AI-Enhanced Exam Generator Program: A Case Study in Live University Exam Settings

LÁNG Blanka, KOVÁCS László, DÖMSÖDI Balázs

Abstract. Creating exams is time-consuming for educators, and despite existing tools, no solution has been universally adopted. This study evaluates EGAL+, a hybrid artificial intelligence and metaheuristics-based exam generation tool, in real university exam settings. Students were randomly assigned to traditional or EGAL+-generated exams. Student performance and exam quality were assessed using objective metrics, and qualitative feedback from teachers and students. Results show that EGAL+ significantly reduces exam preparation time without harming student performance, while improving exam quality through better alignment with teachers' preferences, greater question diversity, and more consistent difficulty. These findings indicate that EGAL+ reduces teacher workload while maintaining or enhancing exam quality, with no observed drawbacks.

Keywords: automated exam generation, AI, faculty satisfaction, multi-objective optimization, harmony search algorithm, empirical analysis

1. Introduction

Preparing exams is labor-intensive and time-consuming for educators; consequently, several software solutions have been developed to streamline the process, allowing teachers the opportunity to focus on tasks that make better use of their expertise. Despite these efforts, no such solution has been accepted universally. Therefore, this study evaluated the effectiveness of a hybrid tool based on artificial intelligence (AI) and metaheuristics, EGAL+ (Exam Generator Algorithm+), in real university examination settings to assess its potential for wider adoption. The study aimed to demonstrate and validate the unique benefits of EGAL+ in a live exam environment.

The authors of this study previously researched this topic in another work [1], aiming to generate different exams and practice tests for students in a class to assess their knowledge. It was recognized that some topics were more important than others, and the exams and practice tests should be as varied as possible to avoid memorization in favor of quality learning. In our research, the quality of an exam or practice test was determined by how well the grouping of questions aligned with teachers' preferences and by the diversity of questions across different groups. Additionally, certain questions would be undesirable to ask together for pedagogical or logical reasons. The manual compilation of the question sets was considered to require a significant amount of time and effort, and teachers had to pay attention to the equal difficulty of the question sets and the diversity between them. Over the years, the topics we examined have varied, but ultimately, the focus remained on exam sets composed of test questions under similar conditions.

EGAL+ is a unique application in the domain of exercise generation, enabling instructors to define, through a preference matrix, which exercises they would like included in their task sets and at what pairwise priority. The completed preference matrix allows the teacher to almost instantaneously generate numerous sets of tasks of the same difficulty from the uploaded database, each set covering the discipline as specified by the teacher in the matrix (i.e., coexistence preferences).

The literature review revealed that no automatic solution had been developed before to address such a case. Therefore, a heuristic solution was applied since an exact algorithm could not solve the problem within a reasonable timeframe. Hence, a harmony search metaheuristic with a double objective function was created, as it maximizes the coexistence preferences and diversity measures simultaneously to solve the problem, yielding satisfactory results almost instantly with a quick run.

Later, a generative AI model was integrated into the application, enabling the question bank to be uploaded quickly.

Thus, instructors could specify in a preference matrix which tasks they would like to include in the task sets with their joint preference. Notably, the sets of questions were diverse and of an overall difficulty set by the teacher and. With all this, this paper aimed to demonstrate and validate the unique advantages of EGAL+ in a university exam.

2. Literature Review

2.1. Overview

The use of metaheuristics and other algorithms in the field of automatic test or exam generation has been the subject of much research. The goal of these solutions is to automatically generate test sets in a fast and efficient manner according to different objectives and criteria, such as difficulty level, diversity, and quality. In this chapter, we review current automatic test set generation solutions and their characteristics and critically analyze their methodological differences, strengths, and limitations in relation to our proposed EGAL+ algorithm.

All of the following articles conclude that producing test question sets is a difficult and time-consuming practice and, therefore, well worth automating. In their solutions, the authors usually formulate several optimization objectives simultaneously, and the fitness functions they use are multi-objective. All authors set quality maximization as a goal, but they differ in what they consider quality. When defining quality, most articles use Bloom's taxonomy [2]. Moreover, they strive for diversity and often attempt to maximize the distance between test sets to obtain different test sets for students. Many of them also aim to generate question sets of varying difficulty levels for groups of students with different levels of knowledge while attempting to give students in the same group question sets of the same difficulty. Further, they attempt to cover as much of the subject area as possible according to suitably formulated objectives. Due to the high computational demand of the problems, metaheuristics are employed (primarily genetic algorithms [GAs]). However, other metaheuristics are also used, such as ant colony optimization and simulated annealing. Several studies also addressed the efficient loading of question banks in addition to the generation of test sets.

Despite these shared goals, the reviewed approaches differ significantly in how these objectives are modeled, prioritized, and optimized, particularly in terms of flexibility, adaptability, and granularity of control.

2.2. Review of Related Works

2.2.1. Genetic Algorithm-Based Approaches

A significant portion of the literature applies genetic algorithms to the problem of automatic test generation, modeling it as a multi-objective optimization task in which candidate test sets evolve toward predefined criteria.

[3] proposed a GA-based auto-generator of examination questions designed to streamline the creation of examination papers that align with outcome-based education (OBE) specifications. The challenge of manually creating high-quality exam questions, especially for novice lecturers, is

addressed through automation. The objective function minimizes the error between the user-defined exam paper criteria and the generated output. Then, the system validates the conformity of the generated questions with OBE criteria, ensuring assessments align with learning outcomes. Compared to EGAL+, this approach similarly relies on optimization but focuses on matching predefined criteria rather than enabling flexible, fine-grained instructor control.

[4] proposed a solution that reduces the manual effort and time educators spend preparing exam papers while ensuring quality and diversity in question sets. Bloom's taxonomy is used to categorize questions by difficulty levels (easy, medium, and hard), ensuring that exams test different cognitive domains. This automated approach ensures efficient, reliable exam paper generation, thus reducing educators' workload while maintaining assessment standards. In contrast, EGAL+ also considers difficulty and diversity, but extends beyond taxonomy-based categorization by introducing pairwise preference modeling.

[5] explored the use of a GA for the automatic generation of assessment tests in higher education, ensuring fairness and differentiation between test versions. This approach addresses challenges related to increasing class sizes, individualization, and the need to minimize cheating by generating structurally diverse tests. A structural metric, DTest, measures the distance between two tests based on question content and answer choices, inspired by the Levenshtein distance. The GA optimizes test differentiation by iterating through processes like selection, crossover, and mutation to maximize the structural distance between tests. While EGAL+ also maximizes diversity between tests, it complements this objective with coexistence preferences, which are not considered in this approach.

[6] presented an automated exam question generator that uses a GA to assist educators in efficiently creating high-quality exam papers. The system focuses on producing multiple-choice questions (MCQs) that adhere to the six levels of Bloom's taxonomy, ensuring balanced cognitive difficulty. The generator uses a question bank containing categorized past exam questions, tagged according to Bloom's taxonomy and specific chapters. The fitness function evaluates the quality of generated question sets based on the distribution across Bloom's taxonomy levels. Compared to EGAL+, this method emphasizes taxonomy balance, whereas EGAL+ introduces an additional dimension of instructor-defined relationships between questions.

[7] introduced an HGA for automated test paper generation. The system ensures compliance with multiple constraints, including question difficulty, discrimination, knowledge coverage, and format requirements. The fitness function prioritizes minimizing deviation from constraints while maximizing test quality, ensuring that generated test papers align closely with user-defined requirements. Although this approach incorporates multiple constraints similarly to EGAL+, it does not allow expressing preferences at the level of individual question pairs.

[8] discussed the use of a GA for generating exam tests from a categorized question bank. The fitness function measures how well the test aligns with the selected categories, ensuring relevance to the intended topics. The method is particularly suited for large question banks, where manual selection would be too time-consuming. In comparison, EGAL+ also supports similar large-scale optimization but provides additional flexibility through its preference matrix.

2.2.2. Hybrid and Adaptive Metaheuristic Approaches

A number of studies extend classical metaheuristic frameworks by incorporating hybridization strategies and adaptive mechanisms, aiming to enhance solution quality, convergence speed, and responsiveness to dynamic educational requirements. These approaches typically combine genetic

algorithms with complementary techniques—such as learning agents, predictive models, or domain-specific heuristics—to overcome the limitations of standalone methods. By introducing adaptivity, these systems can refine their behavior based on historical data, student performance, or evolving constraints, thereby enabling more intelligent and context-aware test generation.

[9] proposed an auto-generating examination paper algorithm based on a hybrid genetic algorithm (HGA) to address inefficiencies and subjectivity in traditional examination paper creation systems in China. The paper concluded that the proposed HGA-based method improves the efficiency and quality of auto-generated test papers, offering a reliable solution for intelligent teaching systems. While EGAL+ also aims to improve efficiency and quality, it differs by prioritizing customization through explicit preference modeling.

[10] introduced an automated exam question set generator (AEQSG) that uses utility-based agents (UBA), learning agents (LA), GA, and Bloom's taxonomy scaling to streamline the creation of exam question sets. This system aims to reduce educators' workload while ensuring high-quality exams aligned with institutional guidelines. UBA selects actions that maximize utility based on user preferences, ensuring that generated question sets meet desired criteria such as difficulty and content distribution; LA improves the system by learning from past exam results, adapting automatically to generate better question sets over time; and Bloom's taxonomy scaling automates the distribution of questions across six cognitive levels. This approach significantly reduces the manual effort involved in exam preparation, thereby improving the consistency and quality of exam papers. Compared to EGAL+, this approach introduces adaptivity through learning, whereas EGAL+ focuses on explicit and interpretable instructor-defined preferences.

[11] proposed a novel approach for automated exam paper generation (EPG) based on predicting the performance of student groups using deep knowledge tracing (DKT), dynamic programming, and GAs. This method aims to produce exam papers with adaptive difficulty levels without manually labeling question difficulty. The quality of generated papers is evaluated based on difficulty, skill distribution, and the distribution of predicted student scores, offering a reliable way to automate exam paper generation while maintaining fairness, accuracy, and efficiency. In contrast, EGAL+ does not rely on predictive models but instead offers deterministic control through its preference matrix.

2.2.3. Other Approaches

Beyond metaheuristic-based solutions, the literature also includes a variety of alternative approaches to automatic test generation, ranging from rule-based systems and database-driven methods to machine learning techniques. These approaches often prioritize simplicity, scalability, and ease of implementation, focusing on efficient question selection and coverage rather than complex optimization. While they may not always achieve the same level of flexibility or solution optimality as metaheuristic methods, they provide practical and accessible solutions for many real-world educational settings.

[12] surveyed various question paper generation systems and highlighted techniques that automate and optimize the creation of exam papers. These systems address the time-consuming and error-prone nature of manual question paper generation. Automated systems utilize databases and algorithms to randomly select questions, ensuring coverage of the entire syllabus and multiple question types (e.g., multiple-choice, numerical, and theory-based questions). The proposed system employs CSV files for keyword extraction from syllabi, which is more efficient than manually

populating question banks. Compared to EGAL+, these approaches provide simpler automation but lack advanced optimization and fine-grained control.

[13] utilized supervised machine learning algorithms to efficiently generate exam questions tailored to candidate skills. The system allows customization of test parameters, such as question selection, format, and difficulty level, to suit specific needs. The use of machine learning classifiers improves the accuracy of skill-based assessments and helps recruiters understand candidate strengths. Compared to EGAL+, this approach relies on learned models and historical data, whereas EGAL+ operates independently of such data and instead provides direct control to instructors through preference specification.

[14] conducted a comparative study of various algorithms and methodologies used for automated question paper generation to improve efficiency, security, and reliability over traditional manual methods. The study examined automatic generation of question papers based on the following criteria: simplicity, speed, constraint handling, and the ability to control question difficulty levels. It provides a detailed analysis of the strengths and weaknesses of each method with respect to these aspects. The study concluded that selecting the appropriate algorithm depends on factors like efficiency requirements, database size, and difficulty control.

2.3. Critical Comparison with EGAL+

The studies reviewed here reached converging conclusions, stating that automating the generation of questionnaires improves the efficiency, fairness, and reliability of educational assessments. Automated systems are designed to simplify the process, reduce human error, allow teachers to focus on teaching rather than administrative tasks, and reduce the manual effort and time spent by teachers on preparing examination papers while ensuring the quality and diversity of the question sets.

Through the reviewed literature, three primary optimization objectives can be identified:

- difficulty control (often taxonomy-based),
- content coverage,
- diversity (distance between test sets)

While these systems often introduce adaptivity to achieve these goals through student modeling or learning agents, they largely overlook instructor-driven customization at a granular level. Specifically, no reviewed method allows the definition of pairwise preferences between individual questions.

In several aspects, our EGAL+ algorithm is similar to the solutions described above. Metaheuristics were applied to achieve the goal, specifically the harmony search algorithm. Moreover, we aimed for diversity, maximizing the distance between tests to obtain different test sets for students. We also sought to generate varying levels of difficulty for groups of students with different levels of knowledge, while students in the same group were given questions of the same difficulty. Our fitness function is also multi-objective: while one goal is to maximize diversity, the other is to maximize coexistence preferences. The only significant difference compared to the algorithms above is the latter point, which is formulated using a preference matrix that contains the instructor's preferences. The essence and significance of the preference matrix are explained in the next chapter. Our solution is unique in that no other solution allows the user to parameterize a preference matrix pairwise on a task level, allowing for full customization over task grouping.

3. Methodology

3.1. Research Questions

The purpose of this study was to demonstrate and validate the unique advantages of EGAL+ in the context of university exams. To achieve these objectives, we posed the following research questions:

1) How do students perform on tests generated with EGAL+ (T2) compared to traditional, manually created tests (T1)? Specifically, is there a significant difference in student performance between T1 and T2 exams?

2) To what extent does the algorithm reduce instructors' workload?

3) Does T2 provide greater diversity and better compliance with user preferences compared to T1?

3.2. Design and Functionality of EGAL+

EGAL+ is a sophisticated program designed to automate the generation of exams for educational assessments, ensuring that exams are diverse, align with user-defined grouping criteria, and maintain balanced difficulty levels across task sequences to promote fairness. The EGAL+ source code is publicly accessible at https://github.com/balazs-domsodi/EGALPP.

The foundation of EGAL+ lies in its implementation of the harmony search metaheuristic algorithm [15]. Inspired by the improvisational techniques of musicians striving for harmony, this algorithm efficiently explores complex solution spaces, identifying near-optimal results.

EGAL+ operates by assembling task sequences from a user-provided question bank. Each task is characterized by a difficulty value (rated on a 1–5 scale) and coexistence preferences for other tasks (rated on a 0–10 scale, where 0 prohibits inclusion and higher values indicate stronger preferences). The user also specifies the desired number of tasks per sequence, the number of sequences to generate, and the target difficulty level (low, medium, or high).

During the generation process, the program determines the three target difficulty levels by evaluating the sum of task difficulties in sequences that adhere to coexistence preferences. Tasks are selected iteratively, with their indexes stored as they are added one at a time to the sequences. This step-by-step approach avoids the inefficiencies associated with generating entire sequences simultaneously, significantly reducing the likelihood of invalid combinations. The compact storage of task indexes minimizes memory usage and enhances the speed of sequence processing.

The generation process seamlessly integrates the determination of difficulty levels and initial sequence creation. As potential difficulty levels are analyzed, valid task sequences are simultaneously generated and classified according to their difficulty. This integrated approach eliminates redundant operations, enabling the program to begin quality enhancement immediately after the user selects a difficulty level.

The quality of a task sequence is assessed based on two factors: its distinctiveness compared to other sequences in the population and its alignment with user-defined coexistence preferences. Sequences that include tasks with higher coexistence preference values and differ more substantially from others are considered of higher quality.

The coexistence preferences are represented by a matrix, denoted as CP. Let ex_length represent the exercise length and p denote a given population element. Using these notations, the coexistence preference value for a specific task is computed from the CP matrix as defined by formula (1).

$$F_{coexistence}(p) = \sum_{i=0}^{ex\_length-1} \sum_{j=i+1}^{ex\_length} CP[p[j]][p[i]] \quad (1)$$

Let h, k be two individuals, and the relative diversity measure between these individuals be the number of different tasks they contain, denoted as d(h, k).

Building on this, the difference value for a given individual p, compared to all other elements t in a population (excluding the element it is intended to replace), sums the distances for each individual in the population. It is denoted as $F_{diversity}(p)$.

Formula (2) represents the sum of two components: one based on preference values and the other on difference values, which, together, define the quality measure for each population element.

$$F_{fitness}(p) = F_{coexistence}(p) + F_{diversity}(p) \quad (2)$$

The fitness function combines two operand values using a single operation, which could be addition, subtraction, multiplication, or division. In this study, addition was selected as the operation because the two operands are considered independent and equally important in the overall fitness. The ratio of the two operands is a subjective decision made by the teacher, similar to how the preference matrix is populated. Like the preference matrix, this ratio is not determined objectively. For this study, the two operands were simply added, as the instructor deemed that this approach represented the correct ratio. In future developments, we plan to enable instructors to customize the ratio of the two operands.

Once the initial population is generated, the program refines the sequences using an iterative optimization process based on the harmony search algorithm. In each iteration, a new task sequence is created, an existing sequence is modified, or a sequence is copied without modification, ensuring that, in all cases, the new sequence remains valid. The process evaluates whether the newly created sequence offers better quality than the lowest-quality sequence in the current population. If so, it replaces the inferior sequence.

To create or modify sequences, the program employs probabilistic methods. New sequences are either generated from scratch or derived by copying and adjusting existing sequences by replacing a task with another from the question bank while maintaining adherence to coexistence preferences and difficulty targets. If no suitable replacements are found within a given time, the sequence remains unchanged.

The program continually evaluates the population's average quality. If the average quality fails to improve by a predefined threshold over a series of iterations, the optimization process halts. Similarly, the process concludes if a generation limit is reached. The result is a final population of task sequences that meet all criteria and maximize quality.

EGAL+ is capable of handling extensive question banks with thousands of tasks and generating high-quality task sequences for hundreds of students within seconds. By iteratively selecting tasks, optimizing storage, and integrating processes, the program achieves exceptional efficiency without imposing limitations on input size or user-defined constraints. This capability makes EGAL+ invaluable for large-scale educational settings.

The primary challenge of using EGAL+ for efficient exam compilation is the initial effort required to create a comprehensive question bank and define preferences, which can be time-intensive for

instructors. This requirement renders the program particularly well-suited for large-scale educational environments, such as online learning platforms and undergraduate courses, where foundational concepts in extensive subject areas are frequently examined. In such settings, a well-constructed question bank can be reused over an extended period, enabling streamlined and consistent exam generation.

3.3. Using EGAL+ for Exam Generation

Exams generated by the program are formatted in the GIFT format, which can then be imported into Moodle, a widely adopted learning management system.

To generate an exam, the program must be started, and the third mode ("3. Generate an exam") must be selected. The name of a question bank file is required, which must be placed in the "databank" folder of the program.

In the question bank, each line must begin with the text of the question itself. This is followed by a tab and an integer on a scale of 1–5, indicating the difficulty level of the question. After another tab, the coexistence preferences for joint inclusion in exams are provided. These preferences consist of values of 0–10 separated by semicolons, where each number represents the preference relative to the preceding questions in order. For the first question, a single placeholder value is used, as there are no preceding questions. Finally, after another tab, optional answer options can be included, separated by tabs. The number of questions per exam and the total number of exams to generate are then specified.

The user interface of EGAL+ can be seen in Figure 1.

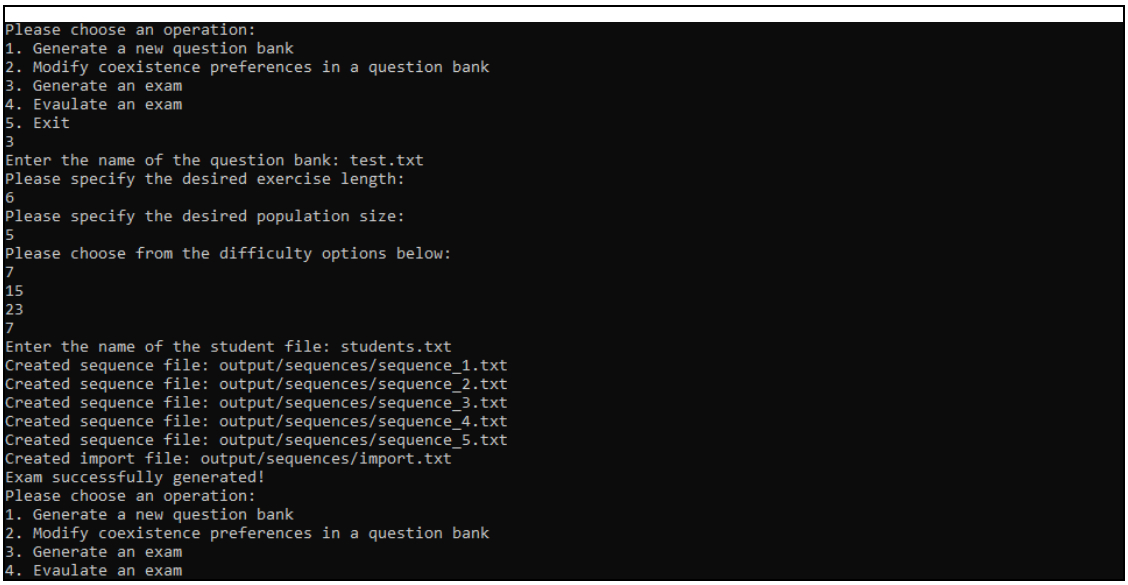

Figure 1: The user interface of EGAL+

Three achievable difficulty levels for the exams are determined by the program. Once a desired difficulty level is chosen, an optional student file is requested, which should include the names or identifiers of the students for whom the exams will be generated in separate lines. The information in the student file is used to create a Moodle import file, ensuring that uniquely identifiable exams are assigned to each student. If a student's file is not provided, the exams will be generated into separate files in a similar format.

After all inputs are provided, the exams are generated. All coexistence preferences are adhered to, restrictions on specific questions are enforced where necessary, the selected difficulty level is maintained, and varied exams are created to meet the specified criteria. The contents of a resulting GIFT file can be observed in Figure 2.

```
$CATEGORY: $course$/top/Student1

Task1 {
    =Task1CorrectAnswer
    ~Task1InorrectAnswer1
    ~Task1InorrectAnswer2
}

Task2 {
    =Task2CorrectAnswer
    ~Task2InorrectAnswer1
    ~Task2InorrectAnswer2
    ~Task2InorrectAnswer3
}

Task7 {
    =Task7CorrectAnswer
    ~Task7InorrectAnswer1
}
```

Figure 2: The contents of a resulting GIFT file

The resulting file can then be imported into a Moodle system.

3.4. Research Context and Research Design

The experiment was conducted at Corvinus University of Budapest during the fall semester of 2024, specifically in November and December, when quarterly exams were held. Students used the Moodle platform for the exams. The tests were administered to four groups of students across four classrooms at different time slots. Altogether, 107 BSc students enrolled in the fifth semester of a seven-semester business information systems program participated.

A single year cohort of students enrolled in the same business information systems program may be treated as a cluster sample from the population of students enrolled in the program in the same institution. Cluster sampling can be considered appropriate when naturally occurring groups (clusters) exhibit internal heterogeneity while remaining comparable across clusters. The validity of using this cohort as a representative sample of the broader student population rests on the assumption of structural equivalence across cohorts.

Specifically, if admission criteria, selection procedures, and program requirements remain stable over time, then each cohort is generated through the same underlying selection mechanism. This consistency implies that each yearly intake approximates a realization from the same population distribution. Consequently, differences between cohorts are expected to be random rather than systematic. Hence statistical methods and tests can be applied to draw conclusions on the population of unobserved students as well.

However, it is important to acknowledge limitations. External factors such as demographic trends, policy changes, or shifts in labor market attractiveness of the program could introduce cohort

effects, potentially undermining representativeness. Furthermore, this cluster sampling method does not make the results generalizable to students of other programs and other institutions. Therefore, the authors aim to tackle this limitation in their future work, extending the test of EGAL+ to students of several programs in multiple universities.

The exams included in the research were theoretical tests from the business intelligence lectures. The first quarterly exam was worth a total of 20 points, while the second was worth 30 points. Both exams featured an initial section comprising 10 MCQs, each worth one point, with four answer options per question, of which exactly one was correct. Students had 10 minutes to answer these 10 questions. To pass the exam, students needed to answer at least six questions correctly; otherwise, the exam resulted in a failing grade. These 10 MCQs covered the fundamental concepts of the course, for which students received a predefined list in advance for their preparation.

Following the multiple-choice section, students tackled essay questions. In the first quarterly exam, there were two essay questions, each worth five points, to be answered within 20 minutes. These questions aimed to test the students' understanding of the deeper relationships within the course material. In the second quarterly exam, students answered four essay questions, each worth five points, within a 30-minute timeframe.

In Moodle, the instructor prepared the following question banks:

- Multiple-choice questions for the first quarterly exam (138 questions);
- Essay questions for the first quarterly exam (46 questions);
- MCQs for the second quarterly exam (83 questions);
- Essay questions for the second quarterly exam (64 questions).

These question banks were partially compiled from questions from previous exams and partially generated using the PrepAI question-generation tool from a PDF version of the textbook assigned as mandatory reading for the course. Questions generated by PrepAI were manually reviewed to filter out overly simple, ambiguous, repetitive, or irrelevant ones.

The preference matrix for the two quarterly exams was filled separately by the instructor to align with the thematic areas of the course. The material covered in the first quarter encompasses the role of business intelligence in enterprises, its technological solutions, and analytical tools, focusing on the following topics:

- Decision support;
- Data analysis environments in enterprises;
- Data warehouses;
- Data types;
- Data quality issues;
- Data standards;
- Online analytical processing; dimensions and fact data;
- Data visualization;
- Big data management.

Within these topics, the co-occurrence of questions is generally not preferred (preference value of 0). However, when a topic is sufficiently broad, slight preferences (values of 1–3) may be permitted. For instance, the topic of data analysis environments in enterprises is quite broad, encompassing various analytical tools applied across corporate functions (hence the preference value of 3 within

the topic). Conversely, questions on dimensions and fact data categorization involve only two definitions addressed in different contexts (hence, a preference value of 0 within the topic).

Between different topics, preferences are generally fully allowed (a preference value of 10), with exceptions in specific cases. For example, data quality issues are often managed using data standards, and addressing these issues is essential during data warehouse construction. Therefore, there are overlaps in the concepts across these topics, so their joint occurrence is preferred less. Data warehouses are part of the enterprise data analysis environment, so these topics are also less preferred together. The data type of variables determines how these variables and their relationships are visualized, which connects to the logic of classifying dimensions and fact data. Consequently, these three topics are also not maximally preferred together. Table 1 presents the preference matrix, reflecting these principles.

	Data analysis environments	**Data standards**	**Data quality issues**	**Data types**	**Data visualization**	**Big data**	**Online analytical processing**	**Decision support**	**Data warehouses**
Data analysis environments	3	10	10	10	10	10	10	10	5
Data standards	10	2	4	10	10	10	10	10	3
Data quality issues	10	4	1	10	10	10	10	10	3
Data types	10	10	10	1	4	10	6	10	10
Data visualization	10	10	10	4	1	10	8	10	10
Big data	10	10	10	10	10	0	10	10	10
Online analytical processing	10	10	10	6	8	10	0	10	10
Decision support	10	10	10	10	10	10	10	2	10
Data warehouses	5	3	3	10	10	10	10	10	1

Table 1: Table preference matrix for the exam in the first quarter

The material for the second-quarter exam covers selected chapters on data mining and machine learning:

- Cross-Industry Standard Process for Data Mining (CRISP) data mining process
- General framework of machine learning
- Supervised machine learning
- Unsupervised machine learning
- Natural language processing

The determination of preference values within the topics follows the same principles as for the first-quarter exam, based on the broadness of the topic. For example, questions within the CRISP topic exclusively focus on the various steps of the CRISP-DM process, conceptually constituting a single element; therefore, the within-topic co-occurrence preference is set to 0. Conversely, the topic of natural language processing spans a wide range of areas, from simple tokenization to sentiment analysis and topic modeling, allowing for a higher within-topic preference value of 4. The preference matrix, designed based on these principles, is presented in Table 2.

	CRISP	**Supervised machine learning**	**General machine learning**	**Unsupervised machine learning**	**Natural language processing**
CRISP	0	10	10	10	10
Supervised machine learning	10	1	5	4	8
General machine learning	10	5	3	5	10
Unsupervised machine learning	10	4	5	1	8
Natural language processing	10	8	10	8	4

Table 2: Table preference matrix for the second-quarter exam

This business intelligence course was chosen as the subject of our study because we required material that changes only minimally over time, given that populating the preference matrix is a time-intensive task. The benefits of using the EGAL+ software are most evident when this work needs to be done only once instead of being repeated each semester. The theoretical content of the two quarterly exams focuses on the general concepts and mathematical principles of business intelligence and data analysis rather than the specific workings and software implementations of algorithms. The latter areas are assessed during the course's practical exam. The general principles of data analysis and machine learning are considered temporally stable. Even incorporating large language models into the curriculum (under the natural language processing topic) required no drastic revisions to the theoretical content, as these models can be treated as supervised learning algorithms conceptually [16]. Additionally, this course is mandatory and taken by a large number of students every year, allowing the instructor to benefit from the software's advantages over several years.

Of the 107 students who took the exams, 53 were randomly assigned to receive traditionally prepared exams. In these cases, the instructor manually designed the exams to reflect their preferences. The remaining 54 students were given test sets generated using EGAL+. For these students, each received a unique test set, as generating individualized sets using the tool required

no additional effort from the instructor. Students were not informed before the exam about how their tasks were generated. At the end of both quarterly exams, immediately after completing their exercises, students rated the difficulty of the multiple-choice and essay sections of the theoretical exam on a scale from 1 to 10, unaware of whether their test sets were EGAL+ generated or traditionally prepared. The difficulty of the practical exams was also rated on a scale from 1 to 10 by the students for reference.

The traditional and EGAL+-generated exams were identical in all aspects except for the generation method:

- Both were created from the same question bank;
- The exams were designed to have equivalent overall difficulty, with individual questions rated on a 1–5 difficulty scale. Summing up these difficulty ratings by individual exams yielded the same result for every student, regardless of whether they received traditional or EGAL+-generated exams. The total difficulties were as follows:
    - First-quarter multiple-choice questions: 28;
    - First-quarter essay questions: 7;
    - Second-quarter MCQs: 28;
    - Second-quarter essay questions: 14;
- Students were also allotted the same amount of time in all cases:
    - First-quarter MCQs: 10 minutes;
    - First-quarter essay questions: 20 minutes;
    - Second-quarter MCQs: 10 minutes;
    - Second-quarter essay questions: 30 minutes;

The sole difference was that, in the second case, the EGAL+ software was used to select questions based on the preference matrix, while in the first case, the instructor manually selected questions, which are as follows:

- MCQs were organized into 10 separate sub-question banks based on difficulty and topic. Students received one randomly selected question from each sub-bank in both multiple-choice sections;
- Essay questions were organized into two sub-banks for the first-quarter exam and four sub-banks for the second-quarter exam, also based on difficulty and topic. Students received one randomly selected question from the appropriate sub-bank for each essay section;

Grading was conducted identically for both groups. The Moodle system automatically evaluated the MCQs by checking if the students' responses matched the pre-determined correct answers. For the essay questions, instructors utilized a standardized grading guide provided to all grading instructors. An example of this guide is provided in the appendices. Notably, the uniqueness of the exams effectively eliminated the possibility of academic dishonesty through the online sharing of answers among students.

3.5. Data Collection

In the research context defined in Subsection 3.3., the following variables were collected in three separate datasets on the students who took the exams. In all the variables, the following indices are applied:

- i: student id = $\{1,2,\dots,107\}$;
- j: exam number = $\{1st\ quarter, 2nd\ quarter\} = \{1,2\}$;

- k: exam type = $\{EGAL+, Manual\} = \{E, M\}$;

First, the objective qualities of the individual student exams were evaluated using the components of formula (2):

- CMC_{ijk}: coexistence value of the multiple-choice exam for student i of type k in quarter j;
- CE_{ijk}: coexistence value of the essay exam for student i of type k in quarter j;
- DMC_{ijk}: diversity value of the multiple-choice exam for student i of type k in quarter j;
- DE_{ijk}: diversity value of the essay exam for student i of type k in quarter j;
- FMC_{ijk}: fitness value of the multiple-choice exam for student i of type k in quarter j;
- FE_{ijk}: fitness value of the essay exam for student i of type k in quarter j;

The second dataset contains the subjective student perceptions of the exams, which was assessed by an anonymous student survey where students were asked to judge the difficulty of each exam item on a scale of 1–5, where 1 means very easy and 5 means very difficult. This survey data cannot be connected to any of our data from other sources as they are completely anonymous. Therefore, the k index cannot be defined for these survey variables.

- PMC_{ij}: difficulty perception from 1 to 5 of the multiple-choice exam for student i in quarter j;
- PE_{ij}: difficulty perception from 1 to 5 of the essay exam for student i in quarter j;
- PP_{ij}: difficulty perception from 1 to 5 of the practical exam for student i in quarter j;

For the survey, 45 responses were received in quarter 1 and 39 in quarter 2. Thus, we have that $n_1 = 45$ and $n_2 = 39$.

To assess how the exam generation types (EGAL+ or Manual) affected the student scores, we obtained a third dataset containing the student scores of each exam, the binary variable indicating whether they took an EGAL+ or a Manual exam, and two control variables for individual student's abilities. Note that we do not need to define the k index for these variables as the exam type is expressed by a separate binary variable in this dataset.

- SMC_{ij}: performance score of the multiple-choice exam for student i in quarter j
- SE_{ij}: performance score of the essay exam for student i in quarter j
- $ST_{ij} = SMC_{ij} + SE_{ij}$: total performance score of the exam for student i in quarter j
- $Type_{ij}$: 1 if student i took an EGAL exam in quarter j, 0 if they took a manual exam
- $Retake_{ij}$: 1 if student i has failed the course in a previous semester, 0 otherwise; the value is the same $\forall j$
- $Time_{ij}$: in a two-week timeframe before the exam, how many times did student i access the course materials on Moodle in quarter j

The control variable $Time_{ij}$ had the limitation that it could not register if more students were preparing for the exam together using the same device. Therefore, this variable likely underestimated the preparation time of some students. However, as course material download was prohibited on Moodle, the variable was not distorted by students preparing offline using their downloaded materials.

In addition to tabular data collection, the business intelligence course leader was interviewed regarding the workload associated with manual and EGAL+-supported exam generation. The course leader has held this position since 2021 and has supervised the exam generation process for

four academic years (2021/22, 2022/23, 2023/24, and 2024/25). They maintain a question bank consisting of 138 MCQs and 46 essay questions for the exam in the first quarter, as well as 83 MCQs and 38 essay questions for the exam in the second quarter. During the academic years from 2021/22 to 2023/24, the course leader manually created sub-question banks from these questions according to the principles outlined in section 3.4. Since these exams are administered during seminars, students take the exams in four different time slots; consequently, the course leader needed to create different versions of these sub-question banks for each seminar to minimize redundancy arising from randomly selected questions. According to the course leader, preparing the exams takes an average of 2.5 hours each year. However, by generating individual exams separately for each student with the support of EGAL+, this time is reduced to approximately 15 minutes. Furthermore, the course leader noted that controlling the sub-question banks for coexistence, diversity, and difficulty is a manual and somewhat subjective process, resulting in exams that may be less effective in these respects compared to those generated by EGAL+, where these aspects are controlled directly and objectively through the algorithm's fitness function.

3.6. Statistical Methodology

First, the objective qualities of the individual student exams were compared through the homogeneity of their distributions. We expected that the qualities of the EGAL+ and manual exams were not significantly different. If they were, we anticipated that the EGAL+ type would show a significantly larger proportion of exams with higher quality in coexistence, diversity, and overall fitness. The homogeneity of the exam quality measure distributions among the EGAL+ and manual types was tested using the two-sample Kolmogorov–Smirnov (KS) test, as proposed by [17], [18], and [19]. The KS test statistic was applied to each element of the variable set $X = \{CMC, CE, DMC, DE, FMC, FE\}$ as follows:

$$D = \sup_x \left| F_{XEj}(u) - F_{XMj}(u) \right|$$

where F_{XEj} and F_{XMj} are empirical distribution functions of a given variable in the set X for the E (EGAL+) and M (Manual) types, respectively, in quarter j. The observed samples had sizes of $n_E = 55$ and $n_M = 52$ respectively for the EGAL+ (E) and manual (M) types. The sample sizes were adequate to achieve sufficient statistical power for the tests [20]. The test statistic and the p-values based on the Kolmogorov distribution for each comparison were obtained via the ks.test function in the R language, using the algorithm proposed by [21] and applying numerical improvements as suggested by [18].

Subjective student preconceptions of the exams were compared by examining whether the distribution of ratings for the multiple-choice and essay exams was independent of the ratings for the practical exams. Since the practical exams did not utilize EGAL+ in any way, it could be assumed that if the subjective ratings of the practical and the other two exam types show a strong level of association, then student perceptions are not influenced by EGAL+, as they rated exams with and without EGAL+ similarly.

The level of association between the two variables in set $Z = \{PMC, PE\}$ and PP was measured by the χ^2 statistic and Cramér's V as proposed by [22] and [23]. The following formulas were calculated for each variable in Z separately:

$$\chi^2_{Z,PP} = \sum_{l=1}^{5}\sum_{m=1}^{5}\frac{f^2_{Z_l,PP_m}}{f_{Z_l}\cdot f_{PP_m}} - 1 \quad (3)$$

$$V_{Z,PP} = \sqrt{\frac{\chi^2_{Z,PP}}{n\cdot \min\{\max(l)-1\,,\max(m)-1\}}} \quad (4)$$

In the formulas, f_{Z_l,PP_m} is the joint frequency of observations that take value l in the variable currently examined from set Z and simultaneously take the value m in the PP, while f_{Z_l}, and f_{PP_m} are the marginal frequencies for the values l and m in variables $Z = \{PMC, PE\}$and PP, respectively. As variables PMC, PE, PP are all measured on a scale from l to 5, both l and m range from 1 to 5. Therefore, the $\min\{\max(l)-1\,,\max(m)-1\}$ part in the $V_{Z,PP}$ formula provided by [23] simplifies to $5-1$, and n denotes the total number of observations. As the $V_{Z,PP}$ measures are calculated for both quarters 1 and 2, we have $n_1 = 45$ and $n_2 = 39$, respectively.

The effect of the exam generation types (EGAL+ or manual) on student scores was modeled using ordinary least squares (OLS) regression. Each variable in the set $Y = \{SMC, SE, ST\}$ was applied as a target variable in a multivariate regression with $Type, Reatake, Time$ being the feature variables of the models. Our regression model is defined as follows.

$$Y_{ij} = c_j + \alpha_j Type_{ij} + \beta_j Retake_{ij} + \gamma_j Time_{ij} + e_{ij} \quad (5)$$

With this model, we can capture the effects of the $Type, Reatake, Time$ variables on the three different exam scores separately for each quarter j. The main variable of interest is $Type$, and it was assumed that its α_j coefficients are not significantly different from 0 in any of the fitted regressions, as this would indicate that the exam generation type has no statistically significant effect on the exam scores of students. The role of the $Reatake, Time$ variables are to control for individual student abilities and efforts taken while preparing for the exams. Moreover, it was assumed that students who needed to retake the course would have significantly lower scores, while students who spent more time preparing for the exams were expected to achieve significantly higher scores. These phenomena should not distort the effect of the $Type$ variable on exam scores due to the random classification of students into EGAL+ and manual groups. However, including these variables in our models eliminated any possible confounding effects from the α_j coefficients caused by different individual student characteristics [24].

The e_{ij} residual term of the models was assumed to be normally distributed, homoscedastic, and serially uncorrelated. Homoscedasticity of the residuals was tested using the White and Breusch–Pagan tests, while serial correlation was examined via the Breusch–Godfrey test, as suggested by [25]. Further, the normality of the residuals was checked using the Jarque–Bera test [26]. Testing whether our linear model specification is adequate was done using Ramsey's RESET test, as proposed by [27] and [25].

4. Results

First, the homogeneity of the coexistence, diversity, and overall fitness score distributions was examined between the manual and EGAL+ groups using the KS test. Table 3 presents the test statistics and p-values derived from the Kolmogorov distribution.

Quality Type	Exam	D Statistic	p-value	Significance
Fitness	Quarter 1 - Essay	1.00000	0.00000	Yes on all levels
Diversity	Quarter 1 - Essay	1.00000	0.00000	Yes on all levels
Coexistence	Quarter 1 - Essay	0.13131	0.78692	No on all levels
Fitness	Quarter 1 - Multiple-choice	1.00000	0.00000	Yes on all levels
Diversity	Quarter 1 - Multiple-choice	1.00000	0.00000	Yes on all levels
Coexistence	Quarter 1 - Multiple-choice	0.30553	0.01430	Yes on 5%, but not on 1%
Fitness	Quarter 2 - Essay	0.98000	0.00000	Yes on all levels
Diversity	Quarter 2 - Essay	0.98000	0.00000	Yes on all levels
Coexistence	Quarter 2 - Essay	0.61091	0.00000	Yes on all levels
Fitness	Quarter 2 - Multiple-choice	0.78431	0.00000	Yes on all levels
Diversity	Quarter 2 - Multiple-choice	0.74510	0.00000	Yes on all levels
Coexistence	Quarter 2 - Multiple-choice	0.51408	0.00000	Yes on all levels

Table 3: Table preference matrix for the second-quarter exam

Our results indicate that the null hypothesis of distribution homogeneity can be rejected at all common significance levels for nearly every exam, except for the coexistence levels of the first quarter's essay exams. In this latter case, we can state that at all levels there are no significant differences in the coexistence distribution of the manual and EGAL+ exams. However, for the first quarter's multiple-choice exam, we cannot make a straightforward decision regarding significant differences in the coexistence distribution between the manual and EGAL+ exams, as the decision depends on the significance level.

Figure 3 examines the empirical density functions of these objective exam quality measures. We found that in every case where these distributions are significantly different at all levels, the difference favors the EGAL+ exams. The EGAL+ values are concentrated at the higher end of the distribution compared to the manual exam values.

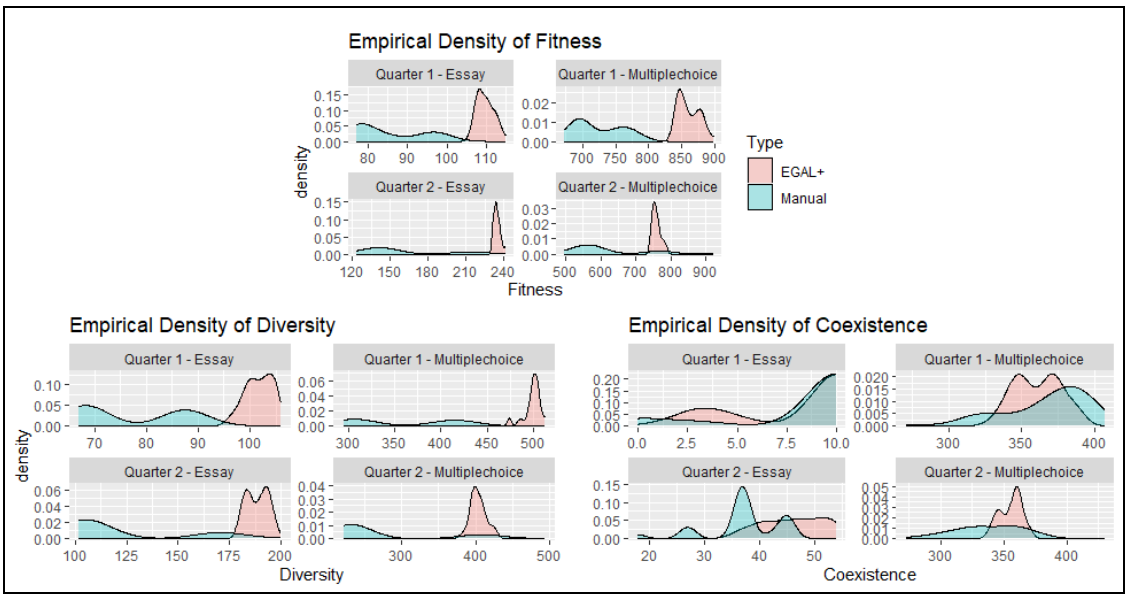

Figure 3: Empirical density functions of fitness, diversity, and coexistence

The only case where the manual exam's objective qualities show concentration at the higher end of the distribution compared to the EGAL+ exams is in the two cases where the differences in the distribution are not considered significant at all levels (the coexistence values of the first quarter's multiple-choice and essay exams). This is further confirmed by the fact that the overall fitness of the EGAL+ exams is significantly higher in these cases. Overall, it can be concluded that the objective qualities of the exams are significantly higher for the EGAL+ compared to the Manual exams in all aspects, except for the first quarter's coexistence levels, where the distributions of EGAL+ exams are not significantly different from those of the manual exams at all levels.

Cramér's V coefficients measuring the consistency of subjective student perceptions between the multiple-choice, essay, and practical exams are provided in Table 4.

Cramér's V	**Quarter 1**	**Quarter 2**
	Practical	**Practical**
Essay	0.657	0.710
Multiple-choice	0.690	0.717

Table 4: Association results of subjective student opinions

The V coefficients indicate that the consistency between the EGAL+-affected essay and multiple-choice exams and the practical exam, which was unaffected by EGAL+, is around 0.7, representing the boundary between strong and moderate consistency [23]. Overall, it can be concluded that subjective student performance is relatively consistent between the EGAL+-affected and unaffected exams. This indicates that if a student perceives an exam unaffected by EGAL+ as easy or difficult, they generally tend to perceive the EGAL+-affected exam similarly. This tendency is evident even for the first quarter's essay and practical exams, where the consistency is at its lowest (0.657). The stacked column chart in Figure 4 shows that students who rated the essay exam

(affected by EGAL+) as high or low in difficulty were also given a high/low score for the practical exam (unaffected by EGAL+) similarly in large proportions.

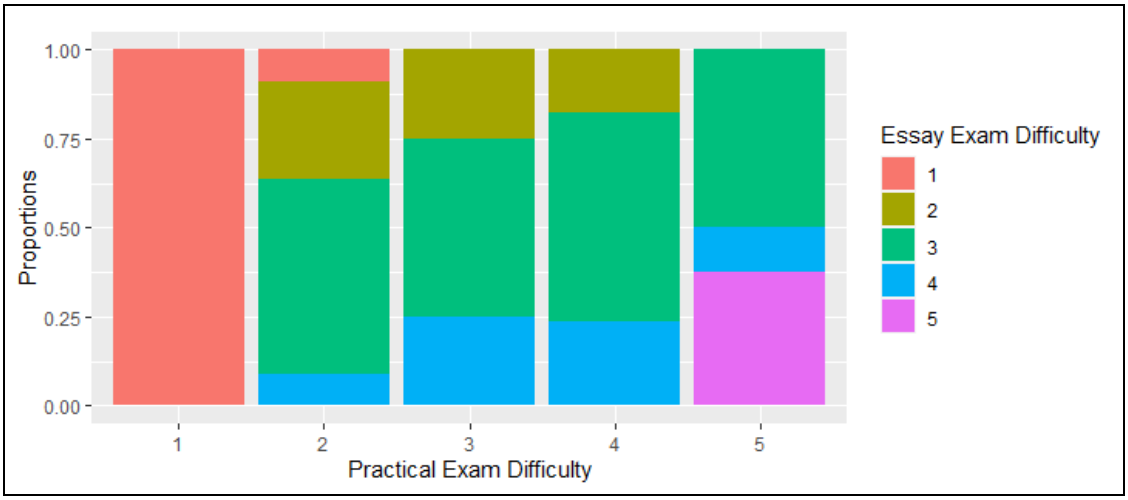

Figure 4: Consistency between practical and exam difficulty ratings in the first quarter

The tendency illustrated in Fig. 4 is consistent across the remaining three pairs and can be considered even stronger as Cramér's V is higher in these cases.

Estimated OLS regression coefficients for the models defined in Equation 5 are presented in Table 5.

	Multiple-choice - Quarter 1		**Essay – Quarter 1**		**Multiple-choice - Quarter 2**		**Essay – Quarter 2**	
	Coefficient	**p-value**	**Coefficient**	**p-value**	**Coefficient**	**p-value**	**Coefficient**	**p-value**
Intercept	6.470	0.00%	4.988	0.00%	7.970	0.00%	14.658	0.00%
Type = EGAL+	0.125	75.99%	0.685	43.39%	-0.157	70.56%	1.207	22.77%
Retake = Yes	-1.143	19.25%	-3.118	9.63%	-0.761	38.77%	-4.086	5.59%
Time	0.017	6.35%	0.043	2.84%	0.020	16.70%	0.023	51.24%
R^2	8.9%		13.3%		4.0%		8.1%	
Global F-test (p-value)	0.1172		0.0284		0.4552		0.1461	
White test (p-value)	0.8430		0.4963		0.9584		0.9410	

Breusch-Pagan (p-value)	0.6275	0.1467	0.7658	0.7819
Breusch-Godfrey (p-value)	0.8858	0.4046	0.5734	0.7962
Jarque-Bera (p-value)	0.1077	0.5211	0.2082	0.3904
RESET (p-value)	0.8503	0.8365	0.5655	0.3935

Table 5: OLS regression summary

Examining the coefficient estimates and the p-values of their partial t-tests, we concluded that none of the α_j coefficients are significant at any common significance levels, indicating that EGAL+ does not influence student performance scores, even if we control for individual student abilities and motivation with the $Retake$ and Time variables.

Overall, the models exhibit weak in-sample fit and mostly insignificant population fit with R^2 values below 10% and global F-test p-values exceeding all common significance levels, suggesting that model explanatory power is not significantly different from 0 in most cases [25]. The only exception is the model where the target variable is the essay performance score in the first quarter, where we have significant explanatory power at the 5% level but not at 1%. The R^2 shows a moderate in-sample fit of 13.3% explanatory power. Therefore, this model alone demonstrates some weak-moderate explanatory power and significant feature effects on some common significance levels. Here, it can be stated that we have the $Time$, and $Retake$ variables have some significant effect on student performance. $Time$ is significant at 5% and $Retake$ is significant only at 10%. However, the exam type (EGAL+ or manual) does not show a significant level of student performance scores even in this best-performing model, confirming it does not influence student performance at all.

Based on their coefficients, we can say that if we take two students with the same exam type (EGAL+ or manual) and retake (yes or no) values, then the student who opened the study materials on Moodle one more time is expected to score 0.043 points more on the essay exam in the first quarter. For the retake coefficient, we can interpret that if we take two students with the same exam type (EGAL+ or manual) and the same time spent studying on Moodle, then the student who already failed the course before is expected to score 3.118 points lower on the essay exam in the first quarter. The $Time$ variable has a similar effect on the multiple-choice scores in the first quarter as well, although only at a 5% level, while the $Retake$ variable is significant only at the 10% level. However, in these cases, the model itself is not significant at any common significance levels based on the p-values from the F-tests.

All four models are diagnostically appropriate. The e_{ij} residuals are homoscedastic and normally distributed at all common significance levels based on the p-values of the White + Breusch–Pagan, Breusch–Godfrey, and Jarque–Bera tests, respectively. Ramsey's RESET test indicates that model specification is appropriate at all common significance levels, with p-values above 10% in all models. This suggests that no nonlinear terms are needed to model student performance.

5. Conclusions

Based on the numerical results of our statistical analysis and the information obtained from the interview with the business intelligence course leader, we can conclude the research questions defined in Subsection 3.1 as follows

- The results of our OLS regression models suggest that student performance scores on the exams are not significantly influenced by the exam type (manual or EGAL+). The exam type variable is highly insignificant at every common level in each regression model result presented in Table 3. Additionally, the association analysis of subjective student perceptions indicates that students find the EGAL+-affected theoretical exams equally difficult as the practical exams that are unaffected by EGAL+. These results support the conclusion that there is no significant difference in student performance between T1 and T2 exams;
- The interview with the course leader suggests that EGAL+ can substantially reduce instructors' workload. In the business intelligence exams under investigation, the time required to prepare the exams is reduced from approximately 2.5 hours to 15 minutes when using EGAL+, according to the course leader;
- Analyzing and comparing the empirical distributions of the coexistence, diversity, and overall fitness values of the manual (T1) and EGAL+-generated (T2) exams reveals that EGAL+ can produce exams with significantly higher quality in all cases except two, where the qualities are generally comparable. In the multiple-choice exams of the first quarter, there are no significant differences in the coexistence distributions between the two exam types. Thus, in this aspect of coexistence, the manual and EGAL+ exams can be considered to have the same quality overall. Additionally, in the coexistence distribution, the essay questions of the first quarter show a significant difference at the 5% level, but not at the 1% level. Observing this difference in the visualized empirical densities in Fig. 3, we observe that this difference generally indicates higher coexistence quality for the manual exams. However, we cannot assert that this difference is significant at all levels. Furthermore, these results do not affect the overall fitness values, where we find significant differences at all levels in favor of EGAL+. Therefore, even if these differences in the empirical distribution are significant, their magnitude cannot be very large if they do not have a significant impact on the overall fitness distributions.

6. Discussions and Limitations

The present study aimed to address a common yet significant challenge in education: the labor-intensive process of manually compiling exam questions. By introducing a tool that harnesses a harmony search metaheuristic to manage the unique use of a preference matrix, which allows instructors to specify detailed priorities for question inclusion, this work not only demonstrates the feasibility of automated exam generation but also shows that such automation can yield tests comparable in validity and difficulty to manually compiled assessments.

The findings indicate that student performance on EGAL+-generated tests was statistically similar to that on manually created tests, thereby validating the quality of the generated question sets and confirming that the assembled exams can preserve the intended difficulty level while adequately covering the course content. Furthermore, the diversity factor, which strives to create several question sets as different, or even unique, as possible, is effective in minimizing opportunities for common cheating strategies, such as the sharing of answers among students.

From the instructor's perspective, EGAL+ significantly reduces the workload associated with exam preparation. By instantly generating numerous high-quality and varied exam sets, the tool liberates educators to devote more time to pedagogically valuable activities, such as personalized student support and curriculum refinement.

Despite these promising outcomes, several areas for further improvement have emerged. First, the current system uses a static assignment of difficulty levels. Therefore, future enhancements could incorporate adaptive mechanisms that update these values based on student performance data. Such feedback loops would refine the question bank over time, thereby ensuring increasingly precise calibration of exam difficulty. Second, while the preference matrix provides a robust framework for incorporating instructor priorities, the relative weighting between diversity and coexistence preferences is presently fixed. Allowing teachers to dynamically adjust this balance during the exam-generation process could further tailor the outputs to specific pedagogical goals. Finally, integration with learning management systems via application programming interfaces and improvements in the user interface could enhance both the accessibility and utility of EGAL+.

In summary, the deployment of EGAL+ in a live university examination setting underscores its potential to enhance exam preparation across a broad range. By effectively reducing teacher workload, enhancing the diversity and compliance of exam content with instructional priorities, and mitigating opportunities for cheating, EGAL+ addresses critical gaps identified in existing research. The continued evolution of this tool promises to not only improve automated exam generation but also contribute significantly to the broader discourse on educational assessment innovation.

Bibliography

1. Láng, B., & Dömsödi, B. (2024). Integration of AI and metaheuristics in educational software: A hybrid approach to exercise generation. *International Journal of Emerging Technologies in Learning, 19(6)*, 38–51. https://doi.org/10.3991/ijet.v19i06.49829
2. Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Handbook 1: Cognitive domain. David McKay.
3. Teo, N. H. I., Bakar , N. A., and Karim, S. (2012). *Designing GA-based auto-generator of examination questions*. In 2012 Sixth UKSim/AMSS European Symposium on Computer Modeling and Simulation (pp. 60–64). https://doi.org/10.1109/EMS.2012.69
4. Rahim, T., Abd Aziz, Z., Rauf, R., & Shamsudin, N. (2017). *Automated exam question generator using genetic algorithm*. In 2017 International Conference on Computer and Communication Engineering (IC3E) (pp. 12–17). https://doi.org/10.1109/IC3E.2017.8409231
5. Ciguené, R., Joiron, C., & Dequen, G. (2019). *Automatically generating assessment tests within higher education context thanks to genetic approach*. In E. G. Talbi & A. Nakib (Eds.), Bioinspired heuristics for optimization. studies in computational intelligence (Vol. 774). Springer, Cham. https://doi.org/10.1007/978-3-319-95104-1_17
6. Shanthi, B. S. A., Harshitha, L. J. R., & Manasa, K. (2019). *Automated exam question generator using genetic algorithm*. International Research Journal of Engineering and Technology (IRJET), 1687–1691.

7. Wang, Y., & Wang, X. (2022). *Test paper automatic generating method based on hybrid genetic algorithm*. In S.C. Chu, J. C. W. Lin, J. Li, & J. S. Pan (Eds.), Genetic and evolutionary computing. ICGEC 2021. Lecture notes in electrical engineering (Vol. 833, pp. 417–426). Springer. https://doi.org/10.1007/978-981-16-8430-2_54

8. Popescu, D. A., Stanciu, G. C., & Nijloveanu, D. (2023). *Application of genetic algorithm in the generation of exam tests*. In V. E. Balas, L. C. Jain, M. M. Balas, & D. Baleanu (Eds.), Soft computing applications. SOFA 2020. Advances in intelligent systems and computing (Vol. 1438, pp. 213–224). Springer. https://doi.org/10.1007/978-3-031-23636-5_12

9. Zhou, C., Lin, L., & Shuai, P. (2018). *Design of auto-generating examination paper algorithm based on hybrid genetic algorithm*. In 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA) (pp. 563–567). https://doi.org/10.1109/ICCCBDA.2018.8386579

10. Rahim, T., Batcha, M., & Abd Aziz, Z. (2020). *Automated exam question set generator using utility-based agent and learning agent*. International Journal of Machine Learning and Computing, 10(1), 164–169. https://doi.org/10.18178/ijmlc.2020.10.1.914

11. Wu, Z., He, T., Mao, C., & Huang, C. (2020). *Exam paper generation based on performance prediction of student group*. Information Sciences, 532, 114–126. https://doi.org/10.1016/j.ins.2020.04.043

12. Joshi, A., Kudnekar, P., Joshi, M., & Doiphode, S. (2016). *A survey on question paper generation system*. In National Conference on Role of Engineers in National Building (pp. 1–4).

13. Rao, P., Kiranmai, T., Samhitha, E., Shiva, R., & Kusuma, S. (2022). *A survey on automated assessment questions generation system using supervised algorithms*. International Journal for Research in Applied Science and Engineering Technology, 10, 1675–1677. https://doi.org/10.22214/ijraset.2022.47700

14. Bobade, M. P., Dandge, S., & Punds, M. (2018). *A study of different algorithms for automatic generation of question paper*. International Science and Technology Journal, 7(5), 27–32.

15. Geem, Z. W., Kim, J. H., & Loganathan, G. V. (2001). *A new heuristic optimization algorithm: Harmony search. Simulation*, 76(2), 60–68. https://doi.org/10.1177/003754970107600201

16. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). *A survey on evaluation of large language models*. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–45. https://doi.org/10.1145/3641289

17. Monge, M. (2023). *Two-sample Kolmogorov-Smirnov tests as causality tests: A narrative of Latin American inflation from 2020 to 2022*. Revista Chilena de Economía y Sociedad, 7(1), 68–78. https://rches.utem.cl/?p=2374

18. Viehmann, T. (2021). *Numerically more stable computation of the p-values for the two-sample Kolmogorov-Smirnov test* (arXiv:2102.0803). arXiv. https://doi.org/10.48550/arXiv.2102.08037

19. Bakoyannis, G. (2020). *Nonparametric tests for transition probabilities in nonhomogeneous Markov processes*. Journal of Nonparametric Statistics, 32(1), 131–156. https://doi.org/10.1080/10485252.2019.1705298

20. Marozzi, M. (2013). *Nonparametric simultaneous tests for location and scale testing: A comparison of several methods*. Communications in Statistics-Simulation and Computation, 42(6), 1298–1317. https://doi.org/10.1080/03610918.2012.665546

21. Schröer, G., & Trenkler, D. (1995). *Exact and randomization distributions of Kolmogorov-Smirnov tests two or three samples*. Computational Statistics & Data Analysis, 20(2), 185–202. https://doi.org/10.1016/0167-9473(94)00040-P

22. Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge. https://doi.org/10.4324/9780203771587

23. Frey, B. B. (Ed.). (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. Sage Publications. https://psych.wisc.edu/Brauer/BrauerLab/wp-content/uploads/2014/04/Murrar-Brauer-2018-MM-ANOVA.pdf https://doi.org/10.4135/9781506326139

24. Békés, G., & Kézdi, G. (2021). *Data analysis for business, economics, and policy*. Cambridge University Press. https://doi.org/10.1017/9781108591102

25. Wooldridge, J. M. (2016). *Introductory econometrics: A modern approach*. Nelson Education.

26. Bayoud, H. A. (2021). *Tests of normality: New test and comparative study. Communications in Statistics-Simulation and Computation*, 50(12), 4442–4463. https://doi.org/10.1080/03610918.2019.1643883

27. Long, J. S., & Trivedi, P. K. (1992). *Some specification tests for the linear regression model.* Sociological Methods & Research, 21(2), 161–204. https://doi.org/10.1177/0049124192021002003

Authors

LÁNG Blanka
Eötvös Loránd University, Faculty of Informatics, Department of Media & Educational Technology, Hungary,
e-mail: langblanka@inf.elte.hu

KOVÁCS László
Corvinus University of Budapest, Institute of Data Analytics and Information Systems, Department of Statistics, Hungary,
e-mail: laszlo.kovacs2@uni-corvinus.hu

DÖMSÖDI Balázs
Corvinus University of Budapest, Doctoral School of Economics and Business Informatics, Hungary,
e-mail: balazs.domsodi@stud.uni-corvinus.hu

About this document

Published in:

CENTRAL-EUROPEAN JOURNAL OF NEW TECHNOLOGIES IN RESEARCH, EDUCATION AND PRACTICE

Volume 8, Number 1. 2026.

ISSN: 2676-9425 (online)

DOI:

10.36427/CEJNTREP.8.1.12403

License