Computer Science and Information Systems 22(1):59–78 https://doi.org/10.2298/CSIS240420068K

A New Course Difficulty Index (*CDf*): Framework and Application

Konstantinos Kelesidis, Leonidas Karamitopoulos, and Dimitris A. Dervos

Dept. of Information and Electronic Engineering, International Hellenic University, Sindos, Greece
kelesidis@iee.ihu.gr
karamitopoulos@ihu.gr
dad@ihu.gr

Abstract. A new framework for the quantification of course difficulty in academic curricula is proposed. The originality of the approach lies in its course-centric nature. A course difficulty index value is calculated (*CDf*), using difficulty indicators that characterize the course as a whole. The difficulty indicators can be tailored to reflect the academic domain considered. A weighting percentage is calculated and it is assigned to each course difficulty indicator, by systematically conducting Principal Component Analysis (PCA) on students' assessment data. Next, the weighted difficulty indicators are used to calculate *CDf* in the form of a composite indicator. In general, the value of the latter varies across courses, and across different offerings of a given course. The *CDf* framework is applied in the case of a university in Greece by utilizing course difficulty indicators which are objective in their nature, like course mean and median grades, passing grade percentages, etc. The dataset used spans a period of thirteen (13) academic years. The findings are used to identify courses that represent "bottlenecks" in student study paths. Subjective course difficulty indicators may also be used, like students' questionnaire data. It is worth noting that the quantification of course difficulty by means of a single index can be used in the calculation of adjusted student scores and, as such, facilitate data mining operations on students' assessment data. All in all, the proposed *CDf* framework and analysis comprise a useful tool for academic policy-making and quality assurance.

Keywords: course difficulty, difficulty index, learning analytics, principal component analysis, exploratory data analysis, composite indicator.

1. Introduction and Motivation

The present research is motivated by the need to analyze student assessment data in order to devise effective policies for educational development. At the Information and Electronic Engineering (*IEE*) department of the International Hellenic University (*IHU*) in Greece, the task comprises the responsibility of the department's Internal Evaluation Group (*IEG*). Exploratory data analysis and mining operations are conducted on student assessment records, i.e. grades assigned to students in course modules (*courses*, for brevity). A valuable outcome is the identification of courses that present notable challenges for students, equivalently: courses that students find difficult to pass, or achieve a high grade in. Such information is useful in many aspects, from designing course curricula (especially in the context of the new online learning paradigms, like smart education

[1]), and helping students identify their preferred course of study [33], to rating students with respect to their academic performance.

The way course difficulty perplexes the task of rating students on the basis of their performance in courses is best illustrated by an example in [9]: nine courses C_i (*i*=1,...,9) and four students S_i (*i*=1,...,4) are listed in Table 1 alongside with grades assigned to the latter in courses they have been assessed in. When a student has not been assessed in a specific course, the corresponding (*Course*, *Student*) cell is left blank.

Table 1. Student Assessment Scores Example ([9])

Course	Student				Average Course Grade
	S_1	S_2	S_3	S_4	
$\mathbf{C}_1$	93			90	91.5
$\mathbf{C}_2$	85		80		82.5
$\mathbf{C}_3$			100	95	97.5
$\mathbf{C}_4$	80	75			77.5
$\mathbf{C}_5$		97		95	96.0
$\mathbf{C}_6$	93			85	89.0
$\mathbf{C}_7$		92	89		90.5
$\mathbf{C}_8$	92	91	90	88	90.3
$\mathbf{C}_9$		91	89		90.0
Average Student Grade	88.6	89.2	89.6	90.6	

For the example considered, using *Average Course Grade*, the nine courses are ranked in descending order of difficulty as follows: $C_4 > C_2 > C_6 > C_9 > C_8 > C_7 > C_1 > C_5 > C_3$. Each one of the four students is seen to have been assessed in five of the nine courses. In accordance with their *Average Student Grade* values, the four students are ordered as follows: $S_4 > S_3 > S_2 > S_1$. However, when compared to each other and ranked on the basis of their grades in courses they have both enrolled in, the students are ordered as follows: $S_1 > S_2 > S_3 > S_4$. This is exactly the opposite to their *Average Student Grade* based ordering. Evidently, it does not suffice to evaluate student performance by just considering grades obtained in courses. It is important to take into account the difficulty of each one course a student has been assessed in.

In Greece, the academic year consists of the *Fall-* and the *Spring-* semesters. Course offerings are semester based. Each semester, students enroll in courses and they participate in two (2) final examinations periods: *examinations Period-A*, and *examinations Period-B*. Upon completion of each final examinations period, students are awarded an overall grade for each course in which they have been assessed. Examinations Period-A for the Fall semester courses runs from January to February. Likewise, June is the month for the Spring semester's examinations Period-A. In September one common examinations Period-B is held for all courses offered during the academic year (Fall- and Spring-semesters, alike). Student performance grades lie in the [0,10] range. Five (5.0) comprises the minimum grade requirement for passing a course. In the *CDf* framework that is to be considered next, when a student is assessed twice in a given course during the same academic year, both grades are accounted for.

Course difficulty may vary due to, for example, a new instructor stepping in, when there are significant updates to the course's educational and training content, or when the instructor of a given course attempts to address a backlog of students who haven't passed the exam in previous attempts. In this respect, course difficulty need be measured per course offering. Once quantified, it can be used for calculating adjusted student scores that incorporate the difficulty of the course at the time of assessment.

One may argue at this point that percentile-based standard scores like the $T-$ and $z-$scores comprise the means to rate/rank the performance of a student without sticking to a raw x grade [27]. The research herewith presented aims for extending the standard scores paradigm (a) by considering the relative difficulty of a course with respect to other courses of a given academic curriculum, and (b) by allowing room for the difficulty of a given course to vary from one of its offerings to the next.

The present treatise is organised as follows: the research goal is considered in the context of the relevant bibliography in Section 2 (*Related Work*). Five (5) objectives that pertain to the set aims and goals are outlined in Section 3 (*Research Objectives*). The academic dataset used is outlined in Section 4 (*The Dataset*). Section 5 (*Methodology*) comprises a detailed presentation of the proposed *CDf* framework and its application in the case of the *IEE* department at *IHU*. The results obtained are presented and discussed in Sections 6 (*Results*) and 7 (*Discussion*). The treatise concludes by summing up and identifying new potential research goals in Section 8 (*Conclusion and Future Work*).

2. Related Work

Higher education establishments exploit course difficulty relating information to shape and implement educational policies with the aim to (a) increase success rates in exams [2,24,28,31], (b) minimize the prolongation of the typical student's study period [10], and (c) promote fairness in student performance evaluation and ranking [36]. Fair student ranking has long attracted the interest of researchers because academic performance is used in most high-stakes decisions as in the determination of eligibility for scholarships, or job employment [36].

Academic analytics involves a wide range of methodologies and techniques utilized by higher education institutions, alongside with their quality assurance procedures and strategic policies for educational development [24]. In academic analytics, course difficulty arises as an issue in methodologies aimed at improving student performance evaluation, curriculum design, and course sequencing. Yet another analytical outcome of great interest is the prediction of student performance in courses that lie ahead in their study path [3,21]. The relevant data analysis and data mining tasks are expected to benefit from the adoption and use of a (single) difficulty index that characterizes each course offering. However, there appears to be no single approach on the quantification of course difficulty in the relevant research literature.

One approach to quantifying course difficulty is to use instruments like questionnaires that utilize Likert-type prompts to collect the students' perceptions of a course's level of difficulty [7][25]. To better assess the perceived course difficulty, questionnaires are also used to collect demographic information that can be combined with Grade Point Average (GPA) scores, and course-related details (workload, assignments, etc.).

The authors in [23] investigate the correlation between course difficulty and student stress during course selection scheduling. The determination of course difficulty comprises a milestone in fulfilling their research objective. They propose four methods for the classification of courses with respect to their degree of difficulty: pairwise comparison, an Analytic Hierarchy Process (*AHP*), *GPA*, and physiological measurements.

A number of studies propose methods for improving the *GPA* score in order to provide a fairer measure of a student's academic performance [9][36][37]. In this respect, adjusted *GPA* scores are calculated for university students, scores that correlate better with pre-admission measures like high school *GPA* and *SAT* scores. The flaws of the (unadjusted) *GPA* scores relate to the fact that grading standards and practices tend to vary from instructor to instructor, and from department to department within the university. The proposed methods involve a predictive model of student grades based on parameters such as the ability of a student, the difficulty, and the discrimination of a course.

More specifically, the authors in [9] have investigated several models for adjusting a student's *GPA* score to account for the difficulty of the courses they have been assessed in. In the simplest and perhaps most useful approach, the model predicts a student's grade in a course as the difference between two parameters: the student's ability and the grading standard index that corresponds to the course (i.e. the course's difficulty), plus an error term. The two parameters are estimated for each one student and for each one course, in a way that minimizes the error term. An analogous linear model is proposed by Vanderbei et al. in [37], with the student's intrinsic overall aptitude and the course's inherent difficulty as parameters. The authors calculate the values of the two parameters in the case where each student has been assessed in each and every one course, as well as in the case where students enrol in selected courses. In [36], the authors utilize a two-parameter logistic model that predicts the grade of a student in a course. Their approach involves one student parameter (*ability*) and two course parameters (*difficulty* and *discrimination*). For each student, grades are predicted for all the courses in the academic curriculum, even for those that the student in question has not enrolled in. A modelled *GPA* value is computed based on the predicted grades. The approach is shown to remove the course-choice drawback of the (unadjusted) *GPA* scores.

In addition, several methodologies have been devised within the framework of Item Response Theory (*IRT*) [22][30], generating measures that have been demonstrated to be more reliable than conventional *GPA* scores in capturing student performance [16][38][39]. *IRT* is used extensively in the field of education to assess and calibrate items within tests, questionnaires, and other instruments. It is also used to score individuals based on their abilities, attitudes, or other underlying traits. In this regard, *IRT* models are used to calculate adjusted *GPA* scores, aiming to provide a more accurate estimation of students' performance (ability) in academic courses.

In [38] the author uses an *IRT* model called the *Graded Response Model* (*GRM*) on undergraduate student assessment data [32]. A number of *K* ($K > 2$) ordinal grade categories are assumed to apply. In *GRM*, the probability of a student to achieve a specific grade or higher in one course is expressed as a function of the student's ability, plus the course's difficulty (referred as *grade category boundary*), and discrimination. An important feature of *GRM* is its explicit parameterization of grade category boundaries for each course. In theory, this enables the model to account for variations in instructor grading patterns [16]. The model proposed in [16] can be regarded as a Bayesian extension to

GRM. The primary motivation behind the adjustment is to utilize the relative rankings of students within courses (instead of absolute grades) as a means for evaluating student performance.

3. Research Objectives

All the approaches discussed in Section 2 (*Related Work*) adopt a student-centric approach when assessing course difficulty. More specifically, course difficulty is calculated (a) by considering student responses to questionnaires, or (b) as a model parameter alongside the student's ability and course discrimination, or (c) through a hybrid combination of (a) and (b). This is achieved by assuming that course difficulty remains consistent across the dataset used. To be exact, some approaches allow for course difficulty to vary, say from one course offering to the next, at the cost of increased model complexity, and processing overhead.

The research herewith reported aims to quantify course difficulty from a course-centric (as opposed to student-centric) perspective. In this context, the intended research objectives are set as follows:

1. Course difficulty is to be quantified by means of a single measure (index)
2. The measure will be calculated using a set of course difficulty indicators
3. The mix of the indicators used may vary across different academic environments and systems
4. Course difficulty may vary from one specific course offering to the next
5. Strong emphasis to be given on the visual presentation of the results

Commenting on objective number one, the need for a single measure aims to facilitate further analytical processing such as the calculation of adjusted student scores, and the prediction of students' performance in courses that lie ahead in their study path. In this respect, the calculation of the course difficulty index is seen to comprise a task of the data preparation for data mining stage.

Objective number two relates directly to the course-centric nature of the approach: the measure need be calculated on the basis of parameters that characterize the course as a whole, not on parameters that characterize each individual student assessed.

Objective number three highlights the need for the relevant framework to seamlessly adapt to diverse application domains. Course difficulty indicators may vary significantly in number and/or nature across different academic establishments or systems.

Objective number four is established to account for course difficulty dependence on instructor teaching and/or grading styles, assessment types (e.g., remote testing necessitated by unforeseen circumstances like the COVID-19 lockdown), etc.

Objective number five stresses the need to visualize the results in order to facilitate exploratory analysis and strategic policy planning.

4. The Dataset

As stated in Section 1 (*Introduction and Motivation*) above, in Greece the academic year comprises two semesters. The academic curriculum delineates the courses that are taught during each semester. Students are assessed based on their performance during the semester and in the final examination. There are two final examination periods to each academic semester. Consequently, a student may receive up to two grades for a course during the academic year.

Table 2 outlines the dataset used. It comprised a total of 199,813 grades assigned to 3,737 students in relation to their enrollment in 81 courses at the *IHU IEE* department. The dataset spans a period of thirteen (13) academic years: from 2009-10 to 2021-22. Several elective courses were added to and/or removed from the department's undergraduate academic curriculum over this 13-year period. To maintain consistency, analysis proceeded by focusing on nineteen (19) core STEM courses comprising the *STEM courses subset*. Still, data from all the courses offered over the 13-year period was used in order to better analyze longitudinal trends and student performance patterns for the STEM courses considered.

Table 2. Students' assessment data: 2009-2021

	All courses	**STEM courses subset**
Assessment scores	199,813	123,850
Courses	81	19
Students	3,737	3,721
(*Course, Academic Year*) instances	565	247

5. Methodology

For the undergraduate program of the *IEE* department at *IHU*, the six indicators listed in Table 3 are taken to shape a course's difficulty profile in the academic year.

Table 3. Six (6) course difficulty indicators

Indicator	**Description**
δ_1	Percentage of course grades in the [0,1] range
δ_2	Average course grade
δ_3	Median course grade
δ_4	Average number of attempts a student makes to achieve a passing grade in the course
δ_5	Percentage of passing grades in the course
δ_6	Percentage of active students in the course: enrolled and assessed vs. enrolled

Table 4 lists two (2) measures used per each one indicator: (a) *Course in ac. year*, the value of which is calculated for the (*course, academic year*) pair considered, and (b) *All courses, all ac. years* which is calculated as an average for the given indicator over all courses and all academic years in the dataset. Alongside with the previous two, the table lists two more variables that will be defined next: the indicator's *Bias (b)*, and its normalized difficulty value (*N_Value*) as calculated for the (*course, academic year*) pair in question.

Beginning with the (binary) *Bias (b)*, its value is set to "0" ("1") to indicate the positive(negative) impact the corresponding δ_i represents to course difficulty. A positive bias implies that as the value of δ_i increases, the course in question becomes more difficult. For example, δ_4 represents a positive bias (*b*=0), since the higher the average number of attempts students make to achieve a passing grade, the more difficult the course is perceived to be. On the other hand, δ_2 represents a negative bias (*b*=1), since the higher the average grade students achieve in a course, the less difficult the latter is perceived to be.

$$d_i = (-1)^b \times 100 \times \frac{x - X}{X} \qquad i = 1, \ldots, 6 \tag{1}$$

where

- x is the value of the given indicator's *Course in ac. year* measure
- X is the value of the given indicator's *All courses, all ac. years* measure, and
- b is the indicator's *Bias (b)* value

For the STEM courses in Table 2, an indicator percent variation (d_i) value per (δ_i, *course*, *academic year*) triplet is calculated as follows:

For example, d_5 is calculated as $d_5 = (-1)^1 \times 100 \times \frac{\mu - M}{M}$, μ and *M* being the percentage of passing grade values for a given STEM course in the academic year considered, and that of the average course across the entire dataset, respectively.

Table 4. Course difficulty indicators: measures, plus bias and normalized values

Indicator	**Measures**		**Bias (*b*)**	**N_Value**
	Course in ac. year	**All courses, all ac. years**		
δ_1	α	A	0	Δ_1
δ_2	β	B	1	Δ_1
δ_3	γ	Γ	1	Δ_3
δ_4	ε	E	0	Δ_4
δ_5	μ	M	1	Δ_5
δ_6	ν	N	1	Δ_6

Next, the Δ_i values listed under *N_Value* in Table 4 are calculated by constructing a d_i-standings list for each δ_i. The list registers all (*course*, *academic year*) instances sorted in descending order, based on their d_i values. Considering the positioning of each (*course*, *academic year*) pair in the d_i-standings list, and using the *Rainbow Ranking* equation from [34], Δ_i is calculated to resume values in the (0,100] range:

$$\Delta_i = 100 - 100 \left(\frac{N_{above(c)}}{C} + \frac{N_{tie(c)}}{2C} \right) \tag{2}$$

where

- c refers to a given (*course*, *academic year*) pair, herewith said to comprise a c instance

- C is the total number of c instances in the d_i-standings list
- $N_{above(c)}$ is the number of c instances ranked higher than the given c in the list, and
- $N_{tie(c)}$ is the number of ties (if any, otherwise: 0) c is involved in, not counting c

Along the lines with the research goals outlined in Section 3 (*Research Objectives*), a single course difficulty value for a given course in a specific academic year can be calculated as a linear combination of the Δ_i (*i*=1,...,6) values calculated via Equation 2, provided that the latter are weighted appropriately:

$$CDf = \sum_{i=1}^{6} \Delta_i \times w_i \tag{3}$$

At this point, it is noted that the introduction of Δ_i (i=1,...,6) complies with objective numbers 3 and 4 outlined in *Research Objectives*. Moreover, a single *CDf* value for a course in a specific academic year is calculated via Equation 3, given the corresponding Δ_i (i=1,...,6) values and their w_i (i=1,...,6) weights. The latter is inline with objective numbers 1 and 2 of the *Research Objectives* section. Remaining to be done is the calculation of the w_i (i=1,...,6) weight values. This is achieved by conducting Principal Component Analysis (*PCA*) as described in the following.

Principal Component Analysis (*PCA*) is a multivariate statistical technique commonly used for dimensionality reduction and data simplification. Its input comprises an $n \times p$ matrix where p is the number of measured variables X_i, and n is the number of observations recorded. The initial correlated variables are transformed into a smaller number of uncorrelated variables ($m < p$), called principal components (*PC*s). This is done by preserving as much as possible of the variation (information) present in the original dataset. For a detailed treatise on *PCA* the reader is referred to [17].

Initially, the number of *PC*s of the *PCA* output is equal to the number of variables p present in the original dataset:

$$PC_j = \sum_{i=1}^{p} \alpha_{ij} X_i \qquad j = 1, \ldots, p \tag{4}$$

The difference made by the *PCA* transformation is that the *PC*s are ordered so that the first few retain most of the variance of the initial variable set X_i. More specifically, PC_1 captures most of the variation in all of the initial variables, PC_2 captures most of the remaining variation, and so on. The a_{ij} coefficients are called *loadings* [5].

In the present study the w_i weight values are calculated on the basis of the loadings derived from applying *PCA* on the academic dataset outlined in Section 4 (*The Dataset*). More specifically, *CDf* is calculated as a composite indicator which constitutes a compilation of individual indicators in order to form a single index the value of which quantifies the multidimensional concept as a whole [26]. The use of *PCA* in the construction of composite indicators enjoys applicability in scientific fields ranging from economics [6][14], and environmental engineering and management [11][13], to road safety [12].

Section 6 (*Results*) details the application of the aforementioned methodology in the current study. The determination of the w_i (i=1,...6) weights proceeds in four steps, as follows:

In **step number one** the correlation structure of the dataset is examined to assess its suitability for *PCA*. First, the Bartlett's test of sphericity [4] is applied in order to test whether the correlations in the dataset (as a whole) are strong enough to justify the application of *PCA*. Next, Spearman's correlation coefficient (r_s) is calculated for every pair of Δ_i (*i*=1,...6) variables. This is done in order to ensure that the *PCA* output can be used for calculating reliable weight (w_i) values for Equation 3.

Step number two involves the application of *PCA* on the academic dataset. Considering Equation 4, the Δ_i (*i*=1,...6) variables are used to derive the principal components PC_j (*j*=1,...,6):

$$PC_j = \sum_{i=1}^{6} \alpha_{ij} \Delta_i \qquad j = 1, \ldots, 6 \tag{5}$$

In Equation 5, the loading (α_{ij}) represents the correlation between PC_j and Δ_i. Equivalently, the squared loading (${\alpha_{ij}}^2$) expresses the variance in Δ_i explained by PC_j [6]. Next, the number (*m*) of *PC*s to be retained need be determined. The relevant literature provides a number of guidelines for determining the number of components to retain without experiencing any significant information loss: the Kaiser criterion [18], *Scree plot* [8], and the variance explained criteria [15], to name a few. More than one criteria are typically used in practice and in this respect.

In **step number three**, the retained principal components are rotated to enhance their interpretability. Rotation comprises a transformation to achieve a *simple structure*, namely one where (i) each variable has a high loading on only one of the retained components, and (ii) each retained component represents high loadings for only some of the variables [29]. This way, the most important variables emerge in the rotated principal component: they are the ones with the larger absolute values for their loadings. At the other end, the least important variables emerge with near zero loadings. In the present study, the *varimax rotation* is applied, as it is the one most commonly used [35].

Step number four involves the utilization of the *PCA* output to calculate the w_i (*i*=1,...,6) weights. Denoting *m* the number of retained components from step two, the total variance of Δ_i explained by all *m PC*s is called *communality* (${h_i}^2$) and it is calculated as follows [15]:

$${h_i}^2 = \sum_{j=1}^{m} \alpha_{ij}^2 \qquad i = 1, \ldots, 6 \tag{6}$$

Δ_i's weight (w_i) is calculated as the ratio of Δ_i's communality over the sum of communalities of all Δ_i (*i*=1,...,6) variables:

$$w_i = \frac{h_i^2}{\sum_{i=1}^{6} h_i^2} \qquad i = 1, \ldots, 6 \tag{7}$$

Summarizing, the w_i component in Equation 3 is taken to represent the proportion of Δ_i's variance explained relative to the variance of all Δ_i (*i*=1,...,6) variables explained by the retained components. This is along the lines of the approach reported to have been implemented in the literature, for example [6] and [26].

6. Results

For the dataset used, the correlations among all pairs of the Δ_i (i=1,...,6) variables were calculated and they are presented in Table 5. It is noted that pairs involving variables from the $\{\Delta_2, \Delta_3, \Delta_4, \Delta_5\}$ set exhibited significant correlations, as measured by the Spearman rank correlation coefficient (r_S), the latter varying from 0.34 to 0.87. Notably, the Δ_1 variable is seen exhibit significant correlations with all other variables, with the exception of Δ_4. In addition, the Δ_6 variable correlates significantly only with Δ_1 (r_S =0.23). Bartlett's sphericity test results (χ^2=757.48, df=15, and p <0.001) implied the presence of correlation patterns among the Δ_i (i=1,..,6) variables. In this respect, the results obtained led to the rejection of the null hypothesis, namely that of an identity correlation matrix.

Considering the above, the Δ_i (i=1,...,6) variables turned out to be suitable for subsequent PCA analysis. In addition, it was decided to consider two more configurations by omitting Δ_6 and Δ_1, respectively. Thus, three PCA configurations were implemented: (a) PCA_{1-6} with Δ_i (i=1,...,6), (b) PCA_{1-5} with Δ_i (i=1,..,5), and (c) PCA_{2-6} with Δ_i (i=2,..,6). This was done for two reasons: (a) PCA_{1-5} is a direct analogue to the heuristic approach reported in [19], and (b) in Table 5, Δ_1 is seen to correlate significantly with Δ_2, Δ_3, Δ_5, and Δ_6, while Δ_6 correlates significantly only with the Δ_1 variable.

Table 5. Spearman's rank correlation coefficient (r_s) values for all (Δ_i, Δ_j) (i, j=1,..,6)

	Δ_1	Δ_2	Δ_3	Δ_4	Δ_5	Δ_6
Δ_1	-	r_s=0.34 (p<0.01)	r_s=0.30 (p<0.01)	r_s=0.07 (p=0.30)	r_s=0.22 (p=0.01)	r_s=0.23 (p<0.01)
Δ_2	-	-	r_s=0.81 (p<0.01)	r_s=0.40 (p<0.01)	r_s=0.87 (p<0.01)	r_s=0.12 (p=0.06)
Δ_3	-	-	-	r_s=0.34 (p<0.01)	r_s=0.70 (p<0.01)	r_s=0.11 (p=0.10)
Δ_4	-	-	-	-	r_s=0.48 (p<0.01)	r_s=0.06 (p=0.37)
Δ_5	-	-	-	-	-	r_s=0.05 (p=0.42)
Δ_6	-	-	-	-	-	-

Beginning with PCA_{1-6}, Table 6 lists the eigenvalues and the corresponding explained variances for all six principal components (PCs) in the PCA outcome. In accordance with Kaiser's criterion, the first two components (PC_1 and PC_2 in Table 6) need be retained since their eigenvalues are greater than 1. The *Scree plot* in Figure 1 is indicative of the notably sharp drop from the first eigenvalue to the second, the rate of decrease remaining small thereafter. For PCA_{1-6}, the findings suggest that analysis may safely proceed by considering only PC_1 and PC_2 which together account for 69.52% of the variance present in the original dataset (the *Cumulative Variance Explained* column in Table 6).

Table 6. Eigenvalues and variance explained for the six PCs in PCA_{1-6}

Component	**Eigenvalue**	**Variance Explained (%)**	**Cumulative Variance Explained (%)**
PC_1	3.05	50.89	50.89
PC_2	1.12	18.63	69.52
PC_3	0.82	13.60	83.12
PC_4	0.65	10.75	93.87
PC_5	0.28	4.66	98.53
PC_6	0.09	1.47	100.00
Total	6.00	100.00	

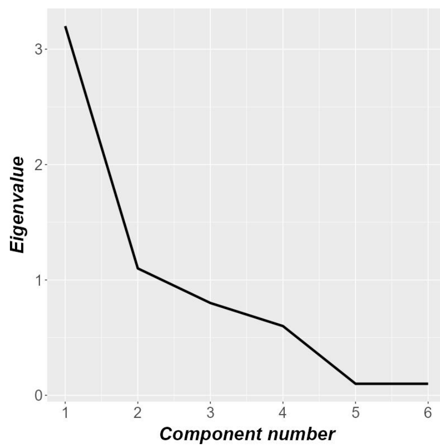

Fig. 1. Eigenvalues of the six principal components (PCs) in PCA_{1-6} (*Scree plot*)

Table 7 presents the PCA_{1-6} findings in the form of loadings associated to the given Δ_i (*i*=1,...,6) variables when the first two principal components are used, alongside with the corresponding communalities and w_i (*i*=1,...6) weights. The *varimax rotation* method was applied on the original PCA_{1-6} outcome. The communality (h^2) for each Δ_i (*i*=1,...,6) is the sum of the squared loadings of the two retained components (cf. Equation 6) For each Δ_i, the corresponding w_i value is calculated by Equation 7. The percentage of cumulative variance explained (information) of the two principal components (PC_1 and PC_2) is also listed under the PCA_{1-6} label.

Table 8 is analogous to Table 7 and summarizes the findings for PCA_{1-5} and PCA_{2-6}. Again, the first two principal components are used. One notes the increase in the cumulative variance explained values: 79.16% and 77.75% for PCA_{1-5} and PCA_{2-6}, respectively, next to 69.52% in the case of PCA_{1-6}. An observation that applies to all PCA outcomes in Tables 7 and 8, is that Δ_2, Δ_3, Δ_4 and Δ_5 are seen to maximize their (information) contribution values in PC_1, whereas Δ_1 and Δ_6 maximize theirs in PC_2 (column: *Loadings*).

Table 7. PC_1 and PC_2 loadings, communalities, and weights for PCA$_{1-6}$

PCA$_{1-6}$ (*cumulative variance explained: 69.52%*)				
Variables	**Loadings**		**Communalities**	**Weights**
	PC$_1$	**PC$_2$**	**(h^2)**	**(w_i)**
Δ_1	0.27	0.70	0.56	0.13
Δ_2	0.92	0.22	0.90	0.22
Δ_3	0.85	0.23	0.77	0.18
Δ_4	0.63	-0.08	0.41	0.10
Δ_5	0.92	0.08	0.84	0.20
Δ_6	-0.08	0.83	0.69	0.17
Total			4.17	1.00

Table 8. PC_1 and PC_2 loadings, communalities, and weights for *PCA*$_{1-5}$ and *PCA*$_{2-6}$

	***PCA*$_{1-5}$** (*cumulative variance explained*: 79.16%)				***PCA*$_{2-6}$** (*cumulative variance explained*: 77.75%)			
Variables	**Loadings**		**Communalities**	**Weights**	**Loadings**		**Communalities**	**Weights**
	PC$_1$	**PC$_2$**	**(h^2)**	**(w_i)**	**PC$_1$**	**PC$_2$**	**(h^2)**	**(w_i)**
Δ_1	0.12	0.90	0.83	0.21				
Δ_2	0.87	0.39	0.90	0.23	0.94	0.07	0.90	0.23
Δ_3	0.79	0.39	0.77	0.20	0.87	0.08	0.77	0.20
Δ_4	0.74	-0.26	0.61	0.16	0.61	0.00	0.37	0.10
Δ_5	0.89	0.21	0.84	0.21	0.92	0.00	0.85	0.22
Δ_6					0.04	1.00	1.00	0.26
Total			3.96	1.00			3.89	1.00

Table 9 summarizes on Equation 3's w_i (*i*=1,...,6) weights as calculated using Equation 7 for PCA_{1-6}, PCA_{2-6}, and PCA_{1-5}, and set heuristically in [19]. PCA_{1-6} and PCA_{2-6} are seen to exhibit similar weighting patterns for Δ_i (*i* =1,...,5). Comparing PCA_{1-5} to PCA_{1-6}, the weight w_1 for Δ_1 increases from 0.13 to 0.21, incorporating the majority of the contribution from the Δ_6 variable (0.17), which is present in PCA_{1-6} but not in PCA_{1-5}. Analogously for PCA_{2-6} and PCA_{1-6}, the weight w_6 for Δ_6 increases from 0.17 to 0.26, incorporating the majority of the contribution from the Δ_1 variable (0.13), which is present in PCA_{1-6} but not in PCA_{2-6}. In all three PCA configurations, the w_4 value for Δ_4 is notably smaller from the rest of the weight values. Consequently, the δ_4 indicator (*Average number of attempts a student makes to achieve a passing grade*) emerges to possess the smallest relative impact on course difficulty. Regarding the weights heuristically assigned to the Δ_i (*i* =1,...,5) variables in [19], there are notable deviations in w_1, w_4, and w_5 compared to those calculated for the three PCA configurations.

Table 9. PCA_{1-6}, PCA_{2-6}, PCA_{1-5} and $Heuristic$ weight values

Variables	**Weights w_i** (i=1,...,6)			
	***PCA*$_{1-6}$**	***PCA*$_{2-6}$**	***PCA*$_{1-5}$**	***Heuristic***
Δ_1	0.13		0.21	0.05
Δ_2	0.22	0.23	0.23	0.20
Δ_3	0.18	0.20	0.20	0.20
Δ_4	0.10	0.10	0.16	0.05
Δ_5	0.20	0.22	0.21	0.50
Δ_6	0.17	0.26		
Total	1	1	1	1

7. Discussion

As stated previously, PCA_{1-5} and *Heuristic* in Table 9 both involve the same set of course difficulty indicators. Their difference lies in the way the w_i (i =1,...,5) weights are determined. In the *Heuristic* configuration, weight values are set heuristically, whereas in PCA_{1-5} weight values are calculated by conducting *PCA* and using Equation 7. The heuristic approach is seen to deviate significantly from PCA_{1-5} in three of the five weights assigned to the Δ_i (i=1,...,5) variables. More specifically: (a) Δ_1 (*percentage of grades in the [0,1] range for the given course in the academic year considered*) is rated to be nearly four times as important in PCA_{1-5} (w_1=0.21) as in the heuristic approach (w_1=0.05), (b) Δ_4 (*the average number of attempts a student makes to achieve a passing grade in the course*) turns out to be nearly three times more important in PCA_{1-5} (w_4=0.16) compared to what it is set heuristically (w_4=0.05), and Δ_5 (*percentage of passing grades for the course in the academic year considered*) is nearly half as important in PCA_{1-5} (w_5=0.21) as in the heuristic approach (w_5=0.5).

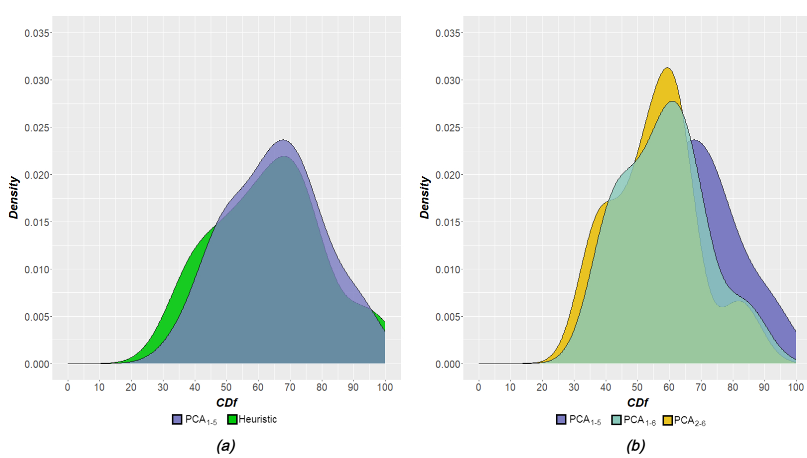

Fig. 2. *CDf* density plots (19 core STEM courses, all years)

Figure 2 presents the *CDf* distribution (density) curves for the nineteen (19) core STEM courses over the 13-year period: PCA_{1-5} and *Heuristic* in Fig.2*a*, PCA_{1-5}, PCA_{1-6}, and PCA_{2-6} in Fig.2*b*. All are left-skewed, with PCA_{1-6} and PCA_{2-6} seen to involve a smaller degree of left-skewness as compared to PCA_{1-5} and *Heuristic*. PCA_{1-5} is seen to lie notably closer to *Heuristic* relative to both PCA_{1-6} and PCA_{2-6}. It is also narrower than the *Heuristic*, i.e. it involves a smaller variance. PCA_{1-6} and PCA_{2-6} appear to be slightly shifted towards smaller *CDf* scores, relatively to the other two. This could be taken to indicate that the δ_6 indicator (*percentage of enrolled students who have been assessed*) represents a lesser impact on course difficulty compared to the δ_1 indicator (*percentage of course grades in the [0,1] range*). This is further supported by the fact that PCA_{1-5} (where δ_6 is not used) represents the highest cumulative variance explained (79.16%) compared to PCA_{2-6} (77.75%) and PCA_{1-6} (69.5%), as shown in Tables 7 and 8. Given these findings, the remainder of this section will exclusively focus on the PCA_{1-5} configuration.

Using PCA_{1-5}, Figure 3 presents two *CDf* distribution curves: one for all courses (the 19 core STEM courses included) and one for the 19 core STEM courses, both over the 13 academic years period considered. The former is significantly more uniform compared to the latter. The STEM courses curve exhibits a notable degree of left-skewness, plus it is clearly shifted towards higher *CDf* scores. This is taken to mean that students tend to face more challenges with the 19 core STEM courses, compared to the other (81-19=62, mostly: elective) courses in the undergraduate curriculum.

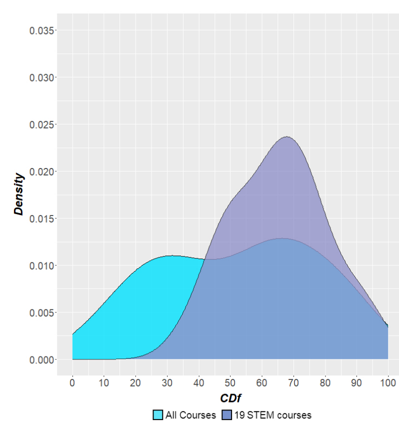

Fig. 3. *CDf* density plots (all years): 19 core STEM courses vs. all courses

The proposed *CDf* framework can effectively identify specific courses that present increased challenges for students. Such courses act as "bottlenecks" that hinder student progress, and prolong the study period. For example, focusing on course C_1, its *CDf* value is marked as a dot on the STEM courses' *CDf* distribution curve in Figure 4, during the 2021-22 academic year. The C_1 dot is seen to be positioned well past the curve's dominant inflexion point. Thus, for the academic year considered (2021-22), with a *CDf* score close to 85, C_1 is undoubtedly ranked among the most challenging core STEM courses.

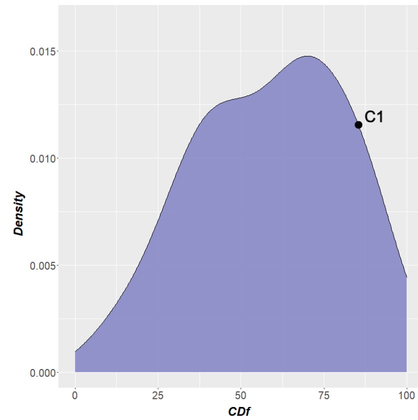

Fig. 4. Course C_1 on the 19-STEM courses' *CDf* curve during the 2021-22 academic year

To obtain a feeling of how C_1's *CDf* has varied during the 2017-2022 period, Figure 5 combines in one graph all five *CDf* distribution curves of the 19 core STEM courses for the academic years considered. It is noted that C_1 has shown a steady increase in difficulty, rising from 52.74 in 2017-18 to 85.36 in 2021-22.

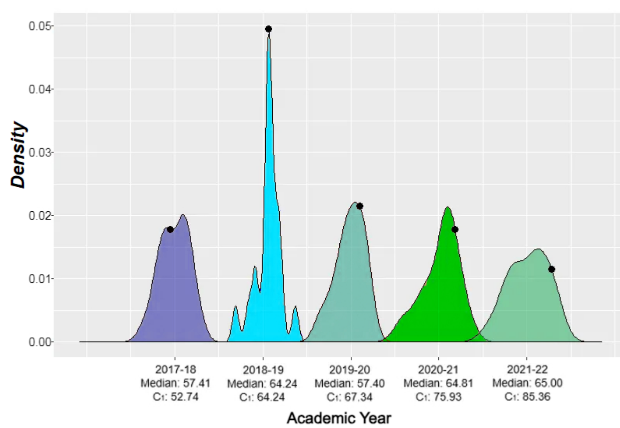

Fig. 5. Course C_1 on the 19-STEM courses' *CDf* curves during the 2017-22 period

Better yet, Figure 6 encapsulates in one graph all the information on the variation of *CDf* scores for two courses (C_1: red, and C_2: green dots on the graph) over the entire 13-year period. With one box plot per academic year the C_1 and C_2 *CDf* scores positioned relative to the former, the reader has a complete picture of how the two courses' difficulty has varied over the 13-year period. Moreover, the graph reveals information on how the difficulty of the 19 core STEM courses (as a group) has varied from one academic year

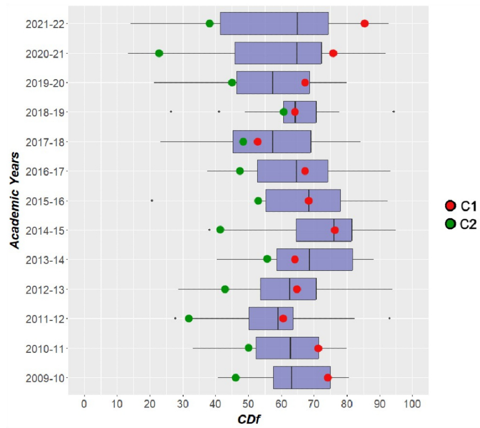

Fig. 6. 13-academic years period: Courses C_1 and C_2 positioned on all 19 core STEM courses' *CDf* box plots

to the next. For instance, by examining their median *CDf* scores, the core STEM courses are seen to have posed the greatest challenge to undergraduate students during the 2014-15 academic year. Also, 2018-19 has been the academic year with the largest number of outliers among the 19 core STEM courses: two (2) in the low (less difficult), and one (1) in the high (most difficult) ends of the corresponding box plot.

Focusing on C_1 in Figure 6, the course's *CDf* score is seen to lie in the right whisker region of the box plot only during the last two academic years (2020-21 and 2021-22). Prior to 2020-21, C_1 consistently remained below or on the Q3 boundary of the STEM courses' box plot. More precisely, C_1 comprised an outlier (in terms of its difficulty) STEM course during the 2020-21 and 2021-22 academic years. This may be indicative of C_1's tendency to gradually evolve into a "bottleneck" course for students. In this context, Figure 6 signals an alert to course instructors and the department regarding the increasing challenge associated with C_1. In response to this, possible actions may include (a) a review of the student assessment procedures, and (b) the identification of one or more other courses in the curriculum that could be designated to comprise prerequisites for C_1. On the other hand, and with the exception of the 2017-18 academic year, course C_2 is seen to always comprise an outlier in the left (less difficult) whisker region of the corresponding box-plots in Figure 6. This too deserves the department's attention.

As stated already, the results herewith presented and commented upon relate to the case of a typical university in Greece. In accordance with research objective number three in Section 3 (*Research Objectives*), the mix of the course difficulty indicators used may vary across different academic environments and systems. Indicators like expected study hours, number of course prerequisites, and independent study requirements also comprise potential candidates to be used as difficulty indicators. U.K. and U.S. universities closely monitor student dropout rates and class sizes for course offerings; these too tend to relate to course difficulty. Last but not least, student feedback and course evaluation survey outcomes can be used in the form of one or more course difficulty indicators, provided that student response rates are consistently high, as in the case of the National Student Survey (N.S.S.) scores in the U.K.

8. Conclusion and Future Work

A new framework for the quantification of course difficulty in academic curricula is proposed. A course-centric index (*CDf*) is calculated using difficulty indicators per course offering. Each indicator is assigned a weight which is determined systematically by conducting Principal Component Analysis (PCA) on student and/or course assessment data. The course difficulty indicators may be either objective (e.g. the percentage of assessed students who achieved a passing grade, the course's mean and median scores, etc.), or subjective (e.g. student questionnaires data). The approach differs from its predecessors in that (a) it is course-centric (instead of student-centric) in nature, (b) the difficulty of a course is assumed to vary from one of its offerings to the next, and (c) the course difficulty indicators can be defined flexibly both in number and type, in order to seamlessly adapt to diverse application domains (universities, schools, and departments).

The proposed *CDf* framework and methodology have been applied in the case of the *IEE* department of the International Hellenic University (*IHU*), where each course is offered once per academic year, in either the Fall or Spring semester. A maximum of six course difficulty indicators were used, all objective in nature. Three PCA configurations were considered, using students' assessment data from nineteen (19) core STEM courses over thirteen (13) academic years. For each (*course*, *academic year*) pair, a single *CDf* score was calculated as a linear combination of the (normalized) course difficulty indicator values, using a matching set of (PCA calculated) weights. As a result, *CDf* density distribution curves were constructed. In relation with research objective number five in Section 3: *Research Objectives*, which dictates that strong emphasis be given on the visual presentation of the results obtained, each course's *CDf* score was positioned on the density distribution curve and on the box plot of all courses in each academic year. This type of graphical output facilitates exploratory analysis and has already proven valuable for the department's Internal Evaluation Group (*IEG*) in identifying potential "bottleneck" courses that may prolong a typical student's study period.

In the future stages of the research, the plan is to:

1. evaluate the proposed *CDf* framework in data mining operations that predict a student's future performance on the basis of their past assessment records [20],
2. consider additional course difficulty indicators (both objective and subjective) and their impact on the efficacy of the *PCA/CDf* output, and
3. develop and offer *CDf* as a prototype open source web service to be used by academic units in Greece and abroad.

Potential improvements of the proposed framework include, for example: (a) the monitoring of an instructor's difficulty profile in cases where the same course is measured to involve a notably different *CDf* value when it is taught by more than one instructors, or when the same instructor teaches both core STEM and elective courses (the former tending to be more challenging to the students, as suggested by Figure 3), and (b) the inclusion of subjective indicators based on data from student feedback and course evaluation surveys with consistently high response rates. The use of such subjective difficulty indicators could unveil hidden insights often missed by objective indicators.

Acknowledgments. The authors thank the anonymous reviewers for their comments and suggestions which have significantly enhanced the readability and integrity of this report.

References

1. Al-Turjman, F., Ivanovic, M.: Guest editorial: Interactive and innovative technologies for smart education. Computer Science and Information Systems 19, vii–x (2022)
2. Arnold, K.E., Pistilli, M.D.: Course signals at purdue: Using learning analytics to increase student success. In: Proceedings of the 2nd International Conference on Learning Analytics and Knowledge. p. 267–270. Association for Computing Machinery, New York, NY, USA (2012)
3. Baepler, P., Murdoch, C.: Academic analytics and data mining in higher education. International Journal for the Scholarship of Teaching and Learning 4(2) (2010)
4. Bartlett, M.S.: A note on the multiplying factors for various $\chi2$ approximations. Journal of the Royal Statistical Society. Series B (Methodological) 16(2), 296–298 (1954)
5. Berlage, L., Terweduwe, D.: The classification of countries by cluster and by factor analysis. World Development 16(12), 1527–1545 (1988)
6. Boylaud, O., Nicoletti, G., Scarpetta, S.: Summary indicators of product market regulation with an extension to employment protection legislation. SSRN Electronic Journal (2000)
7. Bradley, G.: Factors that influence course difficulty. Doctoral dissertation, Iowa State University (1992), `https://lib.dr.iastate.edu/rtd/9820`
8. Cattell, R.B.: The scree test for the number of factors. Multivariate Behavioral Research 1(2), 245–276 (1966)
9. Caulkins, J., Larkey, P., Wei, J.: Adjusting gpa to reflect course difficulty (1996)
10. Chalaris, M., Gritzalis, S., Maragoudakis, M., Sgouropoulou, C., Lykeridou, K.: Examining students' graduation issues using data mining techniques - The case of TEI of Athens. AIP Conference Proceedings 1644(1), 255–262 (2015)
11. Gine, R., Pérez-Foguet, A.: Improved method to calculate a water poverty index at local scale. Journal of Environmental Engineering 136, 1287–1298 (2010)
12. Gitelman, V., Doveh, E., Hakkert, S.: Designing a composite indicator for road safety. Safety Science 48, 1212–1224 (2010)
13. Gomez-Limon, J., Riesgo, L.: A composite indicator to measure agricultural sustainability: Alternative approaches. Decision Support Systems in Agriculture, Food and the Environment: Trends, Applications and Advances (2008)
14. Gomez-Limon, J., Sanchez-Fernandez, G.: Empirical evaluation of agricultural sustainability using composite indicators. Ecological Economics 69(5), 1062–1075 (2010)
15. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E.: Multivariate Data Analysis. Pearson Education Limited, Upper Saddle River, NJ, 7th edition edn. (2013)
16. Johnson, V.E.: An alternative to traditional GPA for evaluating student performance. Statistical Science 12(4), 251–269 (1997)
17. Jolliffe, I.T.: Principal Component Analysis. Springer Science & Business Media, New York, 2nd edition edn. (2002)
18. Kaiser, H.F.: The application of electronic computers to factor analysis. Educational and Psychological Measurement 20(1), 141–151 (1960)
19. Kelesidis, K., Dervos, D., Sidiropoulos, A.: Quantifying the Difficulty of Academic Course Modules. In: 24th Pan-Hellenic Conference on Informatics. pp. 245–249. ACM, Athens Greece (2020)
20. Kelesidis, K., Fotopoulou, N., Dervos, D.: Correlation as an arm interestingness measure for numeric datasets. In: 27th Pan-Hellenic Conference on Informatics. pp. 1–7. ACM, Athens Greece (2023)

21. Komenda, M., Víta, M., Vaitsis, C., Schwarz, D., Pokorná, A., Zary, N., Dušek, L.: Curriculum mapping with academic analytics in medical and healthcare education. PLOS ONE 10(12), 1–18 (2015)
22. Lord, F., Novick, M.: Statistical Theories of Mental Test Scores. Addison-Wesley series in behavioral sciences: Quantitative methods, Information Age Publishing, Incorporated (2008)
23. Ma, S., Akgunduz, A., Zeng, Y.: Course scheduling according to student stress. Proceedings of the Canadian Engineering Education Association (2015)
24. Mat, U.b., Buniyamin, N., Arsad, P.M., Kassim, R.: An overview of using academic analytics to predict and improve students' achievement: A proposed proactive intelligent intervention. 2013 IEEE 5th Conference on Engineering Education (ICEED) pp. 126–130 (2013)
25. Mundfrom, D.: Estimating course difficulty. Doctoral dissertation, Iowa State University (1991), https://doi.org/10.31274/RTD-180813-11387
26. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E.: Handbook on Constructing Composite Indicators and User Guide. OECD Publishing (2008)
27. Neukrug, E., Fawcett, R.C.: The essentials of testing and assessment: a practical guide for counselors, social workers, and psychologists. Cengage Learning, Stanford, CT, 3rd edn edn. (2015)
28. Palmer, S.: Modelling engineering student academic performance using academic analytics. International Journal of Engineering Education 29, 132–138 (2013)
29. Pedhazur, E., Schmelkin, L.: Measurement, Design, and Analysis: An Integrated Approach. Lawrence Erlbaum Associates (1991)
30. Rasch, G.: Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche, Danmarks Paedagogiske Institut, Copenhagen (1960)
31. Romero, C., Ventura, S.: Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10(3) (2020)
32. Samejima, F.: Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement 34(4) (1969)
33. Savic, M., Ivanovic, M., Luković, I., Delibašić, B., Protic, J., Janković, D.: Students' preferences in selection of computer science and informatics studies: A comprehensive empirical case study. Computer Science and Information Systems 18, 54–54 (2020)
34. Sidiropoulos, A., Stoupas, G., Katsaros, D., Manolopoulos, Y.: The rainbow over the greek departments of computer science/engineering: a bibliometric study. In: 21st Pan-Hellenic Conference on Informatics. pp. 1–6. ACM (2017)
35. Tabachnick, B.G., Fidell, L.S.: Using multivariate statistics. Pearson, New York, 7 edn. (2019)
36. Tomkin, J.H., West, M., Herman, G.L.: An improved grade point average, with applications to CS undergraduate education analytics. ACM Transactions on Computing Education 18(4) (2018)
37. Vanderbei, R.J., Scharf, G., Marlow, D.: A regression approach to fairer grading. SIAM Review 56(2), 337–352 (2014)
38. Young, J.W.: Adjusting the cumulative GPA using item response theory. Journal of Educational Measurement 27(2), 175–186 (1990)
39. Young, J.W.: Gender bias in predicting college academic performance: A new approach using item response theory. Journal of Educational Measurement 28(1), 37–47 (1991)

Konstantinos Kelesidis is a Ph.D. candidate with the Department of Information and Electronic Engineering at the International Hellenic University. His research focuses on academic student assessment data analytics. He holds a B.Sc. degree (2009) from the Department of Information Technology of the Alexander Technological Educational Institute of Thessaloniki, and an M.Sc. degree in Web Intelligence (2020) from the Department of Information and Electronic Engineering at the International Hellenic University

in Thessaloniki. Mr. Kelesidis is a highly experienced software developer, working in the software engineering domain since 2011.

Leonidas Karamitopoulos is a lecturer (Laboratory Teaching Staff) with the Department of Information and Electronic Engineering at the International Hellenic University. He holds a B.A. degree in Mathematics from the Aristotle University of Thessaloniki, a M.Sc. degree in Operations Research from the George Mason University, Virginia, and a Ph.D. degree in Time Series Data Mining from the University of Macedonia. His current main research interests lie in the areas of time series data mining and recommendation systems.

Dimitris A. Dervos is Professor Emeritus of Information and Electronic Engineering at the International Hellenic University (I.H.U.) in Thessaloniki, Greece. He holds a B.Sc. in Physics (1979) and a Ph.D. in Computer Science (2000) from the Aristotle University of Thessaloniki, as well as an M.Sc. in Computer Engineering (1982) from the University of Southern California. He has taught Database Technologies and Predictive Data Analytics in Greece and the U.K. A founding member of the Database Technologies Network (DBTechNet), he has lectured on database technologies in Finland, Germany, and Spain. He is a recipient of the 2015 Thomson ISI/ASIS&T Citation Analysis Research Grant. Since retiring in 2023, he supervises Ph.D. candidates and lectures on Predictive Data Analytics at I.H.U. and the University of Macedonia.

Received: April 20, 2024; Accepted: September 20, 2024.