\frac{10}{\mathcal{M}^*(1,x^*,z^*)}\frac{n\mathcal{M}^{**}(t,x,z)}{(\alpha_1+\beta_1+nt)}. \tag{4.12}$$

Expression (4.10) ensures that the marks increase with the number of correct answers and decrease with incorrect answers when $x$ and $t$ remain constant. A measure of this increasing or decreasing rate is obtained from expression (4.11). Finally, expression (4.12) establishes that, when the Bayesian mark is positive, the mark decreases with $t$ when $x$ and $z$ remain constant. When the Bayesian mark is negative, the mark increases with $t$ when $x$ and $z$ remain constant. In conclusion, the transition rules reward or penalize students depending on their academic performance.

## 5. Numerical experiment

For illustration purposes, we demonstrate the model with data collected from 338 students in the Mathematics for Business course of a Business Administration degree at a Spanish university. The

subject is offered in the first year and first semester of the degree, and students are distributed into six groups. The Business degree admits students from different backgrounds, including high school, Science, Humanities, and Social Sciences, all with different mathematical backgrounds. The test consists of a multiple-choice exam with five questions. Each correct answer earns two points, and each incorrect answer subtracts one point. A scatterplot of the data used, showing the joint distribution of $(x, z)$ across the 338 students, is shown in Figure 3.

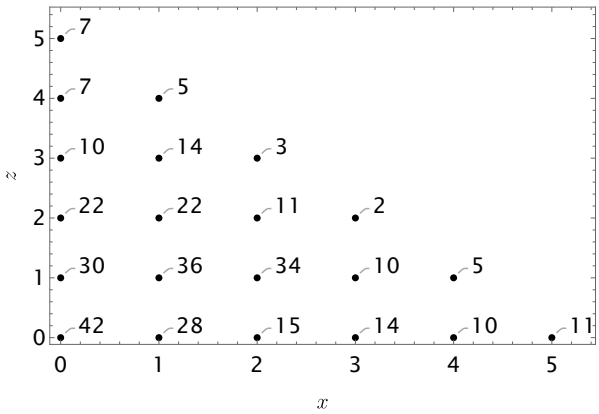


**Figure 3.** Scatterplot of the data showing the joint distribution of $(x, z)$ across the 338 students.

### *5.1. Empirical Bayes approach*

We adopt here an empirical Bayes approach, which can be pictured as a compromise between the Bayesian and classical approaches (see [7, 31]), where the parameters of the prior distributions are estimated from the data. For that, we need the marginal (unconditional) distribution of $(X, Z)$, which is computed by compounding as

$$f(x,z) = \int_0^1 \int_0^1 f(x,z|p_1,p_2)\pi(p_1,p_2)\,dp_1\,dp_2,$$

where $f(x, z|p_1, p_2)$ is given in (3.3) and the prior distribution $\pi(p_1, p_2)$ is given by the product of the two priors in (3.4). This requires some computations and yields

$$f(x,z) = \frac{n!B(\beta_1+n-x,\alpha_1+x)B(\alpha_2+z,\beta_2+x-z)}{(n-x)!(x-z)!z!B(\alpha_1,\beta_1)B(\alpha_2,\beta_2)}. \tag{5.1}$$

The unconditional cross-factorial moment and the marginal means are provided in the Appendix. Then, the unconditional covariance is given by

$$cov(X,Z) = \frac{n\alpha_1\alpha_2\beta_1(n+\alpha_1+\beta_1)}{(\alpha_1+\beta_1)^2(\alpha_2+\beta_2)(\alpha_1+\beta_1+1)}. \tag{5.2}$$

The log-likelihood function is also shown in the Appendix.

### 5.2. *Empirical and theoretical fitted values*

The multiple-choice test done by the 338 students provide the results shown in Table 3. Rows correspond to the number of questions answered by the student, with $X = 0, 1, ..., n$, while columns represent the number of correct answers, $Z = 0, 1, ..., X$. As a consequence, the table exhibits a triangular structure, reflecting the natural restriction $Z \leq X$, which arises from the fact that the number of correct responses cannot exceed the total number of attempted questions.

**Table 3.** Observed (in bold) and expected data (below) for estimated values of parameters. The standard errors of the estimated parameters are shown below the table, in brackets

| $X$ \ $Z$ | 0 | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| 0 | **42** | | | | | | **42** |
| | 39.53 | | | | | | 39.53 |
| 1 | **30** | **28** | | | | | **58** |
| | 30.90 | 32.70 | | | | | 63.61 |
| 2 | **22** | **36** | **15** | | | | **73** |
| | 21.95 | 27.90 | 24.04 | | | | 73.90 |
| 3 | **10** | **22** | **34** | **14** | | | **80** |
| | 14.59 | 19.87 | 20.57 | 16.37 | | | 71.41 |
| 4 | **7** | **14** | **11** | **10** | **10** | | **52** |
| | 8.62 | 12.19 | 13.52 | 12.93 | 9.87 | | 57.13 |
| 5 | **7** | **5** | **3** | **2** | **5** | **11** | **33** |
| | 3.80 | 5.49 | 6.31 | 6.47 | 5.93 | 4.41 | 32.42 |
| Total | **118** | **105** | **63** | **26** | **15** | **11** | **338** |
| | 119.39 | 98.15 | 64.44 | 35.77 | 15.80 | 4.41 | 338.00 |

$\widehat{\alpha}_1 = 1.963\,(0.312)$; $\widehat{\beta}_1 = 2.102\,(0.335)$
$\widehat{\alpha}_2 = 1.591\,(0.342)$; $\widehat{\beta}_2 = 1.503\,(0.319)$
AIC = 1935.18
$\chi^2_7 = 3.72$; $p$-value = 81.08%

Each cell in the table contains two values: The observed frequency (reported in bold) and the corresponding expected frequency (shown below), obtained from the fitted model. The expected frequencies are computed using the probability mass function derived in Eq (5.1), evaluated at the estimated parameter values, which are at the bottom of this table (the standard errors are in brackets). As observed in the sample, of the 338 students, 42 do not attempt any question, whereas 58 students answer exactly one question, of whom 28 provide correct responses and 30 incorrect ones. Overall, the close agreement between observed and expected frequencies indicates a satisfactory fit of the model. We also show the value obtained for the Akaike information criterion (AIC). Note that AIC $= 2(k - \ell_{\max})$, where $k$ is the number of model parameters and $\ell_{\max}$ is the maximum value of the log-likelihood function; see [1] for details. The goodness-of-fit was determined by the standard Pearson chi-squared test statistic, given by

$$\chi^2 = \sum_{i=1}^{n} \frac{(\text{observed}_i - \text{expected}_i)^2}{\text{expected}_i},$$

with the following grouping procedure: The outermost classes are consolidated to produce theoretical class sizes of 5 or larger. It is known that $\chi^2$ follows a chi-squared distribution with $n_0 - k - 1$ degrees of freedom, where $n_0$ is the number of classes considered to compute the value of $\chi^2$. In this case, we obtain $\chi^2_7 = 3.72$ (which is well below 14.07, the 5% critical value of $\chi^2$ with 7 degrees of freedom), with a $p$-value of 81.08%. From the estimated values of the parameters, the estimated covariance obtained from (5.2) is $cov(X, Z) = 1.14892$, versus the sample value of 1.22818.

A comparison between observed and expected frequencies can also be seen in Figure 4.

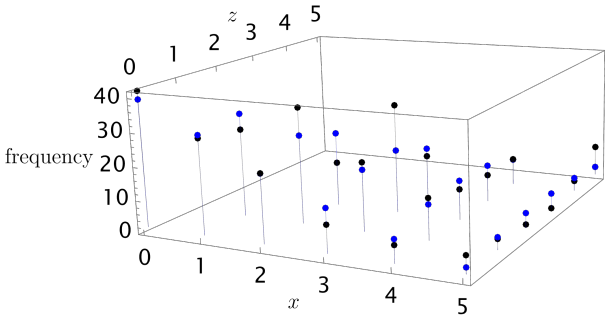


**Figure 4.** Observed (black) and expected frequencies (red) for estimated values of the parameters.

It is well established that estimating hyperparameters from the data and subsequently treating them as fixed quantities may lead to an underestimation of posterior uncertainty. In particular, ignoring the variability associated with the hyperparameter estimates can result in overly narrow credible intervals and an overconfident assessment of point estimates. The statistical literature has proposed several remedies to address this issue, most notably fully Bayesian approaches in which hyperparameters are endowed with prior distributions and inference is performed jointly, typically via Markov Chain Monte Carlo (MCMC) methods, or alternative approaches that explicitly account for this additional layer of uncertainty.

In this setting, however, the empirical performance of the model suggests that the impact of this effect is likely to be limited. As evidenced by the goodness-of-fit measures reported in Table 3, the model provides an accurate representation of the data. Consequently, while a full propagation of hyperparameter uncertainty would in principle be preferable, we believe that its omission does not materially affect the main results in this particular application.

### *5.3. Measuring the heterogeneity of the classroom*

Using the expectation and variance given in (4.5) and (4.6), and the estimated values of the parameters, we obtain that $var[E(X|p_1)] = 1.232$ while $E[var(X|p_1)] = 1.002$. Figure 5 (above) shows the values of expressions (4.7) and (4.8) for different values of $x$. Considering these values (which are relatively high), together with the values of $\gamma(x, t)$ shown in Figure 5 (below) (which are relatively

low), we conclude that the classroom in the example analyzed appears to be heterogeneous, since the Bayesian mark assigns more weight to the expression $h(t, x, z)$.

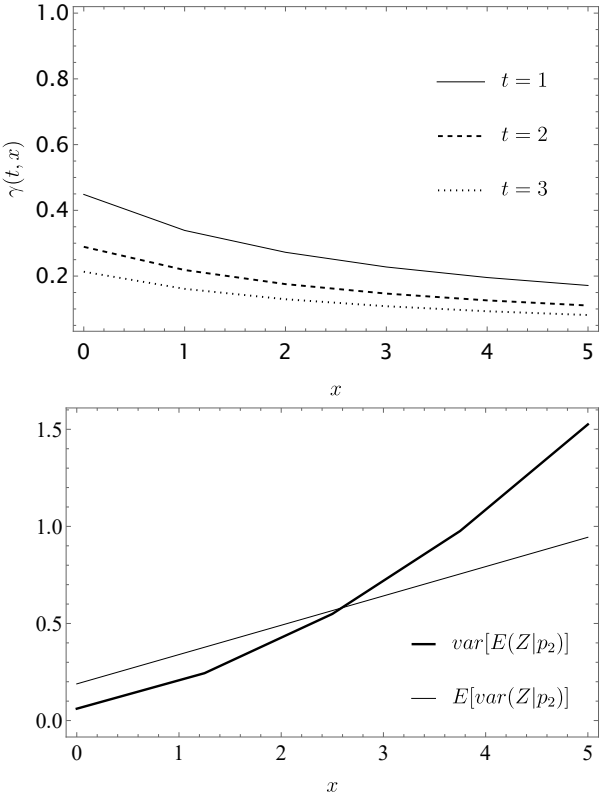


**Figure 5.** Factor measuring the weight of the mark concerning the group (above) and some measures of heterogeneity (below).

### *5.4. Bayesian marks*

The proposed approach enables the computation of grades, potentially motivating students to enhance their performance from one assessment to the next. This strategy may help support students at risk of failing, although such effects are not directly evaluated. Otherwise, there are no rewards if the student remains in the same class and does not improve from one test to the next. To obtain the data in Table 4, the method uses the collective experience of the heterogeneous group.

We assume the tests have five questions, $n = 5$, each with three alternative responses*. These alternatives include only one correct answer, which scores 2 points, $c = 2$; otherwise, each incorrect answer results in $-1$ points, $w = 1$. Furthermore, we assume the existence of only three tests ($t = 3$). It is well known that if the number of possible alternatives in each question is $P > 0$, then $w = \frac{c}{P-1}$ ensures that, if the student answers all the questions randomly, the expected mark is zero.

We now illustrate how the process operates using Table 4. In the first test, $t = 1$, with $x = 5$ answered questions, if all of them are correct, $z = 5$, the student receives the maximum score in the classical (CS) and the Bayesian systems. We assume that if the student responds to all five questions correctly, their score will always be 10 points. On the other hand, if the model provides a negative

*The number of answer options can be generalized, and the choice of 3 responses is illustrative rather than restrictive.

mark, as in the CS, the mark is converted to zero to encourage the student to strive for improvement. For instance, as can be seen in Table 4, when $x = 5$ and $z = 1$, the score would be $-2$; however, it is assigned a value of zero.

**Table 4.** Normalized Bayesian marks and comparisons with the ones obtained under the classical method (CS), shown in the fifth row in bold.

| $t$ | $x = 5$ | | | | | | $x = 4$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $z = 5$ | $z = 4$ | $z = 3$ | $z = 2$ | $z = 1$ | $z = 0$ | $z = 4$ | $z = 3$ | $z = 2$ | $z = 1$ | $z = 0$ |
| 1 | 10 | 7.43 | 4.86 | 2.29 | 0 | 0 | 8.09 | 5.59 | 3.08 | 0.56 | 0 |
| 2 | 10 | 4.78 | 3.13 | 1.48 | 0 | 0 | 5.21 | 3.60 | 1.98 | 0.36 | 0 |
| 3 | 10 | 3.53 | 2.31 | 1.09 | 0 | 0 | 3.85 | 2.65 | 1.46 | 0.27 | 0 |
| Mean | 10 | 5.24 | 3.44 | 1.62 | 0 | 0 | 5.72 | 3.95 | 2.17 | 0.40 | 0 |
| CS | **10** | **7** | 4 | **1** | **0** | **0** | **8** | **5** | **2** | **0** | **0** |

| $t$ | $x = 3$ | | | | $x = 2$ | | | $x = 1$ | |
|---|---|---|---|---|---|---|---|---|---|
| | $z = 3$ | $z = 2$ | $z = 1$ | $z = 0$ | $z = 2$ | $z = 1$ | $z = 0$ | $z = 1$ | $z = 0$ |
| 1 | 6.22 | 3.79 | 1.36 | 0 | 4.40 | 2.07 | 0 | 2.65 | 0.49 |
| 2 | 4.01 | 2.44 | 0.88 | 0 | 2.83 | 1.34 | 0 | 1.71 | 0.31 |
| 3 | 2.95 | 1.80 | 0.65 | 0 | 2.09 | 0.98 | 0 | 0.53 | 0.10 |
| Mean | 4.39 | 2.68 | 0.96 | 0 | 3.11 | 1.46 | 0 | 1.63 | 0.30 |
| CS | **6** | **3** | **0** | **0** | **4** | **1** | **0** | **2** | **0** |

In general, in subsequent tests, $t = 2$ and $t = 3$, if the student obtains the same number of correct answers from one test to the following, they will remain in the same category, as there is no improvement, and the proposed Bayesian rating assigns a slightly lower grade. Focusing on students who consistently respond to, for example, four questions, then $x = 4$. In the first test ($t = 1$), if two of these answers are correct, $z = 2$, the corresponding Bayesian mark would be 3.08 instead of 2 in the classical system. Motivated students can correctly answer 3 questions, $z = 3$, in the following test ($t = 2$), resulting in a mark of 3.60. If this progression continues and reaches $z = 4$ in the final test ($t = 3$), the mark will be 3.85.

Table 5 shows the evolution of two average students over the course and the means calculated using the classical and Bayesian models. For Student 1, in the initial test, where the student answers all the questions but only three correctly, the Bayesian mark would be 4.86 instead of 4 under the classical method.

However, suppose the student does not improve in the following test and answers only four questions, of which two are correct. In this case, the Bayesian mark (1.98), although lower than in the previous test, is almost the same as in the classical method (2.2). Finally, in the third test, after answering five questions with only two correct answers, the marks would be 1.09 and 1 in the Bayesian and classical models, respectively. The final mean score is slightly higher in the Bayesian

model. Student 2 represents a case that begins with very poor grades but improves in the subsequent tests. A graphical representation of this evolution is shown in Figure 6.

**Table 5.** Evolution of the grades of two average students under the classical and Bayesian models.

| Student 1 | $(t, x, z) = (1, 5, 3)$ | $(t, x, z) = (2, 4, 2)$ | $(t, x, z) = (3, 5, 2)$ | Mean |
|---|---|---|---|---|
| Bayesian | 4.86 | 1.98 | 1.09 | 2.64 |
| CS | 4 | 2 | 1 | 2.34 |
| Student 2 | $(t, x, z) = (1, 3, 1)$ | $(t, x, z) = (2, 4, 4)$ | $(t, x, z) = (3, 5, 5)$ | Mean |
| Bayesian | 1.36 | 3.85 | 10 | 5.53 |
| CS | 0 | 8 | 10 | 6 |

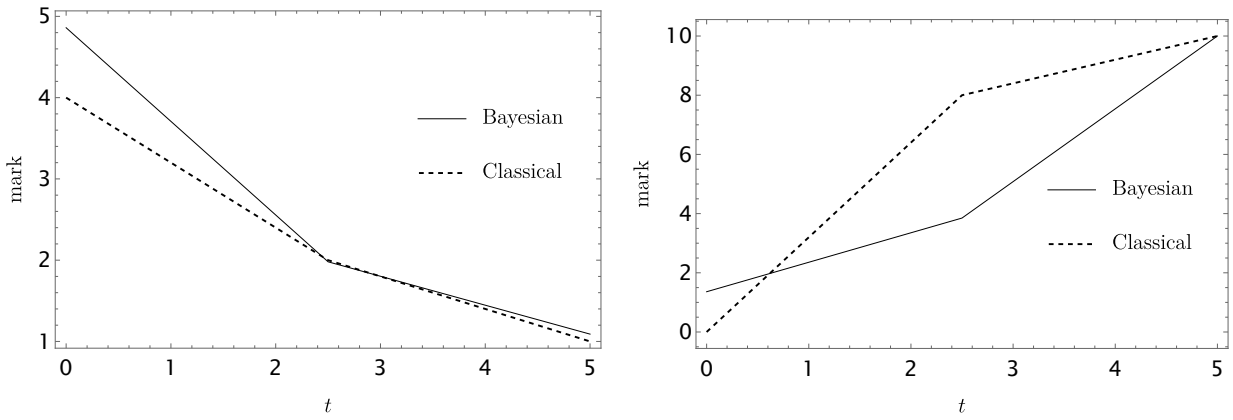


**Figure 6.** Evolution of two students along the course for the classical and Bayesian marks. Student 1 (left) and student 2 (right).

### *5.5. Breaking the independence assumption*

The assumption of independence between the two random variables $p_1$ and $p_2$ is very restrictive. As an anonymous reviewer has pointed out, in practice, students who answer more questions are also more likely to answer them correctly. Although it is not applied here, a dependent bivariate prior can be introduced using a copula. A very simple copula is the one provided by the Sarmanov family of bivariate distributions (see [20]). Specifically, assume that

$$\pi(p_1, p_2) = \pi_1(p_1)\pi_2(p_2)\left[1 + \omega\, \phi_1(p_1)\phi_2(p_2)\right]$$

provided that $\omega$ is a real number that satisfies condition $1 + \omega\phi_1(p_1)\phi_2(p_2) \geq 0$ for all $p_1$ and $p_2$, so that it defines a genuine bivariate probability density function with specified marginals $\pi_1(p_1)$ and $\pi_2(p_2)$. Note that the dependence is captured through the mixing functions and the parameter $\omega$. In practice, various mixing functions may be selected to determine the bounds of the $\omega$ parameter. The researchers in [20] propose the mixing functions given by $\phi_i(p_i) = p_i - \alpha_i/(\alpha_i + \beta_i)$, for $i = 1, 2$. In this case, the posterior density of $(p_1, p_2)$ is a linear combination of products of univariate beta densities, which results in an almost conjugate form [20].

## 6. Conclusions and future directions

In this paper, we proposed a Bayesian framework for modeling and evaluating student performance on multiple-choice tests, explicitly accounting for classroom heterogeneity. The approach is based on a bivariate probabilistic structure that jointly models the number of attempted questions and the number of correct answers, embedded within a statistical decision-theoretic framework.

We made this framework by the combined effect of two stochastic processes; one concerned with the fact of answering or not answering a test question, and if answered, whether the answer was correct or incorrect. The student's performance can be due to many causes. However, fundamentally, it is found in students' subject-matter skills and in variations in their provenance and knowledge in the subject we consider. These differences in the performances are related to the heterogeneity of the population under study (the students).

A key contribution of the study is the derivation of an explicit scoring rule in the form of a Bayes estimator under quadratic loss. The resulting expression for the student's mark combines individual performance and prior (group-level) information, yielding a credibility-type formulation that enables a transparent interpretation of the weighting mechanism. This structure provides a flexible way to balance individual experience with information from the collective, depending on the degree of heterogeneity and the amount of available data.

The empirical results illustrate that the proposed scoring system produces outcomes broadly comparable to those obtained under traditional evaluation methods. Furthermore, the Bayesian approach introduces a dynamic component by incorporating accumulated performance, leading to smoother adjustments in marks across successive assessments.

The proposed model may provide a framework that encourages improved academic performance by incorporating individual and group-level information into the evaluation process. Although the final marks obtained are broadly comparable to those derived from classical scoring systems, the Bayesian approach may yield relatively higher scores for students who do not reach the highest performance levels. Importantly, this adjustment should not be interpreted as an artificial inflation of grades or as a consequence of subjective assessment, but rather as the result of a systematic incorporation of individual outcomes and collective information within a coherent probabilistic framework. In this sense, the proposed methodology is consistent with the findings in [13], as it aims to account for classroom heterogeneity to support a more individualized and potentially motivating assessment process, rather than generating spurious grade inflation.

Student failure is widely recognized as a key challenge in higher education, often associated with low initial performance and a consequent loss of motivation. Poor results in early assessments may discourage students and negatively affect their engagement with the course. In this context, educational institutions frequently implement support mechanisms aimed at mitigating these effects and fostering student persistence. The proposed assessment framework may be interpreted as a complementary tool in this direction, as it incorporates accumulated performance and group-level information into the grading process and is consistent with mechanisms that could mitigate the negative impact of initial low scores. However, this interpretation should be regarded as suggestive, and further empirical analysis would be required to assess the extent to which such an approach may influence student motivation or retention.

Finally, the proposed methodology is consistent with the philosophy underlying the European

grading scale defined within the Bologna Process, where student performance is interpreted in relation to the distribution of results within a group. By integrating individual and collective information, the model provides a probabilistic framework that is compatible with this perspective and raises relevant considerations regarding fairness in the evaluation process. In particular, the approach may contribute to a smoother evaluation of early performance, which could help mitigate discouraging effects in the initial stages of the academic term, although such behavioral implications were not directly evaluated in this study. Furthermore, as with any model-based assessment procedure, its implementation may involve unintended consequences if not carefully designed and communicated. Nevertheless, any potential impact on student retention should be interpreted with caution, as it was not directly evaluated in this study.

Several avenues for future research emerge from this study. First, the analysis could be extended by considering alternative loss functions within the decision-theoretic framework. While the squared-error loss leads to analytically convenient expressions and a credibility-type formulation, other choices (e.g., asymmetric or absolute loss functions) may provide different interpretations of the scoring rule and could be more appropriate in specific educational contexts. Second, the current model assumes independence between the latent parameters governing response behavior and correctness. Relaxing this assumption by introducing dependence structures (for instance, through hierarchical models or copula-based approaches) would enable a more realistic representation of student ability, since more skilled students are likely to attempt more questions and to answer them correctly. Third, researchers could explore connections with alternative modeling frameworks commonly used in educational assessments, particularly item response theory (IRT) models. A comparative analysis could help assess the relative advantages of the proposed approach in terms of interpretability, flexibility, and predictive performance. Fourth, the empirical analysis could be extended using larger and more diverse datasets, including longitudinal data, to better evaluate the dynamic properties of the proposed scoring system and its ability to capture learning progression over time. Finally, an important direction for future research is the empirical assessment of the behavioral implications of the proposed methodology. In particular, experimental or quasi-experimental studies could be designed to investigate whether the use of this scoring system has an effect on student motivation, engagement, and retention. Such analyses would be necessary to formally establish any causal relationship between the proposed evaluation framework and educational outcomes.

## Author contributions

Emilio Gómez-Déniz, Nancy Dávila-Cárdenes and José María Pérez-Sánchez: Conceptualization, Methodology, Validation, Writing-original draft, Writing-review & editing. All authors contributed equally to this work.

## Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

## Acknowledgments

EGD was partially funded by grant PID2021-127989OB-I00 (Ministerio de Economía y Competitividad, Spain

## Conflict of interest

The authors declare no conflict of interest.

## References

1. H. Akaike, A new look at the statistical model identification, *IEEE Trans. Automat. Contr.*, **19** (1974), 716–723. https://doi.org/10.1109/TAC.1974.1100705
2. J. Alberola, E. Del Val, V. Sanchez-Anguix, A. Palomares, M. Teruel, An artificial intelligence tool for heterogeneous team formation in the classroom, *Knowl.-Based Syst.*, **101** (2016), 1–14. https://doi.org/10.1016/j.knosys.2016.02.010
3. M. Bahar, Student perception of shift from homogeneous grouping to heterogeneous grouping at a university class, *Procedia–Social and Behavioral Sciences*, **46** (2012), 1886–1892. https://doi.org/10.1016/j.sbspro.2012.05.397
4. J. Berger, *Statistical decision theory and Bayesian analysis*, New York: Springer, 1985. https://doi.org/10.1007/978-1-4757-4286-2
5. R. Bertolini, S. Finch, R. Nehm, An application of Bayesian inference to examine student retention and attrition in the STEM classroom, *Front. Educ.*, **8** (2023), 1073829. https://doi.org/10.3389/feduc.2023.1073829
6. H. Bühlmann, A. Gisler, *A course in credibility theory and its applications*, Berlin: Springer, 2005. https://doi.org/10.1007/3-540-29273-X
7. G. Casella, An introduction to empirical Bayes data analysis, *Amer. Stat.*, **39** (1985), 83–87. https://doi.org/10.1080/00031305.1985.10479400
8. D. Gabaldón-Estevan, Heterogeneity versus homogeneity in schools: a study of the educational value of classroom interaction, *Educ. Sci.*, **10** (2020), 335. https://doi.org/10.3390/educsci10110335
9. E. Gómez-Déniz, A generalization of the credibility theory obtained by using the weighted balanced loss function, *Insur. Math. Econ.*, **42** (2008), 850–854. https://doi.org/10.1016/j.insmatheco.2007.09.002
10. M. Haas, C. Caprani, B. Van Beurden, Bayesian generative modelling of student results in course networks, *Journal of Learning Analytics*, **10** (2023), 135–152. https://doi.org/10.18608/jla.2023.7957
11. M. Homer, The future of quantitative educational research methods: bigger, better and, perhaps, Bayesian? *Hillary Place Papers*, **3** (2016), 1–12. https://doi.org/10.48785/100/230
12. S. Hooper, M. Hannafin, The effects of group composition on achievement, interaction, and learning efficiency during computer-based cooperative instruction, *ETRD*, **39** (1991), 27–40. https://doi.org/10.1007/BF02296436

13. C. Jephcote, E. Medland, S. Lygo-Baker, Grade inflation versus grade improvement: are our students getting more intelligent? *Assess. Eval. High. Educ.*, **46** (2021), 547–571. https://doi.org/10.1080/02602938.2020.1795617
14. W. Jewell, Credible means are exact Bayesian for exponential families, *ASTIN Bull.*, **8** (1974), 77–90. https://doi.org/10.1017/S0515036100009193
15. D. Johnson, R. Johnson, Learning together and alone: overview and meta-analysis, *Asia Pac. J. Educ.*, **22** (2002), 95–105. https://doi.org/10.1080/0218879020220110
16. M. Johnson, F. Jenkins, A Bayesian hierarchical model for large-scale educational surveys: an application to the National Assessment of Educational Progress, *ETS Research Report Series*, **2004** (2004), i-28. https://doi.org/10.1002/j.2333-8504.2004.tb01965.x
17. T. Kärner, J. Warwas, S. Schumann, A learning analytics approach to address heterogeneity in the classroom: the teachers' diagnostic support system, *Tech. Know. Learn.*, **26** (2021), 31–52. https://doi.org/10.1007/s10758-020-09448-4
18. J. Kim, K. Choi, Closing the gap: modeling within-school variance heterogeneity in school effect studies, *Asia Pacific Educ. Rev.*, **9** (2008), 206–220. https://doi.org/10.1007/BF03026500
19. M. Kubsch, I. Stamer, M. Steiner, K. Neumann, I. Parchmann, Beyond p-values: using bayesian data analysis in science education research, *Practical Assesssment, Research, and Evaluation*, **26** (2021), 4. https://doi.org/10.7275/vzpw-ng13
20. M. Lee, Properties and applications of the Sarmanov family of bivariate distributions, *Commun. Stat.-Theor. Meth.*, **25** (1996), 1207–1222. https://doi.org/10.1080/03610929608831759
21. P. Lee, *Bayesian statistics: an introduction*, Hoboken: John Wiley & Sons, Inc., 1997.
22. O. Lopez, Classroom diversification: a strategic view of educational productivity, *Rev. Educ. Res.*, **77** (2007), 28–80. https://doi.org/10.3102/003465430298571
23. W. Lyu, J. Kim, Y. Suk, Estimating heterogeneous treatment effects within latent class multilevel models: a bayesian approach, *J. Educ. Behav. Stat.*, **48** (2023), 3–36. https://doi.org/10.3102/10769986221115446
24. M. Matthews, Gifted students talk about cooperative learning, *Educ. Leadership*, **50** (1992), 48–50.
25. M. Mosia, F. Egara, F. Nannim, M. Basitere, Bayesian hierarchical modelling of student academic performance: the impact of mathematics competency, institutional context, and temporal variability, *Educ. Sci.*, **15** (2025), 177. https://doi.org/10.3390/educsci15020177
26. M. Mosia, L. Sekonyela, F. Egara, E. Nimy, I. Mabokgole, F. Nannim, Bayesian hierarchical modelling of academic orientation and advising effects on student retention and progression: multi-cohort evidence, *PLoS One*, **21** (2026), e0345001. https://doi.org/10.1371/journal.pone.0345001
27. P. Murphy, J. Greene, C. Firetto, M. Li, N. Lobczowski, R. Duke, et al., Exploring the influence of homogeneous versus heterogeneous grouping on students' text-based discussions and comprehension, *Contemp. Educ. Psychol.*, **51** (2017), 336–355. https://doi.org/10.1016/j.cedpsych.2017.09.003
28. J. Oetzel, Explaining individual communication processes in homogeneous and heterogeneous groups through individualism-collectivism and self-construal, *Hum. Commun. Res.*, **25** (1998), 202–224. https://doi.org/10.1111/j.1468-2958.1998.tb00443.x
29. M. Pozas, V. Letzel, C. Schneider, Teachers and differentiated instruction: exploring differentiation practices to address student diversity, *J. Res. Spec. Educ. Need.*, **20** (2020), 217–230. https://doi.org/10.1111/1471-3802.12481

30. G. Raftu, Methods and techniques of instruction individualization and differentiation. Learning through cooperation or group work, *Bulletin of the Transilvania University of Braşov, Series VII: Social Sciences and Law*, **9** (2016), 83–90.
31. H. Robbins, The empirical Bayes approach to statistical decision problems, *Ann. Math. Stat.*, **35** (1964), 1–20. https://doi.org/10.1214/aoms/1177703729
32. C. Robert, *The Bayesian choice: from decision-theoretic foundations to computational implementation*, New York: Springer, 2007. https://doi.org/10.1007/0-387-71599-1
33. H. Ruskeepaa, *Mathematica navigator: mathematics, statistics and graphics*, 3 Eds., Burlington: Academic Press, 2009.
34. S. Samsudin, J. Das, N. Rai, Cooperative learning: heterogeneous vs homogeneous grouping, *Proceedings of the Asia-Pacific Education Research Conference*, 2006, 1–6.
35. N. Schullery, S. Schullery, Are heterogeneous or homogeneous groups more beneficial to students? *J. Manag. Educ.*, **30** (2006), 542–556. https://doi.org/10.1177/1052562905277305
36. S. Sirin, Socioeconomic status and academic achievement: a meta-analytic review of research, *Rev. Educ. Res.*, **75** (2005), 417–453. https://doi.org/10.3102/00346543075003417
37. J. Thomsen, Exploring the heterogeneity of class in higher education: social and cultural differentiation in Danish university programmes, *Brit. J. Sociol. Educ.*, **33** (2012), 565–585. https://doi.org/10.1080/01425692.2012.659458

## A. Appendix

The unconditional cross-factorial moment is given by

$$E(XZ) = \frac{n\alpha_1\alpha_2\,(n(1+\alpha_1)+\beta-1)}{(\alpha_1+\beta_1)(\alpha_2+\beta_2)(\alpha_1+\beta_1+1)}.$$

Again, by compounding and using (3.5), we obtain the unconditional means, given by

$$\begin{aligned} E(X) &= \frac{n\alpha_1}{\alpha_1+\beta_1}, \\ E(Z) &= \frac{n\alpha_1\alpha_2}{(\alpha_1+\beta_1)(\alpha_2+\beta_2)}. \end{aligned}$$

Let $\vartheta = (\alpha_1, \beta_1, \alpha_2, \beta_2)$ be the vector of parameters to be estimated, and consider a sample consisting of $t$ observations $(\tilde{x}, \tilde{z}) = \{(x_1, z_1), \ldots, (x_t, z_t)\}$ drawn from the probability function (5.1). The log-likelihood is proportional to

$$\begin{aligned} \ell(\vartheta; (\tilde{x}, \tilde{z})) \;\propto\; & -t\left(\log B(\alpha_1,\beta_1) + \log B(\alpha_2,\beta_2) + \log\Gamma(\alpha_1+\beta_1+n)\right) \\ & + \sum_{i=1}^{t}\left[\log\Gamma(\beta_1+n-x_i) + \log\Gamma(\alpha_1+x_i) + \log\Gamma(\alpha_2+z_i)\right. \\ & \left. + \log\Gamma(\beta_2+x_i-z_i) - \log\Gamma(\alpha_2+\beta_2+x_i)\right]. \end{aligned} \tag{A.1}$$

Given the log-likelihood function defined in (A.1), it is evident that the global maximum cannot be obtained explicitly. The procedure followed here consists of starting from a seed point and using the `FindMaximum` function of the Mathematica software package v. 12.0 (see, for example, [33]).

This package offers the possibility of using different procedures to find the optimum, such as `Newton`, `PrincipalAxis`, and `QuasiNewton`, guaranteeing in all cases the same parameter estimates.

Finally, the variances of the maximum likelihood estimates can be obtained from the diagonal elements of the inverse matrix of negative second derivatives of the log-likelihood function, evaluated at the maximum likelihood estimates. It is worth noting that the Mathematica package includes the psi function, which is the derivative of the Gamma function, $\psi(z) = d/ds \log\Gamma(s)$, and which is required in this context. In practical situations, it is convenient to approximate the digamma function by the following expression:

$$\log(\Gamma(z)) \approx \frac{1}{2}\log(2\pi) + \left(z - \frac{1}{2}\right)\log(z) - z + \frac{z}{2}\log\left(z\sinh\left(\frac{1}{z}\right)\right),$$

which is well known in the statistical literature and performs well in practice.

## B. Codes in mathematica

### *B.1. Code for getting the maximum likelihood estimators, standard errors and p-values*

```
n = 5;
logl = Sum[
Log[(Binomial[n, x[[j]]] Binomial[x[[j]],
z[[j]]] Gamma[x[[j]]+\[Alpha]1] Gamma[z[[j]]
+\[Alpha]2] Gamma[nn-x[[j]]+\[Beta]1]
Gamma[\[Alpha]1+\[Beta]1] Gamma[x[[j]]-
z[[j]]+\[Beta]2] Gamma[\[Alpha]2+\[Beta]2])/(Gamma[\
\[Alpha]1] Gamma[\[Alpha]2] Gamma[\[Beta]1]
Gamma[nn+\[Alpha]1+\[Beta]1] Gamma[\[Beta]2] Gamma[
x[[j]]+\[Alpha]2+\[Beta]2])], {j,1,338}];
maxlogl =
FindMaximum[
logl, {\[Alpha]1, 2.9}, {\[Beta]1,
1.100084616737736491 99‘},{\[Alpha]2,
1.2970215133710393‘}, {\[Beta]2,1.3791859133560646 88‘},
MaxIterations->20000]
mle = maxlogl[[2]]
parameters = {\[Alpha]1, \[Beta]1, \[Alpha]2, \[Beta]2};
B = Inverse[-D[logl, {parameters, 2}]/.mle];
param = Length[parameters];
erroreparam = Table[N[Sqrt[B[[j,j]]]], {j,1,param}];
t = Table[N[parameters[[j]]/erroreparam[[j]]]/.mle, {j,1,param}];
pvalues =
Table[2*(1-CDF[StudentTDistribution[n-param],Abs[t[[j]]]])/.
mle, {j,1,param}];
AIC = N[2*param-2 logl/.mle];
```

```
BIC = N[-2*(logl/. mle)+param*Log[n]];
CAIC = N[(1 + Log[n])*param-2 logl/.mle];
listaestimate = parameters/.mle;
listase = erroreparam;
listat = t;
listapvalue = pvalues;
listparameters = parameters;
MatrixForm[
Transpose[{listparameters, listaestimate, listase, Abs[listat],
listapvalue}]]
TableForm[
Table[{listparameters, listaestimate, listase, Abs[listat],
listapvalue}, {i,1,1}],
TableHeadings -> {None, {``Parameter", ``Estimate", ``S.E.",
``t-Wald", ``P>|t|"}}, TableSpacing -> {1,3}]
TableForm[Table[{logl/.mle, AIC, BIC, CAIC}, {i,1,1}],
TableHeadings -> {None, {``Log-likelihood", ``AIC", ``BIC", ``CAIC"}},
TableSpacing -> {1,3}]
```

*B.2. Code for getting the marks*

```
Clear[n]; r = pc = 2; s = pu = 1; n = 5; r =
pc = 2;
mle = {\[Alpha]1 -> 1.963,
\[Beta]1 -> 2.102, \[Alpha]2 -> 1.591,
\[Beta]2 -> 1.503};
M[t_, x_,z_] = ((n (\[Alpha]1 + x) (r (\[Alpha]2 + z) -
s (\[Beta]2 + x - z)))/((\[Alpha]1 +
\[Beta]1 + t*n) (\[Alpha]2 +
\[Beta]2 + x))) ((\[Alpha]1 n (r \[Alpha]2 -
s \[Beta]2))/((\[Alpha]1
+ \[Beta]1) (\[Alpha]2 + \[Beta]2)))^-1;
W[t_, x_, z_] = 10/M[1, 5, 5] M[t, x, z];
Table[W[1, 5, j], {j, 0, 4}] /. mle // N
```

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)