

# Class integration of ChatGPT and learning analytics for higher education

Miguel Civit<sup>1,2</sup>  | María José Escalona<sup>1</sup> | Francisco Cuadrado<sup>2</sup> | Salvador Reyes-de-Cozar<sup>2</sup> 

<sup>1</sup>Departamento de Lenguajes y Sistemas Informáticos, Escuela Superior de Ingeniería Informática, Universidad de Sevilla, Sevilla, Spain

<sup>2</sup>Departamento de Educación y Comunicación, Universidad Loyola, Sevilla, Spain

## Correspondence

Miguel Civit, Departamento de Lenguajes y Sistemas Informáticos, Escuela Superior de Ingeniería Informática, Universidad de Sevilla, Avenida Reina Mercedes S/N, Sevilla 41012, Spain.  
 Email: [miguelcivit@us.es](mailto:miguelcivit@us.es)

## Funding information

European Union, Erasmus+ grant Teaching to Be: Supporting Teachers' Professional Growth and Wellbeing in the Field of Social and Emotional Learning. Grant/Award Number: 626155-EPP-1-2020-2-LT-EPPKA3-PI-POLICY; Spanish Ministry of Science and Innovation MCIN/AEI/10.13039/501100011033/, Grant/Award Number: PID2022-137646OB-C31; European Union

## Abstract

**Background:** Active Learning with AI-tutoring in Higher Education tackles dropout rates.

**Objectives:** To investigate teaching-learning methodologies preferred by students. AHP is used to evaluate a ChatGPT-based studented learning methodology which is compared to another active learning methodology and a traditional methodology. Study with Learning Analytics to evaluate alternatives, and help students elect the best strategies according to their preferences.

**Methods:** Comparative study of three learning methodologies in a counterbalanced Single-Group with 33 university students. It follows a pre-test/post-test approach using AHP and SAM. HRV and GSR used for the estimation of emotional states.

**Findings:** Criteria related to in-class experiences valued higher than test-related criteria. Chat-GPT integration was well regarded compared to well-established methodologies. Student emotion self-assessment correlated with physiological measures, validating used Learning Analytics.

**Conclusions:** Proposed model AI-Tutoring classroom integration functions effectively at increasing engagement and avoiding false information. AHP with the physiological measuring allows students to determine preferred learning methodologies, avoiding biases, and acknowledging minority groups.

## KEYWORDS

cooperative/collaborative learning, data science applications in education, post-secondary education, teaching/learning strategies

## 1 | INTRODUCTION

### 1.1 | Active learning methodologies

In recent decades, Higher Education institutions have faced problems such as demotivation, anxiety, or lack of participation in their classrooms. Although the implementation of support programs by educational institutions has tried to alleviate this situation, the dropout rate, which is still around 30% (Ifenthaler & Yau, 2020; Aina et al., 2022) has not decreased in recent years. Researchers claim that among the most important factors that lead students to give up or fail their studies is the low motivation with which they find themselves during the academic process (Pascoe

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Expert Systems* published by John Wiley & Sons Ltd.

et al., 2020). The European Union, and Spain in particular have been concerned with their dropout rates and have ordered studies into this state of affairs (Commission, Directorate-General for Education, & Culture, 2022; Mellizo-Soto, 2022). Although previous studies suggest that this is prevalent across different universities and areas of study (Shen et al., 2022), with sociocultural factors, student identification with the course, difficulty of the course, and low motivation and self-regulation being some of the main causes of dropping out signaled by the EU (Commission et al., 2022), especially in first-year dropouts. However, areas related to STEMs may have an increased dropout rate due to the initial barrier of technical subjects or lower levels of engagement (Reyes-de C  zar et al., 2023). Consequently, in this study, our focus is on Social Sciences & Humanities subjects and courses.

Furthermore, another problem that worries university students is the prevalence of negative emotional states such as anxiety or burnout syndrome, as well as the daunting effect that these factors have on learning retention and therefore on academic performance (Cooper et al., 2018). Such a scenario presents the challenge of knowing how to listen to both parties, both educational institutions and their role in the teaching process, as well as the needs of students in the classroom. On the one hand, we find the demand from new generations of students who are looking for more reactive, stimulating, and immediate learning. On the other hand, the position adopted by many institutions regarding the learning methodologies implemented has often stalled as traditional. These methodologies, characterized by demanding a passive role from students and where the master class is privileged, tend to receive criticism for not being able to generate engagement, not always reaching good learning outcomes, or not contributing to the satisfaction of students during the teaching process (Bond, 2020).

Faced with this situation, some solutions point to the use of new methodologies in the classroom (B  rte et al., 2023), exploring new teaching approaches that, in addition to meeting the demands of the new generations, favor a more active role for the student in the classroom environment (Aidoo et al., 2022). These efforts include active learning methodologies, which are postulated as an alternative, capable of achieving greater engagement in university students, increased motivation, and, as a consequence, an improvement in their academic performance. In addition, their use opens the possibility of taking advantage of technological advances to create new dynamic and innovative opportunities in learning, offering different narratives and discourses better adapted to the needs of the classroom. Among these are methodologies and strategies such as gamification (Ekici, 2021) or the flipped classroom (Zou et al., 2022), in addition to digital tools such as videos, quizzes, and Kahoot! (Wang & Tahir, 2020), Socrative (Kocak, 2022) or Edpuzzle (Hidayat & Praseno, 2021), among others.

The concept of active learning methodologies has been approached from disciplines as diverse as education, social psychology, or engineering, although without reaching a common consensus on its definition (B  rte et al., 2023). Using a pedagogical approach, according to Hartikainen et al. (2019), an active learning methodology is any teaching method that involves students in the learning process. This is a clearly holistic approach, and under its umbrella we can find many different instructional methodologies and processes, including experiential, collaborative, or cooperative learning, as well as case-based, inquiry-based, problem-based, team-based, or game-based learning. Conceptually, active learning is related to constructivism theories, which consider learning as a knowledge construction process in relation to prior knowledge (Driessen et al., 2020), and which places understanding before memorization (Mitry, 2021). Active learning involves a change in the role of students, going from playing a passive role, as in traditional methodologies, to a clearly active role in the classroom. This also requires students to carry out significant learning activities and think about what they are actually doing and the processes in which they are engaged (Hartikainen et al., 2019).

Active learning methodologies have been in the spotlight in recent years, and, like other educational methodologies, they face a mixture of praise and criticism. Although most specialists consider them a good alternative to traditional methodologies, there are those who claim that they may be a temporary educational trend (Hood et al., 2021). There are several research works that have highlighted the benefits of active learning methodologies, including those that focus on: how active learning methodologies have been shown to be an effective way of teaching in the field of sciences (Theobald et al., 2020), how they constitute an effective way of improving student engagement (Fisher et al., 2021), they increase student academic performance (Bressane et al., 2021) or enhance motivation in the classroom (Castro, 2019). Furthermore, other studies state that students develop and improve important skills such as problem solving, peer collaboration, critical thinking, or analytical reasoning (Bean & Melzer, 2021; Ting et al., 2019).

Another important aspect to consider is the positive perception that students have about active learning methodologies. Students show a special interest in the use of active learning methodologies in the classroom (Brown et al., 2017) since, in general, they consider that they learn more and better while also helping them develop self-regulatory management habits outside of the classroom, such as autonomy and other social skills (Akiha et al., 2018). Even more importantly, they positively valued characteristics such as interactivity or collaboration among equals, as well as the impact on their academic performance (Goode et al., 2022).

Although active learning methodologies present a promising scenario in the educational field, some caution is still advised. Some research works still stand out as contrary voices for several reasons. First, they argue that there exists a lack of empirical research that measures the effectiveness of these pedagogical approaches (Carpenter et al., 2020). However, when these methodologies are compared to traditional ones, some studies claim that student engagement does not increase significantly (Deslauriers et al., 2019) or that its contribution to lifelong learning is not demonstrated by conclusive evidence (Gu et al., 2022).



## 1.2 | Flipped learning and gamification

Within the field of active learning methodologies, two of the most widely used strategies are flipped learning and gamification; thus, they offer a great starting point for comparisons and the implementation of new technologies. In the flipped learning pedagogical approach, instruction moves from the group learning space within the classroom to the individual learning space, and the group space is transformed into a dynamic and interactive environment where the lecturer mediates the interaction between students with the concepts learned (Al Mamun et al., 2022). To succeed in this methodology, the work of the students at home, before the actual class, is essential. The core value of this strategy lies in the collective debate that arises in the class after each student had previously internalized the knowledge by studying individually. The flipped methodology divides the learning objectives into low-order objectives (such as comprehension of content) and higher-order ones (such as information analysis). The low order can be reached autonomously at home, while the high order ones are best achieved through classroom interaction (Fisher et al., 2021).

For some researchers, the flipped classroom represents an educational methodology that reveals numerous advantages. It increases engagement, improves satisfaction, and produces better learning outcomes, in addition to positively influencing student learning retention (Prieto et al., 2021; Bowers et al., 2019; Fischer, 2019). Its power to increase self-efficacy (Mehta et al., 2022) and motivation (Vergara et al., 2020), in addition to developing critical thinking, creativity, communication, and collaboration skills (Lugosi & Uribe, 2022) is highlighted throughout the scientific corpus. It also provides efficient use of teaching time (Dejene, 2019) and promotes richer and deeper learning (Fisher et al., 2021).

Another approach to active learning that has gained significant popularity in recent years is gamification. Game-based learning allows to merge educational content with elements of games, creating an optimal context for problem solving or research (Hammill et al., 2021; Murillo-Zamorano et al., 2021). Some factors supporting this pedagogical strategy are the inclusion of differentiating and motivational elements such as the pursuit of unknown learning results, the possibility of choosing multiple paths towards an objective, greater autonomy in the construction and resolution of problems, collaboration between students/players and elements of competition and chance (Hu et al., 2022). For some authors, gamification brings different benefits to the learning process, including improved student motivation and experiential learning thanks to the interactivity of digital media (Shroff & Keyes, 2017), a better assimilation of complex concepts (Rodriguez et al., 2018), performance improvement and promotion of self-confidence (Li et al., 2021).

For all these reasons, active learning methodologies are an alternative in the pedagogical field, where, even though their scientific credibility has clearly increased in the last decade, there is still the necessity to find formulas that further prove their usefulness and consistency, especially using empirical approaches. Thus, there is a clear need for research on measurement methodologies for active learning (Hartikainen et al., 2019). In this sense, learning analytics are becoming particularly relevant, since the importance and rigor of the data they can offer to the field of education is of great importance to the scientific community. The variety of data that these tools offer, coupled with the possibility of measuring factors that are difficult to measure by other means, can offer great benefits, not only for lecturers and students, but also in the development of active learning environments (Gašević et al., 2022).

## 1.3 | Large language models for education & learning analytics

In the work of Luckin et al. (2016) a future was depicted in which AI was a redefining tool for the field of education. Although it could be argued that it was written before the invention of modern AI technology, it provides a good categorization for the possibilities that AI may have. As technology advances, more research and commercial applications arise that investigate into the relation of AI and education. In particular Wei et al. (2022) explores the integration of AI into college music education and suggests how there exists a necessity to actualize the curricula to include AI technology that could be very beneficial to students. So far many efforts have been focused on the development and improvement of learning platforms through this technology. However, the availability of newly developed AI-based large language models, such as ChatGPT, has made AI personal tutoring and AI expert facilitation a possible reality for mainstream education (Kasneci et al., 2023).

ChatGPT and other large language models (LLMs) offer many possibilities for their use in educational environments. In Murgia et al. (2023) the authors studied the use of LLMs to support children's information discovery in the classroom. Considering their use for querying and navigating information, evaluating the credibility of information, summarizing, and synthesizing information. The authors found that LLMs have the potential to support children's information discovery, but presents a set of challenges related to bias, accuracy, transparency, and safety. The authors recommended that future research focus on developing methods to improve the accuracy and transparency of LLMs and to identify and mitigate biases in their responses.

In Javaid et al. (2023), the potential of LLMs to revolutionize modern education is discussed. It is argued that LLMs can be used in a number of ways, including personalized instruction, adaptive feedback, active learning, work group facilitation, motivation encouragement, promotion of diversity and inclusion, and accessibility improvement. The paper also provides some approaches to handle the challenges stated in Murgia et al. (2023) based on teacher training, context-specific adaptation, human oversight, and collaboration between teachers, researchers, and developers to share best practices, identify challenges, and develop effective strategies for integrating ChatGPT into education.

Moreover, Lai et al. (2023) states how the role of intrinsic motivation in the adoption of ChatGPT for active learning is discussed and considered as a key factor for LLMs acceptance that can be fostered by providing students with opportunities to learn in a way that is engaging and relevant to their interests. The authors provide some specific recommendations for fostering intrinsic motivation in ChatGPT-based active learning interventions. These are related to creating engaging and interactive learning activities, for example, by simulating conversations with experts, playing games, or creating quizzes. It is also important to provide students with choices on how to use LLMs and to connect their use with the students' interests. Formulas for recognizing and rewarding students' interest are also important for the successful use of LLMs in active learning environments.

Another area of trending research interest in the field of Education is Learning Analytics. This is a loosely defined concept with different applications in the field of education and their use and application in educational research, due to its predictive and prescriptive capacity (Romero & Ventura, 2020), is presented as a set of strategies and tools that amplifies the information obtained and allows the corroboration of scientific findings obtained using traditional research techniques. Learning analytics can allow us to refine and reduce the natural bias of some data collected in the social sciences when working with people, perceptions, and interpretations and can be especially useful for evaluating newly developed and trending educational strategies such as different LLMs class implementations. With a focus on Higher Education, Learning Analytics could be defined as "the use, evaluation, collection, and analysis of static and dynamic information about students and learning environments, for modeling, prediction, and optimization in near real time of learning processes and learning environments" (Ifenthaler, 2017). Recently, scientific studies whose results are supported by these metrics have become fairly common. An example of this approach can be found in Glick et al. (2019), who used learning analytic techniques to measure student motivation, or in Ifenthaler et al. (2019) whose study focused on improving the learning process towards better adapted objectives for students.

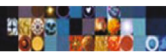
Learning Analytics can also provide a discerning service for areas where there may exist contradicting views or where further research is required. As discussed previously, it seems evident that active learning methodologies are recognized by the scientific community as teaching strategies adapted to the needs of new generations due to their motivational capacity and their capacity for engagement. However, opposing voices point out that student motivation and engagement, by themselves, are not directly related to academic performance or learning. Although it seems logical to think that in a motivating learning context with highly engaged students, success is ensured; lasting and meaningful learning is not always guaranteed. In this sense, new data collection strategies offered by learning analytic tools, with their ability to measure aspects that fall outside Likert scales, tests, interviews, perceptions, and opinions, stand out as well suited tools that allow researchers to delve into aspects that can otherwise be difficult to measure and are well suited for cross-referencing the results, avoiding possible biases. In addition, these tools can serve to test long-standing assumptions held by the scientific community, opening a new stage in the field of education for generating and validating new paradigms.

Furthermore, researchers focused on Learning Analytics are paying increasing attention to multi-criteria decision-making methodologies (Kabassi & Alepis, 2020; Divjak et al., 2022). In particular, the Analytic Hierarchy Process (AHP) (Saaty, 2010) presents a formal way of quantifying the qualitative criteria for different alternatives, thus reducing the subjectivity of the results. Firstly introduced by Thomas L. Saaty in the 1970s, it is based on the theory of positive reciprocal matrices and includes methods to check for inconsistencies. The ability of this method to make decisions by pairwise comparison of uncertain, qualitative, and quantitative factors is an important reason for the selection of this evaluation methodology.

Regarding the use of physiological data in Multimodal Learning Analytics, while there is a well-established research foundation for the use of wearable technology in other areas, for example, in healthcare (Wu & Luo, 2019) its application in the field of education is still gaining momentum. The original Horizon Report (Johnson et al., 2014) in 2014 stated that wearable technologies would become a dominant trend in higher education institutions in the near future, but this is still very much a work in progress. The New Media Consortium Horizon Report: 2022 Higher Education Edition (Pelletier et al., 2022) establishes AI-based predictive analytics as one of the main future research fields. As a result, the study of analytics using all available data, including wearable devices and predictive AI, is attracting the attention of more researchers (Liu et al., 2022).

## 2 | OBJECTIVES AND RESEARCH QUESTIONS

Based on the aforementioned aspects, the purpose of this study is to verify the opinion of the students on different assumed ideas (engagement, anxiety, perceived learning and academic performance) of the teaching-learning processes, especially when comparing active learning methodologies with more traditional approaches. Furthermore, we aim to verify a model for integrating ChatGPT in a gamified class through student validation. Our main hypothesis is that engagement and general student preference are going to be higher for active methodologies than for traditional ones, while in academic performance, the reverse might happen. We also expect that the integration of AI tutoring in a gamified class performs well in all aspects due to the reinforcement of student autonomy. To this effect, data from different sources will be collected (traditional, AHP pairwise comparisons, and other learning analytics such as physiological data processing), and the obtained data will then be cross-referenced. This will allow the opening of a debate on the ideas, perceptions, and opinions of students when faced with these methodologies and the correlation with data measured by quantitative and qualitative means.



We focus on answering the following research questions:

**RQ1.** Are the presuppositions of the students about the different methodologies studied related to their perceived learning, emotional state, and academic results?

**RQ2.** How accurate is the perception of learning by the students compared to their actual academic results? Is there a significant difference between the more standardized methodologies and the ChatGPT class-integration for the field of music study?

**RQ3.** How is the implementation of an active learning method based on a Large Language Model regarded by students?

Thus, with the use of the methodology, tools and intervention design described in the next Section 3 we analyze an implementation of ChatGPT in a gamified class and compare it to two other well-established Teaching-Learning strategies, Flipped classroom and Master class. We analyze the results (Section 4) of different Learning Analytics that were used both during and after the intervention with special consideration to student self-evaluation, and emotional/stress measuring through self-assessment and physiological state measurements (Jiménez-Mijangos et al., 2023). Furthermore, a discussion on the necessity to evaluate the results and engagement of said strategies is provided in Section 5 giving special consideration to the ChatGPT integration and the preference assessment in student-centered AI-Tutoring. As a secondary objective, we will investigate how and if the design of the intervention and data gathering strategy could be used to include students' preferences in the teaching/learning methodology selection for their own curriculum.

### 3 | MATERIALS AND METHODS

#### 3.1 | Sampling and participants

To select the participants of the study all students were contacted by email and an online survey for data collection and time scheduling purposes. The inclusion criteria were:

Not being a student of the researcher implementing the study.

- Being in good physical health that does not hinder good physiological measurement and allows good visualization of a screen device.
- Not being an expert on the selected subjects.
- Having completed all stages of the intervention.
- A higher education student who has completed at least one semester.

All students were asked if they had special needs (such as examination time) for any reason or other special requirements (due to hearing impairment, low mobility, etc.). However, this was not part of the inclusion/exclusion criteria. The study was carried out in class and took place at Universidad Loyola. The study was approved by the Ethics Committee of the Universidad Loyola.

#### 3.2 | Measuring instruments and data analysis strategy

In the intervention, the students were divided into five randomly created groups. Each of them had three classes, each with its own dynamics, and each was followed by a test of similar difficulty. After each class, the students received an SAM-based survey to determine their level of engagement, and after each exam another survey to determine their perceived stress levels during the test. This survey system was selected due to the ease of integrating it with the rest of likert-based surveys. This allowed students to graphically and dynamically reference their emotional state without the need of extensive explanations and pre-training that other self-reporting methods require. To contrast this information and cross-reference with other measurable data, a student per group received an Empatica E4 device to measure their physiological responses.

When all classes and tests were completed, an AHP-based survey was carried out to determine students' preferences and perceptions. AHP is a good for analyzing opinions of small samples and can be used to both identify clusters of students with opinions different from the majority, as well as to reach consensus in decision-making through iterative use (Lai et al., 2002). This makes it an interesting tool, not only when considered as learning analytics, but also if the purpose is to take students' preference into consideration for self-determining teaching/learning methodologies in the class.

Further explanations of the specific methods can be found below in this section and the results of each test and survey can be found in Section 4.

### 3.2.1 | Analytic hierarchical process-based data analysis

The AHP is a multi-criterion decision making (MCDM) approach for structuring multiple choice criteria into a hierarchy, assessing the relative importance of these criteria, comparing alternatives for each criterion, and determining an overall ranking of the alternatives. By structuring the problem in this way, it gains clarity and visibility. AHP has been shown to be suitable for the analysis of data in small groups (Melillo & Pecchia, 2016; Saaty & Özdemir, 2014).

AHP has been used in the education field, among other possible topics, to assess and address problems in Higher Learning Institutions (Anis & Islam, 2015), assess schools (Fu et al., 2022), evaluate online learning technologies such as chatbots (Smutny & Schreiberova, 2020), or even evaluate educational alternatives in a pandemic situation (Hasan et al., 2020). More directly related to our work, AHP has been used to evaluate Flipped Classroom Learning and Gamification in massive online courses. In Mohammed and Daham (2021) Flipped Classroom learning is compared to traditional learning in a secondary school environment using AHP based on the criteria of teaching techniques, learning flexibility, teaching aids, student participation, student achievement and working environment. In Chang and Wei (2016) 40 widely used gamification mechanisms are evaluated in relation to the interaction between the learner-learner, the learner-instructor, and the learner-content in an online setting. Nevertheless, a study centered on the possibility of implementation of this method to choose learning methodologies by students was needed, in particular in the controversial field of music education, where there exists a wide array of multiple solutions and an extensive demand of students for highly engaging interactions.

The possibility of making decisions with small groups and the existence of previously proven applications in the educational field are the main reasons for our decision to implement, after each subgroup finished all classes and tests in the intervention, an AHP-based survey to estimate the students' preferences toward learning criteria and learning methodologies.

Regarding the software tool used to perform the analysis using the AHP methodology, we used AHP-OS (AHP online system) (Goepel, 2018). AHP-OS is an online software tool for AHP also available as an open-source package on GitHub. The evaluation process was carried out by a total of 33 students who completed the study. Each student received an email with a link to conduct a pairwise comparison of all criteria and was asked to provide his opinion using the standard 9-point AHP linear scale shown in Table 1.

In our experiment, we explicitly decided to keep the size of the AHP hierarchy relatively small. Although there are alternatives to reduce the number of comparisons (Leal, 2020), since our students do not have prior experience with this methodology and it is not part of their curriculum, we decided to keep the AHP hierarchy as small as possible. As we show in the following section, this small hierarchy has provided significant information without an excessive workload for students. The AHP hierarchy used in our experiments is shown in Figure 1. These short-form AHP surveys have proven through our experience to be a highly effective and easy-to-replicate method when used as part of a decision-making process where students can get involved in designing their own curriculum and how it can be taught. More complex 7 plus criteria surveys tend to require extensive pre-training of the participants if they have not previously used this method.

It should be noted that AHP and the AHP-OS tool include very thorough data diagnosis components that perform validation of the analysis results. AHP does not require full logical consistency in the judgments; the widely used consistency ratio CR (Teknomo, 2006) is an indicator of logical consistency and, in general, should not significantly exceed the value 1 in individual judgments. The AHP-OS tool provides participants in a group decision an indicator of the inconsistency of their pairwise comparisons and provides suggestions to adjust these comparisons that can be accepted or modified by participants.

In an AHP-based group decision process, it is always possible to calculate a group result. However, the whole-group decision does not make sense in all cases, for example, if half of the decision makers have totally opposite judgments for two criteria; the aggregation will result in equal weight distribution (50/50) for both criteria. In this case, there is no agreement between the decision makers and this can end in a deadlock situation. It can be shown that, given a sufficiently large group, the consistency of the aggregate comparison matrix is guaranteed, regardless of the consistency measures of the individual comparison matrices (Aull-Hyde et al., 2006). If the group of students participating in the experiment is large enough (in most cases, 7 participants are enough Saaty and Özdemir (2014)), the consistency ratio of the CR matrix of the consolidated

**TABLE 1** Saaty's nine-point criteria comparison scale.

| Definition                                             | Relative importance |
|--------------------------------------------------------|---------------------|
| Extremely preferred                                    | 9                   |
| Very strongly preferred                                | 7                   |
| Essentially preferred                                  | 5                   |
| Moderately preferred                                   | 3                   |
| Equally preferred                                      | 1                   |
| Intermediate importance between two adjacent judgments | 2, 4, 6, 8          |

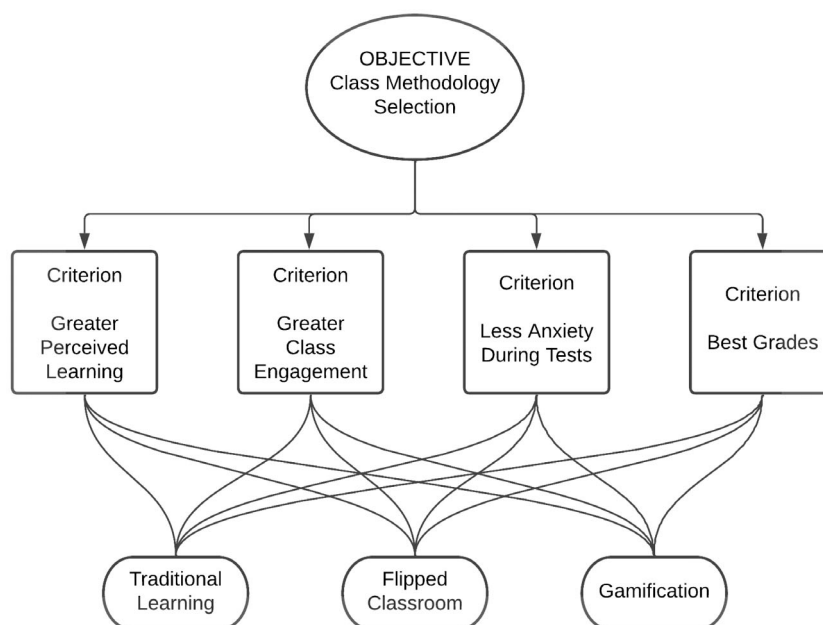


**FIGURE 1** AHP hierarchy.

group will decrease below 10% and will no longer be an issue. This is one of the characteristics that makes AHP suitable for our reduced group testing study.

A very interesting aspect of AHP is the possibility dividing the original group of participants according to the similarity of their judgments and finding clusters of participants with high consensus among themselves, but with low consensus compared to other clusters. This can be used even in relatively small group studies. We will use this feature to divide students into coherent subgroups related to their opinions on educational alternatives. Another benefit of this methodology is that it can be easily used to detect outliers using group clustering, as can be seen through the various tables in the Section 4. There are several alternatives to perform similarity based clustering in AHP. Among them we can find Consensus Index (Dong et al., 2013), Consensus Consistency Matrix (CCM) (Aguarón et al., 2016) and AHP group consensus indicator Goepel (2018).

The AHP consensus indicator  $S^*$  is based on Shannon's entropy (Jost, 2006) and has normalized values in the [0,1] interval. It is an AHP specific version of the original  $S$  similarity indicator. The value zero corresponds to no consensus among the group, whereas one corresponds to full agreement. Consensus indicator values below 0.5 indicate that there is very little consensus within the group, and thus, a high diversity of judgments. Values above 0.75 indicate excellent agreement among group members. Consensus must be clearly distinguished from consistency, as, both small or large groups, consensus could be high or low, reflecting the 'agreement' between group members. The AHP consensus indicator is integrated in the AHP-OS tool and provides good clustering results in cases where subgroups of participant with similar opinions exist.

### 3.2.2 | Emotional state estimation

Engagement in active learning-related activities is a fundamental part of the methodologies mentioned above. Furthermore, stress/anxiety levels are closely related to academic performance during tests and to general student satisfaction (Loderer et al., 2020). This is why estimating emotional states in students during both classes and tests can be an extremely useful research tool. In the study, we used a survey based on self-assessment manikins (SAM) after each class and after each test, which allowed students to self-evaluate their emotional state.

SAM are a picture-oriented approach to measure emotional responses that have also been used successfully in other studies related to music education (Cuadrado et al., 2020). They intend to estimate three central features of emotional responses: perceived valence, perceived arousal, and perceptions of dominance.

To develop a faster and more objective emotional estimate for future experiments that could be able to monitor such states in real time, and, to compare and assess the validity of students' self-assessments, we recorded galvanic skin responses (GSR) and heart interbeat intervals (IBI) on a small subset of students using an Empatica E4 wristband (McCarthy et al., 2016). The recordings were analyzed using statistical parameters calculated with the Kubios heart rate variability (HRV) application (Tarvainen et al., 2014) and the EDA Explorer from MIT medialab (Taylor et al., 2015) for the analysis of GSR. We also used an AI-based stress/engagement estimation network developed in Muñoz-Saavedra et al. (2023). The version of the network used is shown in Figure 2. Traditional (not AI) estimation of stress and engagement is based on linear

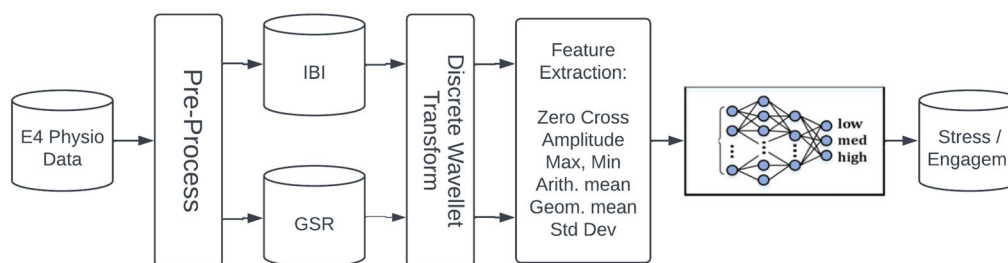


**FIGURE 2** ML stress/engagement predictor.

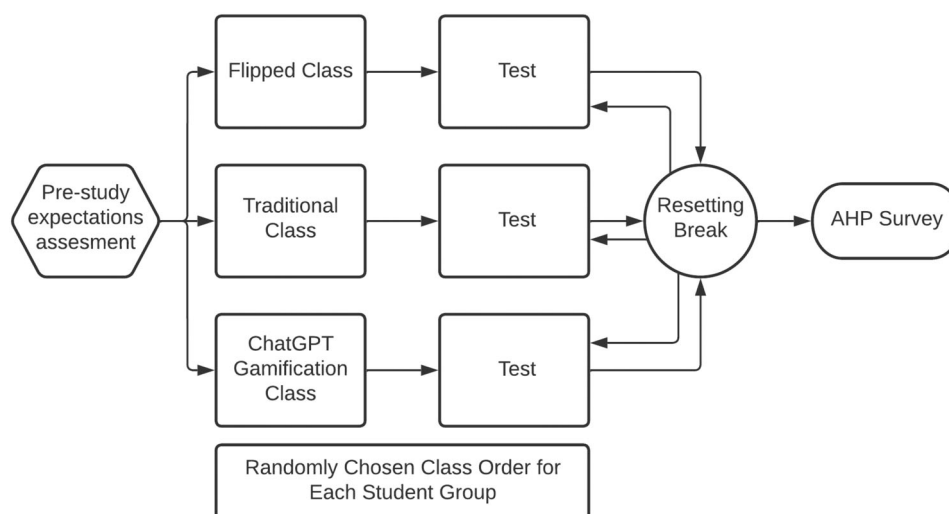


**FIGURE 3** Study design outline.

combinations of the Kubios parasympathetic and sympathetic indices with the highest densities of EDA from the EDA Explorer. The network developed in Muñoz-Saavedra et al. (2023) was originally trained with data obtained with a different sensor. Although with the help of the author we preprocessed the data from the E4 wristband, we were only able to obtain three values for the results (low, medium, and high), we assigned the numerical values 2, 5 and 8 to these results. Having only this three different results means that standard deviations of the estimations tend to be either very low or very high.

As some recent research works show (Giggins et al., 2021) that fourth generation and later Apple watches are more accurate than the certified medical device Empatica E4 when measuring heart rate values, we also performed some preliminary analysis of data from some students who wore their own watches during the experiments and voluntarily provided their Apple health data. Although the results are promising, we did not include them in the study, as these devices do not provide GSR data. Nevertheless, promisingly accurate results point towards the usefulness of these analytics for a wider public, with possible real-time implementations in the near future.

### 3.3 | Intervention design

#### 3.3.1 | Intervention structure

The main objective of this study is to compare students' presuppositions and perceptions of different teaching methodologies, with and without LLMs integration, and to verify how accurate these assumptions are when compared to a wider range of measurement techniques based on more rigorous metrics. For this purpose, a pedagogical intervention was designed based on three different learning strategies: Traditional classroom, Flipped learning, and AI-supported gaming. For the analysis of the self-assessment data, pre-study and post-study measurements were taken, as well as a test to verify knowledge acquisition with each different class of the intervention. Physiological data was measured on a subset of participants throughout the intervention. The outline of the intervention design can be seen in Figure 3.



The study follows a pretest/posttest approach, in which students were given an initial test to evaluate their initial state and assumptions. As the study pursued not only to analyze the responses of the students to the particular methodologies but also to compare their perceptions of the methods to actual measurable data and their test scores; we used a Likert-based survey as a prestudy benchmark.

The study was carried out by dividing the 33 participating students into five randomly created subgroups of similar size. This follows a single-group intervention design with subgroup counterbalance to compensate for possible activity-order bias. Each subgroup performs each different learning activity and a posterior test of similar difficulty to evaluate their academic performance, together with a survey based on Self-Assessment Mannequins (SAM) (Bradley & Lang, 1994) to assess their emotional state. Each of these subgroups performs the activities in a random order. All lessons lasted between 30 and 40 min, so that students could be at the peak of their attention and engagement threshold. Between each test and the subsequent class, a 15-min break was implemented to avoid overloading the participants (Hlas et al., 2019).

After the three methodologies were completed, the students performed an evaluation by answering an AHP-based pairwise comparison web-based questionnaire for both preferred learning criteria and preferred learning methodology alternatives.

### 3.3.2 | ChatGPT class-integration models and class design

From the three different lessons, the traditional learning methodology was used as a well-researched baseline. As with the other “lectures”, we developed a music-based class that required no prior knowledge of the subject and that was based on nontechnical musical aspects (history, musicology, social studies). We did not present more technical aspects, such as singing or music production/technology. The class was taught by a professor of music history and consisted of a classic masterclass followed by relevant questions from the audience. Thus, this lesson followed the classical approach to teaching found in most Higher Education institutions, that is, lectures commonly supported by the overly present PowerPoint or blackboard explanation.

The second lesson, and the first active learning methodology, was designed around the Flipped Classroom. This lecture topic was based on not widely known facts of famous music pieces. Students received highly edited and well-prepared video material to watch and understand by themselves. Later, they were met in class by a professor specialized in music composition with whom they commented on the facts that were more complicated or interesting to them through a group reflection, clarification, and analysis following a similar approach to Kay et al. (2019).

As a last lesson for comparison, an active learning lesson was designed following the principles of the gamification and integrating the use of an AI LLM based tutor. For the development of this class, the models of Hamari et al. (2014) and Zainuddin et al. (2020) were taken as a reference. Based on them, aspects such as challenges, levels, clear objectives, and story/theme were incorporated into playable aspects to provide the highest level of immersion. This was the most interesting class from the educational theorist point of view, as we used ChatGPT 3.5 by OpenAI (Brown et al., 2020) to develop a mystery solving game in which students had to rescue a famous music singer traveling through time and interviewing several composers from different time periods. The approach takes into account the challenges related to bias, accuracy and safety presented in Murgia et al. (2023) and the ideas to improve active learning LLM-assisted education presented in C. Y. Lai et al. (2023). In this particular lesson, the language model is used as a companion and tutor for the student. The student can ask the AI-based tutor any necessary questions to solve the proposed riddles and tasks. As such, ChatGPT provided useful information to the student, highly tailored to the specific asks, and way beyond what a web search would directly provide. Nevertheless, as indicated in Murgia et al. (2023), when developing an application of ChatGPT for teaching and learning, we faced a series of challenges, most importantly the lack of reliability of interactive AI-based language models (Bašić et al., 2023) sometimes giving erroneous or non-relevant information. During the process of creating this lesson, we observed that even when presented with very specific questions, a difference in the way students phrased these questions on ChatGPT could dramatically alter the response they received. Furthermore, as the students develop a conversation with the assistant throughout the class, it was clear that they could induce the AI into states in which it just repeated the same erroneous information in different variations. As we strived for a gamification model where the information available to the student was useful and trustworthy, we decided that the implementation of “checkpoints”, where group decisions had to be made, was a good solution for this challenge. The solution was to implement ChatGPT as an interactive companion (time-traveling robot detective in this particular narrative, to be specific) that can help our students-detectives by giving them any information, translating, or solving any problem they required. This probed to create a much more immersive experience for the students, as they feel like they are talking to an actual robot, as opposed to just a common web search. As the mystery investigation progresses, students are presented with moments in the story when all the information recollected has to be shared among the participants and a global decision had to be achieved. By doing so publicly in front of the teacher/facilitator, there is a safeguard that false information has been pruned out, and therefore a corpus of trustful knowledge, common inside the group, is created and serves as basis for future lessons, assignments, tests, etc.

To summarize the class design, students were exposed to a gamified class that occurred in three different steps that can be repeated any number of times. First, a narrative is given to students in which a problem of their interest, related to the subject and adequately characterized, was presented for them to solve. A second phase occurs in which the students individually use their knowledge and intuition to arrive at a solution, together with the assistance of their AI tutor and companion LLM by asking it any questions that they consider relevant. In a third stage, students put their suggested solutions in common and together with the teacher/facilitator (who may intercede in disclosing additional information,

helping solve problems, etc.) come to the correct answer. At this particular stage false information is pruned, the corpus of “accepted” knowledge by the class is created and the class can advance to the next part of the gamification, solving a new problem, puzzle or riddle, and developing their knowledge of the subject while having the certainty that the information that they have up to that point is trustworthy.

When creating this gamified lesson using an AI-based language model as a companion and guide for students, we followed the Kolb learning cycle for experiential learning (Kolb et al., 2014), with each phase of the cycle corresponding to a phase of the process, as shown in Figure 4. This approach to developing and analyzing experiential-based learning has been widely used, in particular, in the field of music education, as shown in Russell-Bowie (2013), but it remains still a challenge to implement these theories in the subjects of non-technical music (which, as music history, tend to be imparted in non-experimental curricula based on masterclass methodology). Although the Kolb cycle can be entered from many points, the natural sequence for our lesson started with the active experimentation, where students, tasked with a particular riddle or question related to the subject, asked the language model questions and, using different prompts, experimented with the model until they found the correct prompt or prompting strategy and can pass to the next step. As students receive and evaluate the answers given by the AI tutor, they can start to have a concrete experience of the gamification-based solution, as they no longer are searching blindly but now have a clear strategy and can find the results. The next step, reflective observation, was implemented as a dual-layered level, as students not only reflect on the information they gathered individually but also share and discuss it with the rest of the group. As a final part of the cycle, through group analysis, they conceptualize the information gathered by the group and put it to the test in the gamification riddles/puzzles/etc. This step acts as the second fail-safe mechanism, as if the information (previously shared and discussed) provided by the group is not correct, students will not be allowed into the next stage or level of the game, and so they will need to start the cycle again to look for a new answer. If the information is correct, this is added to the pool of reviewed knowledge that the student has after the lesson and serves as foundation for future lessons, tests, etc. The students can then progress to the next stage of the game/lesson and start the cycle again with a new mystery. This was found to be an engaging method of integrating ChatGPT into a class without the risk of inducing students to errors and misinformation and the results described in the Section 4 show that students responded positively to this methodology while achieving great results.

Outside of the approaches considered for the study and intervention, an alternative Flipped Learning method was designed based on the use of ChatGPT as a personal tutor. Figure 5 depicts the process and shows how an intermediate step is added to the process where we must add fail-safes to prevent AI tutors from giving incorrect or misleading information (Murgia et al., 2023), we divide the video lesson into different parts using the Edpuzzle tool (Atherton, 2018) with short quizzes between them. Students are instructed to use the AI language model to discuss possible answers and any points that need further clarification. After the quiz, if successful, the student can go to the next part of the video, and if not, they can choose whether to re-watch the part or ask the AI Tutor for extra help. After all parts are completed, the assignment is finished and a further discussion on the topic can be done in the next class. This proved to avoid biases in ChatGPT due to prompting errors and gave students a good path toward understanding particularly difficult concepts. However, we decided to keep this alternative methodology out of the actual study with subjects, AHP, physiological data, etc., and consequently we decided to leave it for future research. Therefore, the three methods

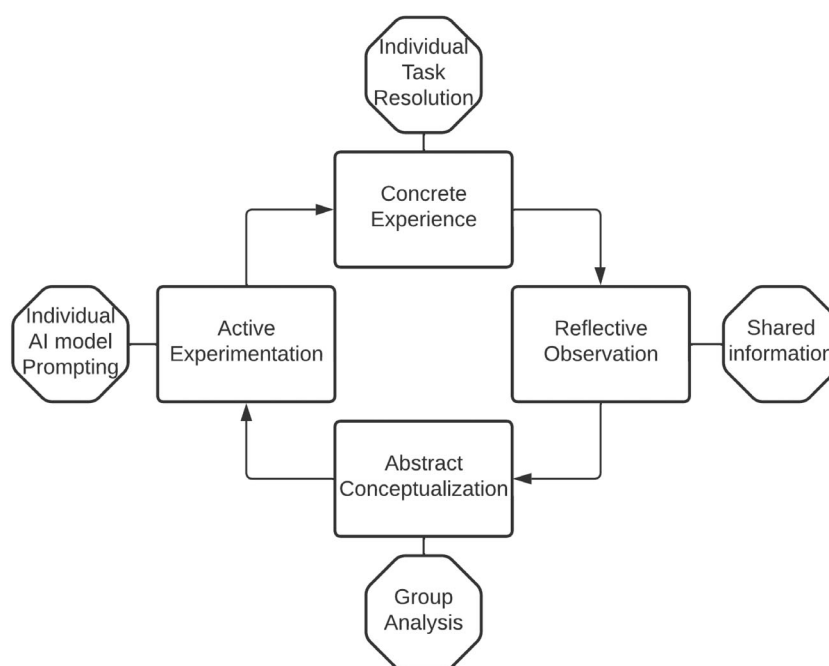


**FIGURE 4** Kolb learning cycle for AI-based gamification.

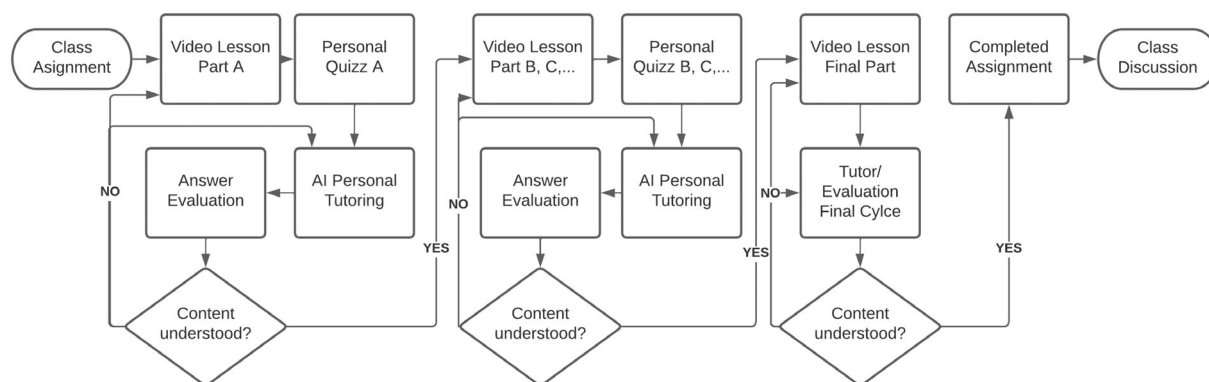


**FIGURE 5** Independent flipped learning process with AI-based tutoring.

compared in the study were a traditional masterclass style one as a benchmark, the non-AI-based Flipped Classroom as a standard active learning strategy, and the previously discussed ChatGPT-Based Gamification.

## 4 | RESULTS

In this section, the results obtained in the different parts of the experiment are analyzed, as well as their validity and the relations between different results from different measuring tools. A further exploration of these relations and relevant conclusions from the cross-study can be found in the Discussion (Section 5).

### 4.1 | Sample description

The study had a total and initial sample of 45 participants from which relevant information was obtained for the study. These students were divided into 3 equally divided different groups to create a counterbalanced study. Each of these groups was assigned the lessons in a different order to avoid biases related to lesson order, order in which the information is provided, tiredness of the students, etc.

The final sample size, reflected in the tables in the Section 4, consisted of a total of 33 students who met all the aforementioned criteria in Section 3.1, as well as completed all phases of the study. The group population consisted of 48% female and 52% male students, with ages between 18 and 24 years, and all enrolled in undergraduate university-level studies. These students all pertain to the same “campus” with all being students of Social Sciences and Humanities courses. The undergraduate program distribution and course distribution was intended to collect a wide population that was very even among courses and programs, to gather a more generalized viewpoint. The distribution among undergraduate programs were: 33% students of Arts & Drama, 30% students of Communication, 36% students of Education. The distribution by academic year was: 24% from first year, 27% from second year, 24% from third year, and 24% from fourth year.

All students were asked to self-report if they needed any extra time or required any adaptations and all reported to have no special needs or requirements. Moreover, all students had already completed at least a semester of their undergraduate education, which provides us with a relative unintrusive measure of their minimum competence and knowledge of their own studies.

### 4.2 | AHP results

To gather the perceptions of the students and draw relevant comparisons, we conducted a multicriteria AHP survey as the final part of our experience with the students to explore their perspectives on various methodologies and the weightings they considered appropriate for each criterion. Initially, we asked students to evaluate the importance of four different criteria in assessing the selection of the class methodology. The consistency ratio (CR) of the group was 0.2% and more than 90% of the students had individual CRs under 10%. The individual and group consistency ratios are more than adequate to obtain significant results.

As presented in Table 2, students placed a significantly higher value on criteria related to teaching-learning experiences, such as perceived learning experience and class engagement, than on those related to test experiences and academic results.

**TABLE 2** AHP criteria group results  $S^* = 56.1$ .

| Objective                   | Criterion                  | Weights global |
|-----------------------------|----------------------------|----------------|
| Class methodology selection | Greater perceived learning | 35.40%         |
|                             | Greater class engagement   | 31.40%         |
|                             | Less anxiety during tests  | 16.10%         |
|                             | Best grades                | 17.10%         |

**TABLE 3** Class methodology selection with equal criteria weight distribution.

| Objective                   | Criterion                  | Weights | Alternatives |         |         |
|-----------------------------|----------------------------|---------|--------------|---------|---------|
|                             |                            |         | Traditional  | Flipped | ChatGPT |
| Class methodology selection | Greater perceived Learning | 25.00%  | 0.417        | 0.258   | 0.325   |
|                             | Greater class engagement   | 25.00%  | 0.215        | 0.245   | 0.54    |
|                             | Less anxiety during tests  | 25.00%  | 0.329        | 0.31    | 0.361   |
|                             | Best grades                | 25.00%  | 0.447        | 0.27    | 0.283   |
|                             |                            | 100%    | 35.20%       | 27.10%  | 37.70%  |

**TABLE 4** Class methodology selection with student's criteria weight distribution.

| Objective                   | Criterion                  | Weights | Alternatives |         |         |
|-----------------------------|----------------------------|---------|--------------|---------|---------|
|                             |                            |         | Traditional  | Flipped | ChatGPT |
| Class methodology selection | Greater perceived learning | 35.40%  | 0.417        | 0.258   | 0.325   |
|                             | Greater class engagement   | 31.40%  | 0.215        | 0.245   | 0.54    |
|                             | Less anxiety during tests  | 16.10%  | 0.329        | 0.31    | 0.361   |
|                             | Best grades                | 17.10%  | 0.447        | 0.27    | 0.283   |
|                             |                            | 100.00% | 34.45%       | 26.43%  | 39.11%  |

After establishing the weightings for each criterion, we proceeded to survey the students to determine which class methodology they believed performed better for each specific criterion. Table 3 shows that when all criteria were considered equally important, the students attributed a similar value to both traditional master class and gamified methodologies for a non-technical music-related class, albeit with a small 2.5% preference for the gamified approach. However, when we applied the weight distribution of the criteria established by the students themselves, there was a clear preference to value gamification higher than the traditional approach, as represented in Table 4. In both scenarios, the classes based on the flipped classroom method received the lowest scores.

#### 4.2.1 | Subgroup clustering

Collected data indicate a relatively high degree of homogeneity ( $S$ ), due to sufficient overlap amongst the samples, which enables us to consider the results properly. However, the consensus indicator for the AHP group ( $S^*$ ) did not reach the desired level. This outcome could have been anticipated as we evaluated a group of students outlining their expectations and priorities for nontechnical music education, with a wide-ranging opinion pool. To address this problem, we used the Goepel (2018) method to classify the group into three clusters, which resulted in a significantly higher consensus. Figure 6 shows how the largest group with 69% of students preferred the previously mentioned class criteria, while the two minority subgroups greatly preferred experiencing less anxiety during tests (subgroup 2 with 16% of the group) and obtaining the best possible grades (subgroup 3 with 13% of the group) as shown in Figure 7. One student remains unclustered as an outlier. The consistency ratio of the surveys performed is below 0.2, which indicates that the participants responded to the questions in a logical manner; suggesting that they took the proper time to reflect on their beliefs and assumptions before answering. Therefore, we can have a reasonable degree of confidence in the validity of their responses.

Clustering is not only useful for recognizing the different subgroups that have strong preferences with regard to the learning criteria. An analysis of the alternatives evaluations (class methodology selection) can be made using the whole group perceptions on methodology, while still

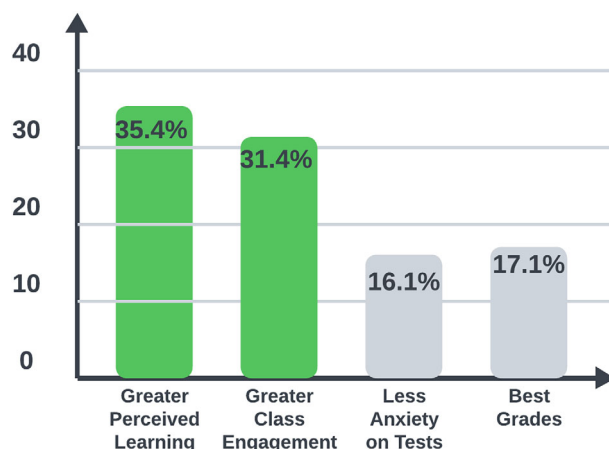


**FIGURE 6** Clustering subgroup 1,  $S^* = 75.9$  (69% of the participants).

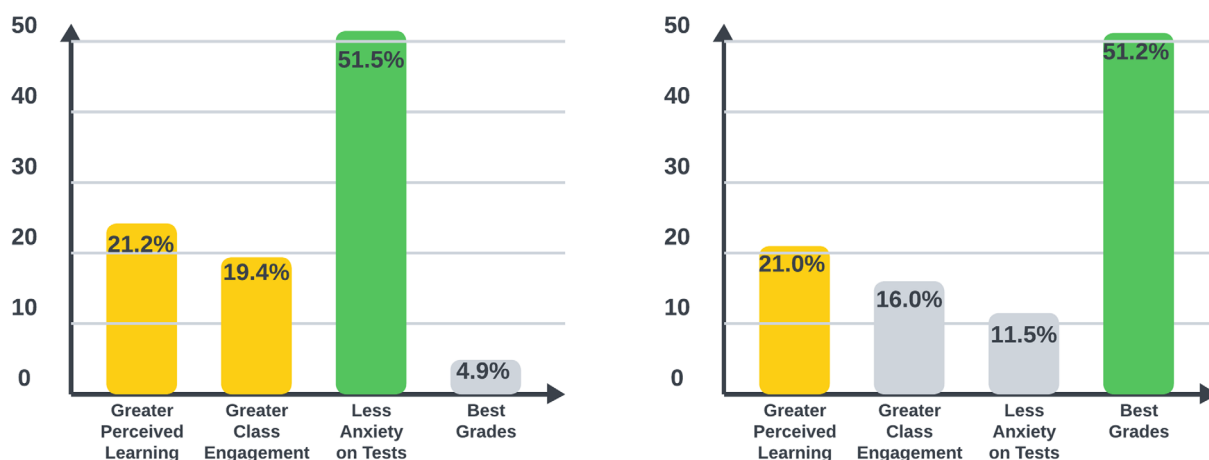


**FIGURE 7** Clustering subgroups 2  $S^* = 84.8$  (16% of the participants) and 3  $S^* = 89.2$  (13% of the participants).

**TABLE 5** Class methodology selection with Cluster 3 criteria weight distribution.

| Objective                   | Criterion                  | Weights | Alternatives |         |              |
|-----------------------------|----------------------------|---------|--------------|---------|--------------|
|                             |                            |         | Traditional  | Flipped | Gamification |
| Class methodology selection | Greater perceived learning | 21.30%  | 0.417        | 0.258   | 0.325        |
|                             | Greater class engagement   | 16.00%  | 0.215        | 0.245   | 0.54         |
|                             | Less anxiety during tests  | 11.50%  | 0.329        | 0.31    | 0.361        |
|                             | Best grades                | 51.20%  | 0.447        | 0.27    | 0.283        |
|                             |                            | 100.00% | 38.99%       | 26.80%  | 34.20%       |

accounting for one specific group criteria preference. In this way, we can have an evaluation by the whole group, while still accounting for minorities preferences. Interestingly, when considering the distribution of the weighting criteria of the subgroup in group 3, which represented a minority of 13% of the students, the traditional methodology was the clear preference, as shown in Table 5. However, it should be noted that even for this group, the ChatGPT-based gamified methodology was considered a well-regarded alternative. The surveys reveal that in this specific area of study, students placed greater value on their classroom experiences than on their test results. Consequently, they preferred methodologies that emphasized active participation in classroom and on-site learning experiences, rather than the flipped methodology, which is based on independent study outside the classroom environment.

### 4.3 | Academic performance

Table 6 presents the findings on the academic performance of each methodology. The results indicate that the traditional approach of a master class followed by student questions performed significantly better in factual grades compared to the other two methodologies. A note for consideration; due to the selected subject and knowledge area, a non-technical music lessons that did not require previous knowledge, the academic performance in this study is greatly identified with memorization and the ability to recall and reproduce data. As further indicated in the limitations of the study Section 6.2, results may vary when applying this study to a logic-based subject. This topic may require future research to be generalized.

Analysis of variance (ANOVA) was conducted on students' perceptions of their outcomes and their actual academic performance, revealing that the differences between the methodologies were statistically significant in both instances ( $p < 0.01$ ). In particular, students' perceptions of the effectiveness of the traditional methodology in achieving high grades, determined in the AHP surveys (Table 4) were consistent with their actual academic performance.

### 4.4 | Stress/engagement estimations

Table 7 displays the average self-reported levels of engagement and stress among all students who participated in the experiment, as measured by SAM during classes and tests, respectively. The SAMs used in the study employ five different pictorial representations, ranging from a scale of 1–9, where odd numbers correspond to each manikin from lowest to highest rating, and even numbers correspond to intermediate values. Anova analysis of the students' self-assessment data shows that moderately higher stress values for the gamified intervention are statistically significant ( $p = 0.004$ ) while the higher engagement reported is not statistically significant.

The physiological data was captured using the Empatica E4 wearable device and is presented in Table 8. By comparing the self-assessment data on the levels of engagement collected from students who also provided physiological data, the findings suggest that the active learning methodologies were more engaging for students. Anova analysis performed on these data shows statistical significance with  $p = 0.15$ . Even though this analysis corresponds to a small set of students, the fact that we get a statistically significant advantage for the gamified option together with the additional fact that in self-assessments, either for the whole population or for the students that participated in physiological monitoring, this option always get better engagement results clearly suggests that this engagement improvement is real. A larger study will be necessary to provide better support for this statement. According to the self-assessment manikin (SAM) data, the mean level of engagement was 8.0 for the gamified and flipped classes, while the traditional approach yielded a mean level of engagement of 7.6. The Kubios HRV and EDA explorer were used to analyze physiological data and revealed marginally better (but statistically significant) outcomes for the gamified class over the flipped class. The machine learning system regarded all results as almost equal. Upon analyzing the stress-related data, it was observed that the students reported marginally lower stress levels during the flipped test, followed by the gamified test and the traditional approach, according to their self-assessments. When examining the measured data with Kubios and EDA explorer, identical results were obtained. However, the machine learning approach favored the ChatGPT-based gamification test, but regarded the stress levels in the other two scenarios as equal.

**TABLE 6** Student test scores.

|             | Mean | SD |
|-------------|------|----|
| Traditional | 81   | 13 |
| Flipped     | 67   | 16 |
| ChatGPT     | 64   | 23 |

**TABLE 7** Self-assessment mean engagement and stress.

|             | Engagement (Class) |     | Stress (Test) |     |
|-------------|--------------------|-----|---------------|-----|
|             | Mean               | SD  | Mean          | SD  |
| Traditional | 7.2                | 1.3 | 4.0           | 2.2 |
| Flipped     | 7.5                | 1.4 | 4.0           | 2.2 |
| ChatGPT     | 7.6                | 1.5 | 5.5           | 1.8 |

**TABLE 8** Analytics physiological data results for stress and engagement.

|                   | Session       | 1 | 2 | 3 | 4 | 5 | Mean | SD   |
|-------------------|---------------|---|---|---|---|---|------|------|
| Traditional class | Engage SAM    | 7 | 9 | 7 | 7 | 8 | 7.6  | 0.89 |
|                   | Engage HR/EDA | 6 | 7 | 6 | 6 | 7 | 6.4  | 0.55 |
|                   | Engage ML     | 8 | 8 | 8 | 8 | 8 | 8    | 0.00 |
| Traditional test  | Stress SAM    | 1 | 6 | 3 | 8 | 8 | 5.2  | 3.11 |
|                   | Stress HR/EDA | 3 | 7 | 6 | 7 | 7 | 6    | 1.73 |
|                   | Stress ML     | 2 | 5 | 5 | 8 | 8 | 5.6  | 2.51 |
| Flipped class     | Engage SAM    | 7 | 8 | 7 | 9 | 9 | 8    | 1.00 |
|                   | Engage HR/EDA | 6 | 7 | 6 | 7 | 8 | 6.8  | 0.84 |
|                   | Engage ML     | 5 | 8 | 8 | 8 | 8 | 7.4  | 1.34 |
| Flipped test      | Stress SAM    | 1 | 7 | 4 | 5 | 7 | 4.8  | 2.49 |
|                   | Stress HR/EDA | 3 | 6 | 3 | 6 | 7 | 5    | 1.87 |
|                   | Stress ML     | 2 | 8 | 5 | 8 | 5 | 5.6  | 2.51 |
| ChatGPT class     | Engage SAM    | 7 | 9 | 8 | 7 | 9 | 8    | 1.00 |
|                   | Engage HR/EDA | 7 | 7 | 7 | 7 | 8 | 7.2  | 0.45 |
|                   | Engage ML     | 8 | 8 | 8 | 8 | 8 | 8    | 0.00 |
| ChatGPT test      | Stress SAM    | 6 | 6 | 2 | 3 | 8 | 5    | 2.45 |
|                   | Stress HR/EDA | 5 | 7 | 3 | 4 | 7 | 5.2  | 1.79 |
|                   | Stress ML     | 5 | 5 | 2 | 5 | 8 | 5    | 2.12 |

**TABLE 9** Student expectations prior to study.

| Best grades |       |             | Greater class engagement |     |
|-------------|-------|-------------|--------------------------|-----|
|             |       |             | Mean                     | SD  |
| Traditional | 31.7% | Traditional | 6                        | 1.7 |
| Flipped     | 29.8% | Flipped     | 5.6                      | 2.1 |
| ChatGPT     | 38.5% | ChatGPT     | 8                        | 1.3 |

It is worth mentioning that for the group that participated in the physiological measuring, the self-reported values of engagement (which as mentioned match well with the data obtained from the wearable device) are higher than those of the average population while the stress levels are lower. This happens for all the methodologies analyzed.

The ANOVA analysis of the ML system data was hampered by the fact that the current version of the system only provides three values for the results. This leads to high standard deviation values that clearly condition the statistical significance of the results. It is necessary to improve this system to use it in future studies or look for alternative solutions (e.g., facial expression recognition Sailer and Schlag (2021)).

#### 4.5 | Student expectations cross-study comparison

It is relevant to compare the expectations of the students prior to intervention with the other measurements throughout the study. A prestudy assessment of the students' initial beliefs regarding different class methodologies was conducted through a survey using Likert scales to compare the expected best grades and expected levels of class engagement. The findings indicate a wide range of student understandings and show reasonably different evaluations than those of the more consistent AHP evaluation results. Moreover, these pre-study surveys allow us to have a baseline of the original perspectives that the students had on the methodologies, which among other factors, could be influenced by the many exposures to them through their academic lives.

Table 9 presents a summary of the results of the pre-study survey, which placed the gamification methodology above others in terms of grades, while the post-study comparative AHP surveys (Table 4) showed a significantly different result. This highlights the importance of using techniques that can address potential biases when using non-comparative measurement tools, such as Likert scales, to collect information on subjective aspects or those with strong emotional implications (Kreitchmann et al., 2019). Expected class engagement values were closer to the

physiological study values (both significant at  $p < 0.15$ ), but still show the lower correlation of the Likert survey to physiological measurements compared to the SAMs and other measurement tools that prompted participants to reflect on their responses.

## 5 | DISCUSSION

### 5.1 | AI-tutoring in teaching-learning methodologies and preference assessment

Commencing the study, expectations of the students were recorded (Table 9), these measurements show an initial preference for gamification. When discussing RQ1, it can be seen that students strongly believed that this AI-based gamification will be the best strategy to achieve greater engagement and obtain the best academic scores. As Table 6 shows, the students were far from their actual experience, as they achieved the best scores in the traditional methodology. Although small, this study shows how these experiences are relevant both to student and academic staff, as they inform the actual performance of the methods studied in a comparative manner, free of initial biases. As seen in Table 4 and further discussed in Section 4, the students were very good at assessing how the different strategies academically perform when exposed to actual experience and not a simple explanation and when compelled to be coherent through a method such as the pairwise comparison in AHP. The present findings align with the perspectives of several scholars who suggest that, despite their wide use in the fields of education and psychology, they may be inadequate for assessing emotional or subjective dimensions (Kreitchmann et al., 2019), especially when such factors are associated with life experiences or have a direct bearing on the lives of participants (Yuan et al., 2019). Moreover, this study further evidences the usefulness of AHP for evaluating students' perceptions Merelo Guervos (2024), in particular when teachers may want to implement different criteria weightings depending on the particular scenario; such as using the more informed weighting of an already completed academic course to compensate the perspectives of new students.

Regarding their initial assessment of the potential for engagement for each methodology, actual pre-study survey and post-study survey show a very similar picture, where gamification is way ahead of similarly ranked traditional and flipped classes. Although the results are good for the actual metrics for the flipped classroom (as seen in Table 8) students still rated it much closer to the poorly rated (and performed when measured) traditional class. This effect clearly requires further investigation, but we theorize that the ability of students to gauge emotional responses is much more influenced by previous experiences, in particular, when they are uncomfortable. This points towards the importance of rigorously applying active learning methodologies, as bad experiences could consequently affect the emotional impact and effectiveness of future exposures to a similar methodology.

Furthermore, as can be observed through this investigation, students showed a preference for criteria and methodologies associated with classroom experiences mediated by an instructor. Although the flipped classroom approach has a significant role within the classroom, where the teacher explains the content and discusses important topics with the students, it requires an initial phase led by the student, which, in the absence of adequate levels of motivation and self-regulation, can result in a sense of isolation during the learning process. From a sociocultural perspective, where interactions between students and instructors or between students themselves are essential for the learning process, the flipped classroom approach may warrant further revision (Glăveanu et al., 2019; Cunningham, 2023). To address the aforementioned issues or, at the very least, stimulate further discussion, this study presents a model of flipped classroom which is assisted by an Artificial Intelligence-based tutor. This model uses ChatGPT, which serves as an AI personal tutor (Luckin et al., 2016), due to its natural language communication capabilities. ChatGPT could accompany the student throughout their learning process, preventing them from getting stuck due to lack of understanding, uncertainty regarding the next steps, and other learning-related challenges. The AI's ability to interact with students in a simple and direct manner can also help mitigate feelings of isolation or loneliness, thus improving students' well-being and improving their receptiveness to learning. As such, the surveys collected in the results Section 4 clearly support the idea that bidirectional communication (in the implemented case of gamification with ChatGPT), through an AI assistant, can positively impact students' perceptions of their education. This finding aligns with the idea of learning by imitation, where conversation and exchange of ideas form the basis of learning (Salas, 2001). It can well be a case similar to e-reading, where the innovation of tactile screens allowed for a more natural interaction; and where LLMs, through conversation, allow for more successful models of Flipped classrooms. Although this model was not tested in the field study, it theoretically tackles some of the issues in the methodology highlighted by the study and, therefore, was included in the method 3 and further research is required on its validation and implementation.

These findings are consistent with previous studies that emphasize the need to carefully examine the assumptions underlying active learning and its consequences for pedagogy (Carpenter et al., 2020; Deslauriers et al., 2019). Although active learning approaches appear to enhance student motivation and engagement, such benefits do not invariably result in better academic performance or, as observed in this investigation, in student preference. With the results of the AHP survey showing little preference for the out-of-class experience of flipped classrooms, and with academic results for both active learning methods rating below the traditional class. The introduction of novel measurement tools that mitigate subjective biases, along with the gradual integration of learning analytics into educational research (Pelletier et al., 2022; Liu et al., 2022), offers new opportunities to refine teaching and learning strategies deployed in educational institutions, by improving our understanding of the intricate processes at play in the classroom.



Regarding RQ2, we first need to discuss the different values students give to perceived learning and academic results, both at a criterion level and for each individual alternative in the AHP survey. In Table 2 we can see how students in general valued much more highly their experience of learning, that is, how much they felt they were learning, rather than the academic scores they receive for a given methodology. As with emotional criteria, students seem to value learning experience over results, and the in-class criteria seem very relevant to them compared to the exam or external criteria. Such is a logical stance, as they spend a great part of their academic experiences inside classes and therefore take this experience into great consideration. Regarding each particular class methodology, Table 4 shows how for all methodologies students rated their perceived experience similarly to their actual grades with the interesting exception of gamification, which was the only one in which perceived learning was rated higher than expected marks.

Regarding RQ3, the particular case of the gamification was, in our experience, due to the higher complexity of the actual gamification based on ChatGPT method, where the class, as shown, is highly engaging, but the information is not spoon fed to the students and therefore it may be harder for them to predict which part of the information is relevant to the actual examination (even if most of it is relevant from a learning point of view). Students feel that the more they learn, the better the grades, and their test results are fairly consistent with their perceptions for each individual class. It is clear that more straightforward classes, where the teacher may even underscore which information is more relevant towards the examination, give a better correlation between perceived learning and academic results, but not necessarily better overall results on any of these fronts. Although the results of this study may not provide a definitive answer to the debate, with the preferred and best rated class by engagement (ChatGPT Gamification) not having the best academic results, they are consistent with previous literature that highlights the need for caution when implementing gamification in educational settings (Thibault & Hamari, 2021). The authors warn against the risk that students mistake gamification for pure entertainment and remark that successful use of gamification in the classroom requires proper training and maturity on the part of students (Queirós & Pinto, 2022). To maximize the effectiveness of disruptive methodologies, particularly those based on games, it is important to gradually introduce them and incorporate them into the educational system from an early stage, allowing their full potential to be realized over time (Bekk et al., 2022). The proposed implementation of ChatGPT is not different; highly engaging, but future teacher and developers should take into consideration the need for proper introduction and the necessity for clear directions from the faculty. Participants in the present study had limited prior exposure to these methodologies, which could partially explain the results obtained. These observations highlight the possibility of extracurricular competencies that could be gained from the implementation of ChatGPT gamification. With good engagement, not bad academic results, especially when compared to the other active learning methodology, and with the best rating by students, this strategy may allow students for greater socialization, better in-class experiences, and may actually tackle motivation issues. Furthermore, complexity is in part caused by the extra burden of user interaction with the technology, which increases students' knowledge of the STEM in a practical manner. This extra difficulty may call for the design of better, education-oriented LLMs, which would greatly decrease the Human-Computer interaction barrier.

Regarding the importance of the different proposed criteria the whole group considered (see Table 4), as well as the majority cluster, that perceived learning and class engagement are much more important than anxiety during tests and best grades. This opinion is very clear inside the first cluster, which represents 69% of the students, since the results of their pairwise comparisons in AHP assign weight factors higher than double for the first two criteria than for the other two. Using the criteria weighting for the whole student group, or the one of the majority cluster, the best valued methodology would be that based on a gamified approach followed by the traditional class technique. However, when a greater weight is given to the actual academic results, as happens with students included in cluster 3 and probably with many higher education institutions, the traditional class methodology would be the selected option, as can clearly seen in Table 5. It seems quite clear that both traditional classes and the ChatGPT-based gamified methodology are well valued by students (which again positively validates RQ3). Traditional classes are primarily valued due to their impact on grades and perceived learning, while the gamified approach is primarily valued to provide greater class participation while not performing poorly on neither anxiety nor perceived learning. Probably, a mixed approach using gamification in those topics where the lecturer knows that student engagement is lower could be an alternative worth of further research. In our study, neither the global group nor any of the clusters seems to favor the flipped-class methodology; some hypotheses have already been discussed previously in this section with the aim of fostering critical inquiry and advancing our knowledge of the aforementioned phenomenon.

When focusing on the implementation of the ChatGPT-based gamification learning model, it was already discussed in Section 3 the necessity for the introduction of checkpoints that allow us to ensure that the information given by the AI is accurate and not misleading. Regarding RQ2 and the differences in the perception of learning for the ChatGPT-based methodology compared to their actual results, Table 6 shows that the students performed similarly to the other active methodology, while Table 4 shows how they had greater engagement, good perceived learning, and went confidently to the exam, having less anxiety during the tests. The capability to create new models for active learning using the new AI technologies is only going to increase and improve in coming years and is a field that requires further research, but are already providing good alternatives for highly engaging active learning classes, capable of a high order of autonomous learning with the support of the AI and which offer good academic results. Both the implemented gamification and the proposed flipped classroom models, which were based on the use of ChatGPT, can serve as starting points for designing many different iterations of classes that follow these methodologies while safeguarding the integrity of both students and academic institutions by corroborating that no false information is passed to the student at the end of the process.

## 5.2 | Learning analytics for education

It can be seen how surveys, particularly in the field of education, can leave behind some minority groups with strong feelings, opinions, or idiosyncrasies. Comparing the general group criteria in the Table 2 to Figure 7 we can see how the general survey is reasonably close to the opinion of the majority (subgroup 1) but very far from the other two minority groups. Tools such as the clustering one used in this study are a very needed addition to the researcher/teacher, as otherwise, we are at a very high risk of losing diversity in classes and curricula due to ignorance of the existence of these minority groups. Through our research, we believe that this is one of the major advantages of the proposed tools for evaluating different learning methodologies. In particular, in a scenario where student dropout rates are increasing, involving students in the decision-making processes and design of their own curriculum is a strategy worth considering. These “simple” pairwise comparisons, together with clustering tools, provide the necessary tools for self-determining the best method or combination of methods for a given subject and a particular group of individuals while avoiding natural biases associated with student making overly emotional decisions.

The SAM data in Table 7 matches well the selection of the AHP-based methodology shown in Figure 3 as the selected gamified methodology was perceived by the students as the approach with the highest average engagement, the traditional approach was the second alternative both for the AHP and SAM-based engagement ranking and the flipped classroom was perceived as the poorest choice in both cases.

The physiological data analysis with Kubios and the EDA explorer has given results with good correlations with the students' self-assessments. In the case of the ML analysis, as the system was trained with data from very different sensors, we had to reduce the output of the system to three levels (low, medium, high); this means that any output that could really belong to any intermediate level can have very significant influence in the results. In the current version, the analysis by Kubios/EDA explorer has provided more accurate results than the ML system. In the future, we will retrain our ML-based system with Empatica E4 data to improve the quality of the predictions produced by these systems. These physiological data, cross-examined with the self-assessment results, constitute a foundation on which decisions regarding the curriculum of Higher Education Institutions and the learning methods decided for individual subjects can be made with greater information and certainty. Further research needs to be made (and is being done) on the implementation of such systems in real-time. Such tools would open many possibilities in regards to their implementation in active learning methodologies and in particular in ones that could make use of said data to inform the AI tutors that could guide, modify and collaborate in these active learning models.

## 6 | CONCLUSIONS

The need for new curricula in Higher Education Institutions that can adapt to the particular needs of a group of students, or even to the individual, is clear. Active learning traditionally proposes engaging experiential learning, where students can feel more connected to the subject. Through our study, we find out how, even though such engagement is usually present, there is a risk for the methodology to not perform to a standard or, as was the case with the Flipped method, students may not particularly like it.

The AHP method, together with clustering tools, proved to be very useful in determining the preferred class learning method by the student. It not only avoided biases, but helped signal which minority groups may have a strong identity or preference that could otherwise be overshadowed by the majority. This is particularly relevant when treating with students that have special requirements, who in many cases are under-reported. In the particular case of students with so-called hidden disabilities (Wolf, 2001), these clustering and analysis tools help us to better identify and address the needs and preferences of these groups. If we want our campuses to be more accessible and reduce dropout rates, the use of tools such as these is incredibly useful. Furthermore, the possibility to have statistically consistent results with a small population allows the method to be used for student self-determining their learning strategies, making content decisions on their curriculum, etc. This self-determining path where students are equipped with analytical tools and are given the opportunity to form part of the institution decision-making process is an area that we believe could yield great results but requires further research with a larger pool of students through a longer period of time.

Other learning analytics are based on collected physiological data. From the small collected subset of physiological data, we can see, although this clearly needs more empirical corroboration, that the physiological data obtained from the Empatica wristband can be used as an estimator for the students' engagement during classes and their stress during the tests. This contributes not only towards the analysis of different dynamics in the class but to a future where wearable technology can be used to monitor the state of individual students in such a way that can inform the teacher of different options and even adapt in real-time user-specific feedback suited for each learning individual. This is particularly useful for AI-Based active learning, where data could be used by the AI to automatically tailor the experience to the individual user.

Our approach to generating possible models to integrate ChatGPT into active learning-based classes was supported by user evaluation (as can be seen in Table 4) and the actual user engagement experience and results. It paves the way towards new models that can integrate newly developed AI tools minimizing the risk they possess to students to a minimum, and open the possibility for further implementation with real-time learning analytics.



As a final thought, throughout our study, both at a criteria evaluation level and as an alternative evaluation level (3 & 2) those students preferred class group experiences over remote or individual ones. They valued greater criteria related to the class (engagement and perceived learning) than those that offer results beyond the scope of the class, and they negatively valued the flipped learning methodology, which is arguably the most individualistic and requires more time outside the classroom. Although it clearly requires further research and a larger student pool to make the results more relevant, let it serve as a warning in a post-pandemic time where remote independent learning is becoming more and more ubiquitous.

## 6.1 | Future work

This study point towards the preference of students for in-class experiences over autonomous learning. Determining whether this is applicable across many study areas, as well as for different students with special needs, requires further study with a wider, cross-cultural sample pool. Understanding the implications of the application of Large Language Models for students with disabilities or those who require additional support or tutoring is a very interesting and important area of study, and the expanding of this research to include differently categorized groups can help develop more inclusive and accessible software.

Furthermore, our study points towards a barrier of human-computer interaction for LLMs integrated in the classroom. Although highly regarded by the students, separating the gamification methodology itself from the LLM integration could eliminate a confounding variable and offer additional insights on this interaction barrier. Nevertheless, it is clear that the development of specific LLMs for education is an area that requires future research both from the academia and the industry.

The development of specific educational LLMs would greatly help teachers in avoiding biases and misinformation. The use of methodologies and structures such as the gamification and flipped classroom strategies proposed in Section 3.3.2 could be integrated in these tools, allowing teachers to select specific knowledge that is available at a given time and helping student develop their skills and knowledge at a consistent pace. This integration of educational methodologies and its implication for decreasing the human-computer interaction barrier in teaching-learning environments. This is a useful and interesting emerging area with implementations starting to become available such as Sibilio et al. (2023) where an actual human professor is virtually modeled which can facilitate the interaction. Nevertheless, this area is very novel and requires further research and development of tools.

Moreover, as previously discussed, this study focus on a non-technical arts & Humanities subject (Music history). Although the proposed implementation of both the study results as well as the methodology for ChatGPT integration in class could be easily done in related areas, further study and examination may be required to extend these strategies and results to knowledge areas related to STEMs, where technical background, logic and other specific skills may be very relevant. These areas could greatly benefit from the implementation of AI tutoring in Active Learning, but further inquiry can be required to apply it.

Finally, the impact of LLMs as personal AI tutoring devices in students' self-learning methodologies is an area that requires future work. In Section 3.3.2, we presented a possible integration of ChatGPT in a flipped classroom methodology. Our hypothesis is that this types of integrations not only provide students with feedback and guide, while avoiding false information; but also can improve their motivation and aid with their self-regulation in autonomous scenarios. Nevertheless, this strategy was not included in the field study, and its implementation and impact on engagement levels requires further investigation.

## 6.2 | Limitations of the study

This research represents a limited study on the subjects of learning methodologies and their specific relations to academic results and perceived learning that we intend to further expand with a wider population in the future. Although the use of the AHP instrument is a great tool for small group testing due to, among other qualities, the built-in consistency analysis and the saturation of information per respondent (Melillo & Pecchia, 2016), further information can be consulted in Section 3.2.1. This makes the validation of the ChatGPT integration, as well as the measures taken, very reliable. However, conclusions extracted from the data in relation to learning methodologies and their own idiosyncrasies related to learning criteria have limited empirical support when applied to a wider global population due to the limited size of the study. The subject of the study requires further research to be able to generalize the conclusions further. Furthermore, this research was carried out with the participation of "normative" students, with no special needs and who all have successfully completed some academic subjects. As suggested in the previous future work Section 6.1, this study should be implemented attending to different student profiles and we advise not to carelessly generalize its results for populations with different needs, such as students with ADHD or very non-average IQs.

A close identification has been made in this study between the ability to recall data and academic performance. Although relevant in the context of the study (music history), we acknowledge that this can be a reductionist approach and further study is required to be implemented in logic-based areas. A point for consideration is the content of the lessons. They were carefully crafted to all have very similar data but a broader

study should carefully consider the possibility of further counterbalancing the study with the content of the lessons, this would avoid a possible confounding variable, that we have tried to minimize, but would require to triple the sample size.

Due to limitations in the availability of wearable medical devices, the number of physiological measurements registered is reduced as only a subset of participants could be tested. The implementation of the Deep Learning-based signal analysis to obtain stress and engagement results cross-references well with the self-reported data from the SAM surveys. However, the actual triangulation of the data between AHP surveys, SAM self-assessments, and physiological data processed requires further investigation with a wider population for the results regarding class/test experiences to be generalized.

## ACKNOWLEDGMENTS

We greatly thank Dr. Luis Muñoz-Saavedra for his guidance using the AI tool Muñoz-Saavedra et al. (2023). This collaboration was of the utmost value to the researchers.

## FUNDING INFORMATION

This work was partially funded by the European Union Erasmus+ grant Teaching to Be: Supporting Teachers' Professional Growth and Wellbeing in the Field of Social and Emotional Learning (626155-EPP-1-2020-2-LT-EPPKA3-PI-POLICY). This publication is part of the project PID2022-137646OB-C31, funded by the Spanish Ministry of Science and Innovation under grant MCIN/AEI/10.13039/501100011033/ and by the European Union.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID

Miguel Civit  <https://orcid.org/0000-0003-4310-6377>

Salvador Reyes-de-Cozar  <https://orcid.org/0000-0001-9602-8988>

## REFERENCES

- Aguarón, J., Escobar, M. T., & Moreno-Jiménez, J. M. (2016). The precise consistency consensus matrix in a local ahp-group decision making context. *Annals of Operations Research*, 245, 245–259.
- Aidoo, B., Macdonald, M. A., Vesterinen, V.-M., Pétursdóttir, S., & Gísladóttir, B. (2022). Transforming teaching with ict using the flipped classroom approach: Dealing with covid-19 pandemic. *Education in Science*, 12(6), 421.
- Aina, C., Baici, E., Casalone, G., & Pastore, F. (2022). The determinants of university dropout: A review of the socio-economic literature. *Socio-Economic Planning Sciences*, 79, 101102.
- Akiha, K., Brigham, E., Couch, B. A., Lewin, J., Stains, M., Stetzer, M. R., Vinson, E. R., & Smith, M. K. (2018). What types of instructional shifts do students experience? investigating active learning in science, technology, engineering, and math classes across key transition points from middle school to the university level. In *Frontiers in education* (Vol. 2, p. 68). Frontiers Media SA.
- Al Mamun, M. A., Azad, M. A. K., Al Mamun, M. A., & Boyle, M. (2022). Review of flipped learning in engineering education: Scientific mapping and research horizon. *Education and Information Technologies*, 27, 1261–1286.
- Anis, A., & Islam, R. (2015). The application of analytic hierarchy process in higher-learning institutions: a literature review. *Journal for International Business and Entrepreneurship Development*, 8(2), 166–182.
- Atherton, P. (2018). 50 ways to use technology enhanced learning in the classroom: Practical strategies for teaching. *Ways to Use Technology Enhanced Learning in the Classroom*, 50, 1–288.
- Aull-Hyde, R., Erdogan, S., & Duke, J. M. (2006). An experiment on the consistency of aggregated comparison matrices in ahp. *European Journal of Operational Research*, 171(1), 290–295.
- Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). Better by you, better than me? chatgpt-3 as writing assistance in students essays.
- Bean, J. C., & Melzer, D. (2021). *Engaging ideas: The professor's guide to integrating writing, critical thinking, and active learning in the classroom*. John Wiley & Sons.
- Bekk, M., Eppmann, R., Klein, K., & Völkner, F. (2022). All that glitters is not gold: An investigation into the undesired effects of gamification and how to mitigate them through gamification design. *International Journal of Research in Marketing*, 39(4), 1059–1081.
- Bond, M. (2020). Facilitating student engagement through the flipped learning approach in k-12: A systematic review. *Computers & Education*, 151, 103819.
- Børte, K., Nesje, K., & Lillejord, S. (2023). Barriers to student active learning in higher education. *Teaching in Higher Education*, 28(3), 597–615.
- Bowers, J., Smith, W., Ren, L., & Hanna, R. (2019). Integrating active learning labs in precalculus: *Measuring the value added*. *Investigations in Mathematics Learning*, 11(1), 1–15.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59.

- Bressane, A., Bardini, V. S. D. S., & Spalding, M. (2021). Active learning effects on students' performance: a methodological proposal combining cooperative approaches towards improving hard and soft skills. *International Journal of Innovation and Learning*, 29(2), 154–165.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
- Brown, T. L., Brazeal, K. R., & Couch, B. A. (2017). First-year and non-first-year student expectations regarding in-class and out-of-class learning activities in introductory biology. *Journal of Microbiology & Biology Education*, 18(1), 18.
- Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students' (mis) judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition*, 9(2), 137–151.
- Castro, R. (2019). Blended learning in higher education: Trends and capabilities. *Education and Information Technologies*, 24(4), 2523–2546.
- Chang, J.-W., & Wei, H.-Y. (2016). Exploring engaging gamification mechanics in massive online open courses. *Journal of Educational Technology & Society*, 19(2), 177–203.
- Commission, E., Directorate-General for Education, S., Youth, & Culture. (2022). *Education and training monitor 2022–spain*. Publications Office of the European Union. <https://doi.org/10.2766/149993>
- Cooper, K. M., Downing, V. R., & Brownell, S. E. (2018). The influence of active learning practices on student anxiety in large-enrollment college science classrooms. *International Journal of STEM Education*, 5(1), 1–18.
- Cuadrado, F., Lopez-Cobo, I., Mateos-Blanco, T., & Tajadura-Jiménez, A. (2020). Arousing the sound: A field study on the emotional impact on children of arousing sound design and 3d audio spatialization in an audio story. *Frontiers in Psychology*, 11, 737.
- Cunningham, R. W. H. L. (2023). *Creating interactive sociocultural environments for self-regulated learning* (p. 17). *Issues and educational applications*.
- Dejene, W. (2019). The practice of modularized curriculum in higher education institution: Active learning and continuous assessment in focus. *Cogent Education*, 6(1), 1–16.
- Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., & Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proceedings of the National Academy of Sciences*, 116(39), 19251–19257.
- Divjak, B., Svetec, B., Horvat, D., & Kadoić, N. (2022). Assessment validity and learning analytics as prerequisites for ensuring student-centred learning design. *British Journal of Educational Technology*, 54(1), 313–334.
- Dong, Y., Xu, W., & Xu, W. (2013). An automatic method to reach consensus in a local context for ahp group decision making. *European Journal of Industrial Engineering*, 7(4), 456–474.
- Driessen, E. P., Knight, J. K., Smith, M. K., & Ballen, C. J. (2020). Demystifying the meaning of active learning in postsecondary biology education. *CBE Life Sciences Education*, 19(4), ar52.
- Ekici, M. (2021). A systematic review of the use of gamification in flipped learning. *Education and Information Technologies*, 26(3), 3327–3346.
- Fischer, B. A. (2019). Fact or fiction? designing stories for active learning exercises. *Journal of Political Science Education*, 15(2), 179–190.
- Fisher, R., Perényi, A., & Birdthistle, N. (2021). The positive relationship between flipped and blended learning and student engagement, performance and satisfaction. *Active Learning in Higher Education*, 22(2), 97–113.
- Fu, Y., Zhang, M., Nawaz, M., Ali, M., & Singh, A. (2022). Information technology-based revolution in music education using ahp and topsis. *Soft Computing*, 26(20), 10957–10970.
- Gašević, D., Greiff, S., & Shaffer, D. W. (2022). *Towards strengthening links between learning analytics and assessment: Challenges and potentials of a promising new bond* (Vol. 134). Elsevier.
- Giggins, O. M., Doyle, J., Sojan, N., Moran, O., Crabtree, D. R., Fraser, M., & Muggeridge, D. J. (2021). Accuracy of wrist-worn photoplethysmography devices at measuring heart rate in the laboratory and during free-living activities. *2021 43rd annual international conference of the IEEE engineering in medicine & biology society (embc)*. 6970–6973.
- Glăveanu, V. P., Ness, I. J., Wasson, B., & Lubart, T. (2019). Sociocultural perspectives on creativity, learning, and technology. *Creativity under duress in education? Resistive theories, practices, and actions*, 3, 63–82.
- Glick, D., Cohen, A., Festinger, E., Xu, D., Li, Q., & Warschauer, M. (2019). Predicting success, preventing failure. In *Utilizing learning analytics to support study success* (pp. 249–273). Springer.
- Goepel, K. D. (2018). Implementation of an online software tool for the analytic hierarchy process (ahp-os). *International journal of the analytic hierarchy process*, 10(3), 469487.
- Goode, E., Nieuwoudt, J., & Roche, T. (2022). Does online engagement matter? the impact of interactive learning modules and synchronous class attendance on student achievement in an immersive delivery model. *Australasian Journal of Educational Technology*, 38(4), 76–94.
- Gu, X., Tan, Y., Wu, X., Cai, Z., Lai, Q., Cheng, M., & Zhao, Q. (2022). Active versus passive strategy in online creativity training: How to best promote creativity of students with different cognitive styles? *Thinking Skills and Creativity*, 44, 101021.
- Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work?—a literature review of empirical studies on gamification. In *2014 47th hawaii international conference on system sciences*. 3025–3034.
- Hammill, J., Nguyen, T., & Henderson, F. (2021). Encouraging the flip with a gamified process. *International Journal of Educational Research Open*, 2, 100085.
- Hartikainen, S., Rintala, H., Pylväs, L., & Nokelainen, P. (2019). The concept of active learning and the measurement of learning outcomes: A review of research in engineering higher education. *Education in Science*, 9(4), 276.
- Hasan, L. F., Elmetwaly, A., & Zulkarnain, S. (2020). Make the educational decisions using the analytical hierarchy process ahp in the light of the corona pandemic. *Engineering Applications of Artificial Intelligence*, 2, 1–13.
- Hidayat, L. E., & Praseno, M. D. (2021). Improving students' writing participation and achievement in an edpuzzle-assisted flipped classroom. *EDUCAFL: Journal of Education of English as Foreign Language*, 4(1), 1–8.
- Hlas, A. C., Neyers, K., & Molitor, S. (2019). Measuring student attention in the second language classroom. *Language Teaching Research*, 23(1), 107–125.
- Hood, S., Barrickman, N., Djerdjian, N., Farr, M., Magner, S., Roychowdhury, H., et al. (2021). "I like and prefer to work alone": Social anxiety, academic self-efficacy, and students' perceptions of active learning. *CBE Life Sciences Education*, 20(1), ar12.
- Hu, Y., Gallagher, T., Wouters, P., van der Schaaf, M., & Kester, L. (2022). Game-based learning has good chemistry with chemistry education: A three-level meta-analysis. *Journal of Research in Science Teaching*, 59(9), 1499–1543.
- Ifenthaler, D. (2017). Learning analytics design. In *The sciences of learning and instructional design* (pp. 202–211). Routledge.

- Ifenthaler, D., Mah, D.-K., & Yau, J. Y.-K. (2019). Utilising learning analytics for study success: Reflections on current empirical findings. In *Utilizing learning analytics to support study success* (pp. 27–36). Springer.
- Ifenthaler, D., & Yau, J. Y.-K. (2020). Utilising learning analytics to support study success in higher education: a systematic review. *Educational Technology Research and Development*, 68(4), 1961–1990.
- Javaid, M., Haleem, A., Singh, R. P., Khan, S., & Khan, I. H. (2023). Unlocking the opportunities through chatgpt tool towards ameliorating the education system. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 3(2), 100115.
- Jiménez-Mijangos, L. P., Rodríguez-Arce, J., Martínez-Méndez, R., & Reyes-Lagos, J. J. (2023). Advances and challenges in the detection of academic stress and anxiety in the classroom: A literature review and recommendations. *Education and Information Technologies*, 28(4), 3637–3666.
- Johnson, L., Adams Becker, S., Estrada, V., Freeman, A., Karpilis, P., Vuorikari, R., & Punie, Y. (2014). Horizon report europe: 2014 schools edition. In Luxembourg: Publications Office of the European Union, & Austin. The New Media Consortium.
- Jost, L. (2006). Entropy and diversity. *Oikos*, 113(2), 363–375.
- Kabassi, K., & Alepis, E. (2020). *Learning analytics in distance and mobile learning for designing personalised software* (pp. 185–203). *Advances in Learning Analytics*.
- Kasneji, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103, 102274.
- Kay, R., MacDonald, T., & DiGiuseppe, M. (2019). A comparison of lecture-based, active, and flipped classroom teaching approaches in higher education. *Journal of Computing in Higher Education*, 31, 449–471.
- Kocak, O. (2022). Asystematic literature review of web-based student response systems: Advantages and challenges. *Education and Information Technologies*, 27(2), 2771–2805.
- Kolb, D. A., Boyatzis, R. E., & Mainemelis, C. (2014). Experiential learning theory: Previous research and new directions. In *Perspectives on thinking, learning, and cognitive styles* (pp. 227–248). Routledge.
- Kreitchmann, R. S., Abad, F. J., Ponsoda, V., Nieto, M. D., & Morillo, D. (2019). Controlling for response biases in self-report scales: Forced-choice vs. psychometric modeling of likert items. *Frontiers in Psychology*, 10, 2309.
- Lai, C. Y., Cheung, K. Y., & Chan, C. S. (2023). Exploring the role of intrinsic motivation in chatgpt adoption to support active learning: An extension of the technology acceptance model. *Computers and Education: Artificial Intelligence*, 5, 100178.
- Lai, V., Wong, B. K., & Cheung, W. (2002). Group decision making in a multiple criteria environment: A case using the ahp in software selection. *European Journal of Operational Research*, 137, 134–144. [https://doi.org/10.1016/S0377-2217\(01\)00084-4](https://doi.org/10.1016/S0377-2217(01)00084-4)
- Leal, J. E. (2020). Ahp-express: A simplified version of the analytical hierarchy process method. *MethodsX*, 7, 100748.
- Li, R., Lund, A., & Nordsteien, A. (2021). The link between flipped and active learning: a scoping review. *Teaching in Higher Education*, 28(8), 1–35.
- Liu, Z., Ren, Y., Kong, X., & Liu, S. (2022). Learning analytics based on wearable devices: A systematic literature review from 2011 to 2021. *Journal of Educational Computing Research*, 60(6), 1514–1557.
- Loderer, K., Pekrun, R., & Lester, J. C. (2020). Beyond cold technology: A systematic review and meta-analysis on emotions in technology-based learning environments. *Learning and Instruction*, 70, 101162.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for ai in education*. Pearson Education.
- Lugosi, E., & Uribe, G. (2022). Active learning strategies with positive effects on students' achievements in undergraduate mathematics education. *International Journal of Mathematical Education in Science and Technology*, 53(2), 403–424.
- McCarthy, C., Pradhan, N., Redpath, C., & Adler, A. (2016). Validation of the empathica e4 wristband. *2016 IEEE embs international student conference (isc)*. 1–4.
- Mehta, S., Schukow, C. P., Takrani, A., Ritchie, R. P., Wilkins, C. A., & Faner, M. A. (2022). Understanding student characteristics in the development of active learning strategies. *Medical Science Educator*, 32(3), 615–626.
- Melillo, P., & Pecchia, L. (2016). What is the appropriate sample size to run analytic hierarchy process in a survey-based research. *Proceedings of the the international symposium on the analytic hierarchy process, london, uk*. 4–8.
- Mellizo-Soto, M. F. (2022). Análisis del abandono de los estudiantes de grado en las universidades presenciales en españa. Ministerio de Universidades.
- Merele Guervos, J. J. (2024). Formative evaluation and learning analytics for agile teaching. Retrieved from <https://hdl.handle.net/10481/89811>
- Mitry, M. M. (2021). Translating constructivism into pedagogy from philosophy to practice: Active project-based learning. *The International Journal of Humanities Education*, 19(1), 39.
- Mohammed, H. J., & Daham, H. A. (2021). Analytic hierarchy process for evaluating flipped classroom learning. *Computers, Materials & Continua*, 66(3), 2229–2239.
- Muñoz-Saavedra, L., Escobar-Linero, E., Miró-Amarante, L., Bohórquez, R., & Domínguez-Morales, M. (2023). Designing and evaluating a wearable device for affective state level classification using machine learning techniques. *Expert Systems with Applications*, 219, 119577.
- Murgia, E., Abbasiantaeb, Z., Aliannejadi, M., Huibers, T., Landoni, M., & Pera, M. S. (2023). Chatgpt in the classroom: A preliminary exploration on the feasibility of adapting chatgpt to support children's information discovery. *Adjunct proceedings of the 31st acm conference on user modeling, adaptation and personalization*. 22–27.
- Murillo-Zamorano, L. R., López Sánchez, J. Á., Godoy-Caballero, A. L., & Bueno Muñoz, C. (2021). Gamification and active learning in higher education: is it possible to match digital society, academia and students' interests? *International Journal of Educational Technology in Higher Education*, 18, 1–27.
- Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2020). The impact of stress on students in secondary school and higher education. *International Journal of Adolescence and Youth*, 25(1), 104–112.
- Pelletier, K., McCormack, M., Reeves, J., Robert, J., Arbino, N., Dickson-Deane, C., et al. (2022). *2022 educause horizon report teaching and learning edition*. EDUC22.
- Prieto, A., Barbarroja, J., Alvarez, S., & Corell, A. (2021). Effectiveness of the flipped classroom model in university education: a synthesis of the best evidence. eficacia del modelo de aula invertida (flipped classroom) en la enseñanza universitaria: una síntesis de las mejores evidencias. *Revista de Educación*, 391, 143–170.
- Queirós, R., & Pinto, M. (2022). The dangers of gamification. *Advanced research in technologies, information, innovation and sustainability: Second international conference, artiis 2022, santiago de compostela, spain, september 12–15, 2022, revised selected papers, part i*. 151–162.

- Reyes-de Cózar, S., Merino-Cajaville, A., & Salguero-Pazos, M. R. (2023). Avoiding academic burnout: Academic factors that enhance university student engagement. *Behavioral Science*, 13(12), 989–1007. <https://doi.org/10.3390/bs13120989>
- Rodriguez, M., Diaz, I., Gonzalez, E. J., & Gonzalez-Miquel, M. (2018). Motivational active learning: An integrated approach to teaching and learning process control. *Education for Chemical Engineers*, 24, 7–12.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3), e1355.
- Russell-Bowie, D. (2013). Mission impossible or possible mission? changing confidence and attitudes of primary preservice music education students using kolb's experiential learning theory. *Australian Journal of Music Education*, 2, 46–63.
- Saaty, T. L. (2010). *Mathematical principles of decision making (principia mathematica decernendi)*. RWS publications.
- Saaty, T. L., & Özdemir, M. S. (2014). How many judges should there be in a group? *Annals of Data Science*, 1, 359–368.
- Sailer, M., & Schlag, R. (2021). Linking teachers' facial microexpressions with student-based evaluation of teaching effectiveness: A pilot study using facereader™. 7th international conference on higher education advances (head'21). 555–562.
- Salas, A. L. C. (2001). Implicaciones educativas de la teoría sociocultural de vigotsky. *Revista educación*, 25(2), 59–65.
- Shen, X., Xu, H., Feng, J., Ye, J., Lu, Z., & Gan, Y. (2022). The global prevalence of burnout among general practitioners: a systematic review and meta-analysis. *Family Practice*, 39(5), 943–950.
- Shroff, R. H., & Keyes, C. J. (2017). A proposed framework to understand the intrinsic motivation factors on university students' behavioral intention to use a mobile application for learning. *Journal of Information Technology Education. Research*, 16, 143.
- Sibilo, M. (2023). Personal digital tutor. In *La semplicità: proprietà e principi per agire il cambiamento*. Morcelliana.
- Smutny, P., & Schreiberova, P. (2020). Chatbots for learning: A review of educational chatbots for the facebook messenger. *Computers & Education*, 151, 103862.
- Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios hrv-heart rate variability analysis software. *Computer Methods and Programs in Biomedicine*, 113(1), 210–220.
- Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. 2015 37th annual international conference of the IEEE engineering in medicine and biology society (embc). 1934–1937.
- Teknomo, K. (2006). Analytic hierarchy process (ahp) tutorial. *Reveledu. com*, 6(4), 1–20.
- Theobald, E. J., Hill, M. J., Tran, E., Agrawal, S., Arroyo, E. N., Behling, S., et al. (2020). Active learning narrows achievement gaps for underrepresented students in undergraduate science, technology, engineering, and math. *Proceedings of the National Academy of Sciences*, 117(12), 6476–6483.
- Thibault, M., & Hamari, J. (2021). Seven points to reappropriate gamification. In *Transforming society and organizations through gamification: From the sustainable development goals to inclusive workplaces* (pp. 11–28). Springer.
- Ting, F. S. T., Lam, W. H., & Shroff, R. H. (2019). Active learning via problem-based collaborative games in a large mathematics university course in hong kong. *Education in Science*, 9(3), 172.
- Vergara, D., Paredes-Velasco, M., Chivite, C., & Fernández-Arias, P. (2020). The challenge of increasing the effectiveness of learning by using active methodologies. *Sustainability*, 12(20), 8702.
- Wang, A. I., & Tahir, R. (2020). The effect of using kahoot! for learning—a literature review. *Computers & Education*, 149, 103818.
- Wei, J., Karuppiah, M., & Prathik, A. (2022). College music education and teaching based on ai techniques. *Computers and Electrical Engineering*, 100, 107851.
- Wolf, L. (2001). College students with adhd and other hidden disabilities. *Annals of the New York Academy of Sciences*, 931, 385–395. <https://doi.org/10.1111/j.1749-6632.2001.tb05792.x>
- Wu, M., & Luo, J. (2019). Wearable technology applications in healthcare: a literature review. *Online Journal of Nursing Informatics*, 23(3). 1–15.
- Yuan, J., Tian, Y., Huang, X., Fan, H., & Wei, X. (2019). Emotional bias varies with stimulus type, arousal and task setting: Meta-analytic evidences. *Neuroscience & Biobehavioral Reviews*, 107, 461–472.
- Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, 30, 100326.
- Zou, D., Luo, S., Xie, H., & Hwang, G.-J. (2022). A systematic review of research on flipped language classrooms: theoretical foundations, learning activities, tools, research topics and findings. *Computer Assisted Language Learning*, 35(8), 1811–1837.

## AUTHOR BIOGRAPHIES

**Miguel Civit** is an associate professor at Universidad Loyola in the department of Education & Communication. His current Ph.D. thesis is on applications of artificial intelligence for music education in the Computer Science Department of Universidad de Sevilla.

**María José Escalona** is a full professor at the Computer Science Department of Universidad de Sevilla and the director of the ES3 Research Group. She has published numerous articles in highly regarded journals in several areas related to software engineering, in particular her specialty is software requirements, early quality assurance, and early testing.

**Francisco Cuadrado** has been Dean of Communication and Education at Universidad Loyola. His research interests include music education, emotion-enhanced learning and media literacy.

**Salvador Reyes-de-Cozar** is a professor at Universidad Loyola in the department of Education & Communication. His research interests include Educational Methodology, Engagement in Education, Social Cultural Approaches and New Narratives.

**How to cite this article:** Civit, M., Escalona, M. J., Cuadrado, F., & Reyes-de-Cozar, S. (2024). Class integration of ChatGPT and learning analytics for higher education. *Expert Systems*, 41(12), e13703. <https://doi.org/10.1111/exsy.13703>