Gamification and SQL: an empirical study on student performance in a Database course

MIGUEL EHÉCATL MORALES-TRUJILLO*, University of Canterbury, New Zealand
GABRIEL ALBERTO GARCÍA-MIRELES, Universidad de Sonora, México

Gamification, defined as the integration of game elements into non-gaming environments, is a promising approach to increase student engagement, motivation and performance in Computer Science and Software Engineering education. This viewpoint encouraged the creation of QueryCompetition, a web system that allows students to practice SQL in a competitive environment. Our goal was to obtain empirical evidence on how such gamifying elements as challenges, points and leaderboards, integrated into QueryCompetition, affect student performance, motivation and user experience. We carried out a quasi-experimental study with two groups in a Database course. One group used a gamified version of QueryCompetition having access to points and leaderboards while the other used a non-gamified version with no access to the above mentioned elements. Quantitative and qualitative data were collected through tests and a survey. The results showed that there was a statistically significant improvement in student performance in the gamified group as compared to the non-gamified group. In addition, a higher motivation was observed in the gamified group. The empirical evidence presented in this paper supports the claim that inclusion of challenges, points and leaderboards together with the competitive nature of QueryCompetition impacts positively on student performance and motivation to practice SQL.

CCS Concepts: • **General and reference → Empirical studies**; • **Social and professional topics → Computational science and engineering education**.

Additional Key Words and Phrases: gamification; game-related method; empirical study, computer science; software engineering; SQL

ACM Reference Format:
Miguel Ehécatl Morales-Trujillo and Gabriel Alberto García-Mireles. 2020. Gamification and SQL: an empirical study on student performance in a Database course. *ACM Trans. Comput. Educ.* 1, 1, Article 1 (January 2020), 30 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Gamification is the use of game design elements in non-game contexts [13]. It has become a fast-growing approach [18] [32] [58] for improving learning experience and increasing learner motivation in a wide variety of settings [13] [20] [27]. Despite the currently growing research on new gamification designs in education, little has been done on how to evaluate such implementations [26], in fact, few of them offer empirical evidence on their impact [27].

More notably, in research on gamification, there is a lack of empirical evaluation [2] [20] and of a systematic methodology to incorporate gamification in education and thus improve user behavioral aspects, such as engagement and

Authors' addresses: Miguel Ehécatl Morales-Trujillo, miguel.morales@canterbury.ac.nz, University of Canterbury, Christchurch, New Zealand; Gabriel Alberto García-Mireles, Universidad de Sonora, Hermosillo, Sonora, México, gabriel.garcia@unison.mx.

performance [39] [58]. Particularly, current literature is not conclusive about the improvement of student performance when using games in a learning environment [14] [56].

Regardless of its popularity, existing empirical evidence on gamification applied to Computer Science (CS) and Software Engineering (SE) is quite preliminary or even immature [27] [35] [50] [58]. At the moment, there is a need for improving its scientific rigor [40], for carrying out large-scale empirical studies on the use of gamification in SE education [50], and in particular, for performing external replications to corroborate and generalize the findings obtained [57]. Given the complexity of CS and SE topics, specificities of teaching SE in universities, and the variety of domains related to SE practices, there is a necessity to explore the benefit of using games in supporting SE related learning outcomes [38] [25].

With the aim of addressing the aforementioned needs, an empirical study for evaluating effects of gamification in CS and SE education was designed. Our study focuses on Database Management, a specific subdomain within CS and SE disciplines. Database Management is under the Computing Foundations knowledge area of the SWEBOK [4] and is related to defining and manipulating data through Structured Query Language (SQL). SQL is a domain-specific programming language designed for data management and the prevalent language used to query relational databases in SE endeavors.

In particular, this study explores how gamified elements of QueryCompetition (QC) [34] influence SQL skills of students. QC is a gamified web system that allows students to practice SQL while solving challenges, competing with others, getting points and observing their progress on a leaderboard. The creation of QC was motivated by the necessity of adding another practical component that would expand the practical exposure to SQL and database management systems while increasing students' engagement and motivation through participating in a competition with their classmates. The objective of this paper is to explore the effect of challenges, points and leaderboards on three aspects involved in learning: student motivation, user experience and performance.

This paper is organized as follows: Section 2 presents related work on gamification in SE education. Section 3 describes QC and its game elements. The overview of the study, research questions and methodology are detailed in Section 4. The results are analyzed in Section 5, while the discussion and conclusions are presented in Sections 6 and 7 respectively.

2 RELATED WORK

In order to analyze relevant facets of the related work, this section presents a deeper insight into game-related methods and game design elements. Besides, it addresses empirical studies that explore the use of game-based learning (GBL) and gamification applied to databases. Lastly, game elements and extend to which they are used in the context of education are explored.

2.1 Game-related methods in SE

Games have increasingly become a resource of choice when looking for better learning outcomes and higher motivation in students [23]. Considering a diversity of games used for different purposes [28] [47], we adopted the concept of a game-related method in SE which refers to "any approach where games are used to support the process of teaching or learning software engineering" [35].

Particularly, game-related methods in SE education include game development, GBL, and gamification. Within the game development category, the objective is to learn software design and construction practices while developing games [35]. As for GBL, it refers to any approach where games, including serious games, are used for learning purposes

[12] [35] [47]. Lastly, gamification involves the identification, extraction, and application of individual game elements or limited, meaningful combinations of those elements [28]. A growing interest has been observed in applying gamification into diverse domains, such as health, education, human resource productivity, users' data collection, and customer participation, among others [12] [14] [20] [21]. It is believed that gamification can influence users to improve their motivation, to trigger behavioral changes, and to promote better collaboration [14].

In the context of education, gamification has become one of the predominant areas of research [12] [20], in which effects on student motivation and learning are explored [44]. It has been perceived that gamification may have the potential for motivating students thus improving both learning processes and outcomes [15] [23]. Empirical studies on gamification in a tertiary level education appear to be the most frequently reported in literature; they fall mainly within CS and information technology domains [14].

The literature review we conducted shows that few empirical studies address CS specific sub-domains of SE or Database Management. For instance, Darejeh and Salim [12] report five papers in the SE domain that focus on the usage context of a programming tool, software documentation and requirements engineering. Indeed, the majority of systematic reviews that address gamification in SE are short of studies conducted around database related topics [19].

Other literature reviews on gamification in SE report that this area of research is in initial stage and there is a need for sound empirical research in order to understand the way gamification can be useful for education and industry [2] [39]. Likewise, none of these reviews mention Database fundamentals or SQL.

2.2 Empirical research on game-related methods in databases

Only seven papers were identified after conducting a search for empirical studies that involve both the use of game-related methods and databases. The former included only GBL applications (e.g. serious games) and gamification. These papers are summarized in Table 1. We identified only three studies that focus on motivation, performance [1] and student engagement [43] [45], which are topics of our interest.

Rahman et al. [43] studied the effectiveness of using Kahoot! and Quizizz tools to improve student engagement in a Database Design course. The gamified intervention based on Kahoot! and Quizizz was used to assess students' understanding of the topics covered during lectures. At the end of semester, a survey was applied and the results showed that 96% of students felt that gamification helped them increase their engagement: they took notes and listened more carefully during lectures. Gamification resulted appealing to the students due to the easiness of use of the platform [43].

At a postgraduate education level, Abdool et al. [1] conducted a study based on a gamification intervention in the course named "Database Systems: Principles and Design" in the University of the West Indies. The purpose was to investigate the integration of specific game-based elements and pop-culture references in order to increase student interest and motivation through a system that delivered content and activities. The authors reported a positive reaction towards motivational factors, however student performance was reported unchanged as compared to previous courses [1].

Finally, Santhanam et al. [45] examined the impact of competition on learning and engagement with emphasis on playing against opponents with different skill levels. The students were trained in database design and a trivia-based mini-game was played after each video-based training module. A competitive component in form of a simulated fellow competitor was added to the game which provided messages so that real trainees felt that they were competing against a real person. As a result, trainees who played against lower-skilled simulated competitors reported higher self-efficacy beliefs and better learning outcomes while those who faced equally-skilled competitors reported higher levels of engagement [45].

Table 1. Empirical studies on game-related methods applied to databases

ID	Study purpose	Game-related method	Gamified activity focus	Sample size	Outcomes
[48]	To study effects of gamification on students' learning styles.	GBL – role playing	Learning SQL	120 university students	Use of games can achieve better learning outcomes than paper-based instruction. No difference in learning effectiveness (text-based learning style or picture-based learning style).
[49]	To study fluctuation of learning styles in GBL.	GBL – role playing	Learning SQL	60 university students	Participants who had text-based learning style tend to change their learning style. Changing learning style is correlated with the number of mistakes participant committed.
[55]	To study students' perceptions towards an adaptive online learning system.	GBL – story, missions	Learning SQL	39 university students	Students perceived that the developed system can respond to their learning goals and it was easy to use.
[43]	To study effectiveness of using a gamification technique to improve students' engagement.	Gamification – Kahoot! and Quizizz	Assessment	50 university students	Ninety six percent of students felt that gamification helped them increase their engagement during lectures. Students were attracted by gamification due to the easiness of use of the platform.
[6]	To report an experience of teaching databases with help of gamification.	Gamification – treasure hunts, competition, missions, rewards	Learning main features of four NoSQL databases	30 members of an IT department	Positive perception about learning new technology, more collaboration between testers, however programmers complained of challenges being too easy and were reluctant for non-programmers to join their team.
[1]	To study integration of specific game-based elements and pop-culture references in order to increase student interest and motivation.	Gamification – boss fight, experience points, collaboration	Learning database principles and design	28 postgraduate students	Highly positive reaction towards motivational factors, however student performance stayed the same as in previous courses.
[45]	To study effects of competition on learning and engagement.	Gamification – quizzes mini-game	Learning database design	182 business students	Students who played with lower-skilled simulated fellow competitor reported higher self-efficacy beliefs and better learning outcomes. Those who faced equally-skilled competitors reported higher levels of engagement.

Based on the scarcity of the above presented empirical studies, the need of sound research on game-related methods applied to databases is evident. In addition, it can be noticed that existing research is centered on such psychological aspects as engagement and motivation while more "measurable" aspects, like student performance are left aside. Moreover, the effectiveness of game-related methods is measured through participant perceptions and no data are collected at a pre-treatment stage of the study, i.e. pre-test, to show progression of students' knowledge and skills.

2.3 Game design elements

Gamification in education makes use of elements traditionally thought as game-like or "fun" in order to promote learning and engagement [23], those elements are known as game design elements. Different categories of game elements exist, e.g. [12] [13], [47], however, no commonly agreed classification of game design elements is found in the education domain [15]. Moreover, the gamification terminology is still unsettled [14] and lacking of standardized definitions of game elements in the field of SE [35].

For the purpose of this paper, the game design elements integrated in QC will be described as based on the model of dynamics, mechanics and game components by Werbach and Hunter [61]. This model is commonly used in both education and SE domains [2] [14] [15].

In the Werbach and Hunter model, the dynamics provide a motivation to use a gamified system through such features as constraints, emotions, narrative, progression and relationships. The mechanics, on the other hand, represent elements that drive the player involvement; these set up rules of the gamified system and manage interactions between players. Some examples of mechanics are: challenges, changes, competitions, cooperation, feedback, resource acquisition, rewards, transactions, turns and win states. The mechanics implement one or more dynamics, while dynamics can be assessed as the way users responds to the mechanics.

The third part of the model represents the game components; these are the tactics used to achieve the goals described by mechanics and dynamics. Game components include user interface elements: achievements, avatars, badges, boss fights, collections, combat, content unlocking, gifting, leaderboards, levels, points, quests, social graph, teams, and virtual goods, among others.

In the context of competitions, where participants are supposed to improve their skills, it is suggested that positive feedback, optimal challenges, progressive information, intuitive controls, points, levels and leaderboards are appropriate game mechanics [3]. Within the SE education domain, progression is considered to be the most used dynamics while the most reported mechanics are challenges, feedback and rewards [2]. As for the game components, in SE education, leaderboards, points and levels are the most reported game components [2], while in higher education, points, challenges and rankings are reported to be the most used [31].

In general terms, the most game design elements tested in empirical studies are points, leaderboards, badges and challenges [20]. The combination of points, badges and leaderboards tends to be the most commonly used subset of game design elements [14]. Although no explicit rationale for selecting these particular elements is given in the empirical studies, a possible explanation might be related to how easy they are to implement, their similarity to the classroom assessment model [14], and their reliability [12].

The gamified system in our study contains several game design elements. The dynamics are represented by progression and specific time frames. The players are presented with queries of different levels of difficulty and new contents to move onto in specific time frames, which is suggested for a better engagement at every stage [61].

On the other hand, the game mechanics are represented by challenges and competition in our gamified system. Challenges are puzzles or other tasks that require some effort to solve, they might involve time, skill or creativity; and, by overcoming them, competence or mastery is displayed [61]. Competition is defined as “a structure in which one player or group wins and the other loses” [61].

As far as the game components are concerned, our system works with leaderboards and points. Leaderboards are defined as “visual displays of player progression and achievement in rank order among some group of players” [61], while points are defined as “numerical representations of game progression” [61]. This model is taken a basis for presenting QC in Section 3.

2.4 Solutions similar to QC

An additional literature review was performed in order to retrieve similar work on SQL and gamification specifically. We ran a search in four databases (Scopus, IEEE Xplore, ACM Digital Library and Wiley Online Library) obtaining 862 results. After filtering, four candidate papers were selected; three of them ([6] [48] [49]) had been found in the first literature review described in Section 2.2. The fourth paper [36] investigates the potential of gamification for motivation and user engagement in SQL-Tutor [37], an intelligent tutoring system.

In addition to this, a Google search on the topic was carried out and four meaningful results were found: SQL Zoo [53], SQL Murder Mystery [52], SQL Island [51], and GalaXQL [17]. A fifth game, Schema Verse [46], was found but it was unplayable and therefore excluded from the analysis.

The four games and SQL-Tutor are web-based systems created with the aim to be an SQL self-directed learning resource. The key differences between QC and the aforementioned systems are that they are one-player web systems and no competition aspect is involved in any of them. Besides, their tutorial-like approach with brief explanations of the concepts, examples followed by exercises aims more at learning while the aim of QC is to provide practising environment. However, there are several aspects to consider in order to improve QC as based on the web systems we found. For example, adding a story line to the game would be beneficial and might increase the engagement factor. This can be done by relating the challenges to each other and creating a sequence supported by dialogues and animations. Moreover, in SQL-Tutor, hints have shown to be of use to improve student engagement; adding links to SQL official documentation can also be useful.

3 THE GAME: QUERYCOMPETITION

QC is a web system that creates competitions during which students solve queries using SQL in a gamified environment. QC does not teach SQL, its main purpose is to provide students with an environment to practice their SQL abilities. SQL related knowledge is expected to be covered by lecturers and tutors during lectures and lab/tutorial sessions. QC competitions are meant to be used as an additional learning resource for students apart from lectures, lab/tutorial exercises and assignments.

The system has been used since 2016 in Database courses by around 600 students in total. Using the system allows them to practice query solving abilities in a competitive environment. In the next subsections, we will describe QC in terms of its dynamics, mechanics and game components, in other words, following the Werbach and Hunter [61] model.

3.1 QC Dynamics and Mechanics

Each competition in QC is divided into rounds, during which the participants have to solve challenges, in the form of queries, presented by the system. The intrinsic nature of a competition motivated the inclusion of points and leaderboards into QC. Participants receive points based on the correctness, efficiency and response time of the solutions they provided. The competitions are customizable: choosing the number of rounds, defining the point system, adding pools of queries, and specifying the database to be used is possible.

Once a round is open, the participants can enter the competition home page. From there, the participants can download the database diagram and a backup of the database to try and solve their challenges in a local environment. They can also go directly to open the challenges to solve them. During one round, the participants can solve the challenges in the order they want (see Figure 1).

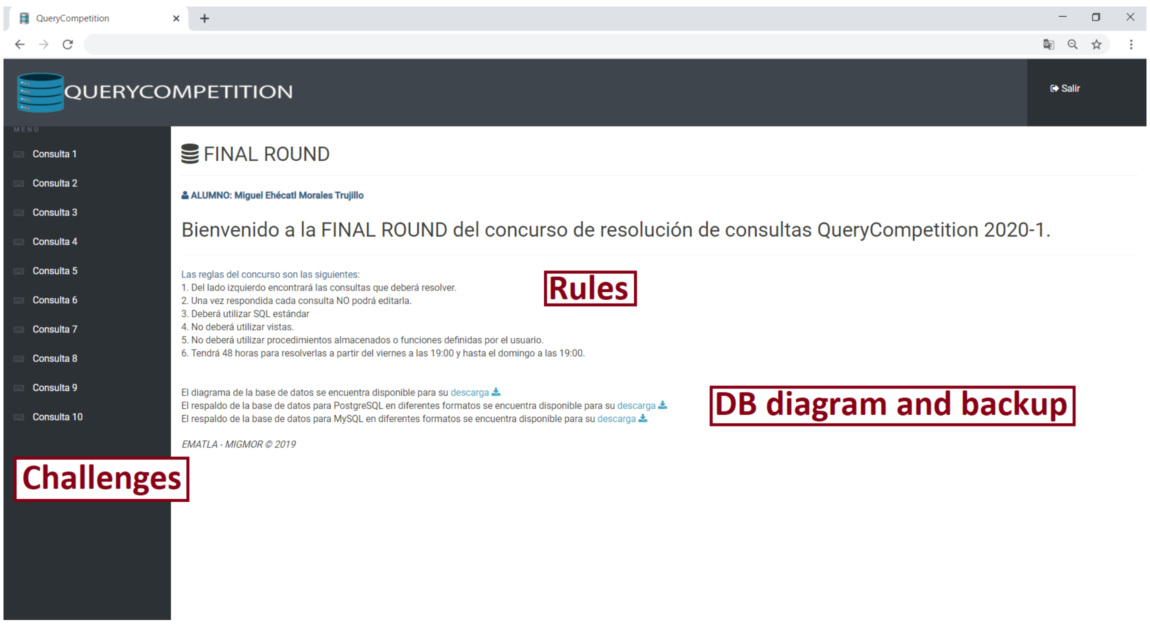

Fig. 1. QC home page.

Once they open a challenge, a solving time counter starts ticking until they submit an answer (see Figure 2). If the answer is correct, 25 points are granted, otherwise no points are granted. For correct answers, their efficiency and the time taken to submit are also taken into account. If the answer is provided in less time than the average of the group, 10 extra points are awarded. Similarly, 5 extra points are granted if the query is more efficient than the average of the group.

Based on the awarded points a global leaderboard is made accessible to participants. The results are published in the system showing a ranking with points obtained by each student. The leaderboard is the main game element related to competition and it allows for public recognition among participants. In fact, Vassileva and Sun [59] explored the use of community visualization to stimulate participation, while Marín et al. [33] stated that social comparison and competition increases student participation.

3.2 QC Game Elements

The game elements in QC are: challenges, points and leaderboards. The challenges are presented in the form of natural queries that participants have to solve by providing an SQL statement to retrieve expected data. The challenges are

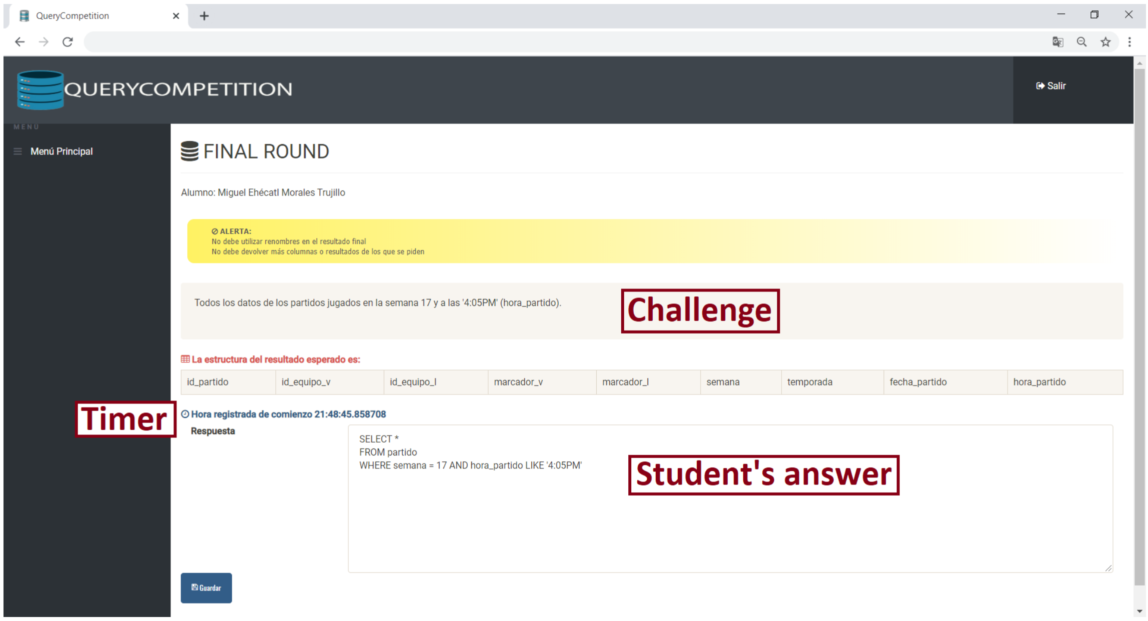

Fig. 2. Screenshot of a challenge presented to the student.

classified according to its complexity in 10 levels (from level 0, the easiest, to level 9, the hardest), see Table 2. The level of complexity of the challenges increases through the game.

The level of complexity of the queries was determined by the expected components of the model answer that must follow the ISO/IEC 9075:2016 Information technology – Database languages – SQL standard. The selection of query components was based on the Data Manipulation Language elements defined in the syllabus of the course. The queries and their levels of complexity were defined by the first author of this paper together with a lecturer with six years of experience in teaching databases at tertiary level.

QC levels of complexity are significantly more complex than the ones presented in [48] and [49], where only levels 0, 1 and 2 were considered. An example of a model answer is shown in Figure 3.

```sql
-- 9.3. Nombre del entrenador cuyo equipo es el más ganador de local.
SELECT nombre_entrenador, apellido_entrenador
FROM entrenador JOIN equipo_entrenador ON entrenador.id_entrenador = equipo_entrenador.id_entrenador JOIN equipo ON equipo_entrenador.id_equipo = equipo.id_equipo
WHERE equipo.id_equipo IN ( SELECT equipo.id_equipo
                            FROM equipo JOIN partido ON equipo.id_equipo=partido.id_equipo_l
                            WHERE marcador_l > marcador_v
                            GROUP BY equipo.id_equipo
                            HAVING COUNT(id_partido) = (    SELECT MAX(numpartido)
                                                            FROM (  SELECT equipo.id_equipo AS nombreEquipo, COUNT(id_partido) AS numpartido
                                                                    FROM equipo JOIN partido ON equipo.id_equipo=partido.id_equipo_l
                                                                    WHERE marcador_l > marcador_v
                                                                    GROUP BY equipo.id_equipo) AS t1));
```

Fig. 3. Model answer for the query “Find the full name of the coach of the team that has won most home games”.

QC has two controls embedded in the form of timers and rounds. Each round has a fixed duration, and in the case of this study, it was decided to have 48-hour rounds. However, the duration can be adjusted by the teaching team. Two other timers are present in QC, one is activated once the student opens a challenge and is deactivated when the answer is submitted or the round finishes. The second timer is used to measure the execution time of each answer.

Table 2. Level of complexity of the queries based on their SQL components

Level	SQL query structure
0	SELECT FROM
1	SELECT (aggregate functions) FROM GROUP BY
2	SELECT FROM WHERE (complex conditions)
3	SELECT FROM (complex join operations) WHERE (complex conditions)
4	SELECT FROM (complex join operations) WHERE (nested query involving an aggregate function)
5	Level 3 + (set operations) + Level 3
6	Level 3 + GROUP BY ORDER BY
7	Level 3 + GROUP BY HAVING
8	Level 7 + Level 4 (as a nested query)
9	Level 8 + Level 7 (as a nested query)

Students are awarded points for three different aspects: correctness of the answer, response time, and answer efficiency. By response time, we refer to how long it took the student to answer a challenge while answer efficiency is determined by how long it took the system to execute the code written by the student. Lastly, a leaderboard based on the points awarded is created and displayed to the students.

In order to efficiently mark students' answers, we created a script that assesses the following aspects:

- **Answer provided:** checks if an answer was provided, which is useful in order to know if the student tried to solve the query.
- **Basic content:** if an answer was provided, it checks if specific SQL keywords are present in the answer. For the purpose of this project, an answer is considered to have the basic content if it contains the `SELECT` and `FROM` keywords.
- **Syntax:** it checks if the statement follows the SQL syntax rules and can be executed without errors.

- **Correctness:** it checks if after executing the student's answer, the retrieved result set is the same as the expected result set obtained from the model answer.
- **Response time:** it obtains the difference between the time when the student opened the query and when they submitted an answer.
- **Efficiency:** it obtains the execution time of the student's answer.
- **Structure:** it checks if the retrieved columns and their order are the expected ones.

The marking script was developed by the first author of this paper and was tested in conjunction with two lecturers using a pool of 1,300 answers. The pool of answers was obtained from previous years of using QC. The testing process consisted in comparing the mark returned by the script and the mark assigned during a manual review. The only issue found during the test was related to the `ORDER BY` clause and will be discussed in Section 6.2.

The main modules of QC that support the above described mechanics are:

- **User manager:** this module allows the teaching team to register participants and create a user profile that includes: student ID, name, group, email and password.
- **Competition manager:** this module manages competitions through defining the rounds and challenges to be presented to the participants. The instructor is able to establish the number of rounds and the opening and closing times of each of them. In regards to the challenges, the instructor should provide queries in natural language, the model answer and the level of difficulty.
- **Marking:** this module handles the marking process of the submitted answers. Currently the marking process is executed on demand at any time of the competition and consists of three steps: executing submitted answer, allocating points for submitted answer, and creating/updating reports and leaderboards. A detailed explanation of the data and steps required for this process is presented in Appendix C.
- **Configuration:** this module handles the database connection to the internal QC database (users and queries), and to the external database, where the queries are executed when being marked.

QC allows instructors to use databases of their preference to design and run queries. In this study, we used an American Football database that stores information about teams and their coaches, players and their stats, games, scores, stadiums and cities; it was created by the first author of the paper for teaching purposes.

4 STUDY DESIGN

This section presents an overview of the study, its hypotheses and variables. Also, the research question together with a detailed description of the subjects who participated in the study, the steps of the study, and the data collection methods and sources are described.

4.1 Study overview

Our experience of using QC in Database groups since 2016 with around 600 students has shown that points and leaderboards have positive effects on students, with a visible increase in motivation, interest in the subject and performance. However, existing literature on the matter provides inconclusive results [2], varying from such negative effects as the increase of anxiety or the destruction of the quality of friendship [54] to positive effects such as improving student performance, increasing motivation [8] [5], and stimulating student engagement and persistence [9].

Aiming at supporting our initial observations, we designed and conducted this study which assesses the impact of QC game elements, i.e. challenges, points and leaderboards, on student performance, motivation and user experience.

During the study, we observed and surveyed two groups. The non-gamified group used QC as a web system for solving queries without any game elements, while the gamified group used QC as a platform for a query-solving competition, and had access to the points and leaderboards. Both groups received feedback from their respective lecturers and tutors on their answers in terms of correctness and mistakes, if any. The participation in the study was voluntary.

The independent variable was the QC environment the groups used: gamified and non-gamified while the dependant variables were considered to be motivation, user experience and performance. The following research hypotheses were outlined:

- **H1.** Students in the gamified group will have a **higher motivation** than those in the non-gamified group.

Motivation is defined as the willingness to do or the enthusiasm for doing something [7]. It is well known that motivation plays a key role in student learning [42]. In the current study, motivation will be measured using the attention, relevance, confidence, and satisfaction (ARCS) model [24], which is a commonly used motivational design model [30].

- **H2.** Students in the gamified group will have a **better user experience** than those in the non-gamified group.

User experience can be measured in terms of user performance and perceived satisfaction [29]. For the purpose of this study, we defined the user experience including the following dimensions: immersion, challenge, competence, fun and social interaction [60], which will be measured through a survey.

- **H3.** Students in the gamified group will **perform better** than those in the non-gamified group.

Lastly, the performance will be measured using two metrics, one is objective and the other is subjective. The former is based on the correct answers obtained in two tests while the latter will be gathered from the students' perceptions regarding their learning.

4.2 Research Question

The research question was defined following the recommendations stated in [14] on how to formulate a research question for studies on gamification in education. The overall goal of this project is to understand whether game design elements, such as challenges, points and leaderboards, make an impact on the process of university students learning SQL, particularly, on student performance, motivation and user experience. To achieve this goal, we sought an answer to the following research question:

How do challenges, points and leaderboards impact student performance, motivation and user experience in students using QC?

While a wide majority of studies report an increase of engagement and motivation when applying gamification strategies [11] [16], there is little empirical support for improvement of student grades or academic performance as described in Section 2.2. We will analyze the student performance by comparing the results obtained in a pre-test and post-test in a gamified and non-gamified educational environments. The motivation and user experience will be analyzed using the results gathered in a survey. The survey consists of a standardized questionnaire developed by von Wangenheim et al. [60], which evaluates "how the students felt about the training or learning experience" [60] and it already includes the ARCS model.

4.2.1 Subjects. The participants of this study were students from a Mexican university enrolled in two Database courses. The gamified group (G) consisted of 65 students while the non-gamified (N) of 74 students. Students from both groups were asked to use QC to practice their SQL abilities. The students in the G group were told that this activity was a competition and had access to the points and leaderboards, while the students in the N group did not. The median age of the participants was: 21.5 (average: 21.9) for the G group and 21.0 (average: 21.6) for the N group.

Both groups followed the same syllabus and pursued the same learning outcomes, but had different lecturers. In regards to the learning outcomes, the relevant to this study are:

- Implementing relational databases using SQL.
- Using SQL to query, populate, update and manage databases.

The G and N groups' courses ran 16 weeks during the same semester in parallel. Lecturers had equivalent experience and knowledge in teaching the course, four and six years respectively. Both have a Master degree in the Information Technology area and a Bachelor degree in Computer Science within the same University.

4.2.2 Steps of the Study. The empirical study consisted of three stages distributed during five consecutive weeks. The stages of the study were:

Phase 1 (both groups): a pre-test was applied one week before the competition started during a regular lab/tutorial session supervised by the lecturer and tutors. The pre-test consisted of 10 queries ordered by level of complexity (0-super easy to 9-extremely hard). Students had two hours to solve the test. The objective of the pre-test was to discover and compare the students' previous knowledge of SQL, if any, in order to identify differences in knowledge that could affect the final results. Students were not expected to have any previous knowledge of SQL. The pre-test queries can be consulted in Appendix A.1.

Phase 2 (Gamified group): a 3-round competition was carried out. In each 48-hour round, the students had to answer three queries individually. The first round included queries from levels 1 to 3, the second round from levels 4 to 6 and the third round from levels 7 to 9 (See Table 2). The rounds ran during three consecutive weekends, starting on a Friday and ending on a Sunday. Once each round ended, the answers were marked, feedback was provided to the students, points were awarded and the leaderboard was published.

Phase 2 (Non-gamified group): students were asked to solve three queries during three consecutive weekends, they also had 48 hours per weekend to solve them. The set of queries was the same as for the G group. However, in contrast to the G group, they were neither granted points nor had access to leaderboards thus using the system only as a means of practicing queries and submitting their answers. Once each round ended, the answers were marked and feedback was provided to the students.

Phase 3 (both groups): a post-test was applied during the following week after phase 2 during a regular lab/tutorial session supervised by the lecturer and tutors. It consisted of 10 queries ordered by level of complexity (0-super easy to 9-extremely hard). Likewise, students had 2 hours to solve them. In addition to the test, students were asked to participate in a three-section survey that collected information regarding their user experience, emotions and general information. The post-test queries can be consulted in Appendix A.2.

All the above tasks were performed online through QC (phase 1: pre-test; phase 2: rounds; and phase 3: post-test) and Qualtrics (phase 3: survey), which is a software that enables users to collect and analyze data online. Each participant was assigned a unique identifier and no personal data except for their age were collected.

The course outline and the schedule of the study phases are detailed in Appendix B.

4.3 Data collection

The study ran during the semester 2020-I (August 2019 to January 2020) in two groups of a Database course at the Science Faculty of the National Autonomous University of Mexico. The total duration of the case study was five weeks: it started with the application of the pre-test, followed by three consecutive rounds, closing up with the application of the post-test and the survey. All the materials, survey, tests and the system were provided in Spanish.

Three main aspects were measured in the study: performance, motivation and user experience. The student performance was measured comparing the results of the pre-test, post-test (phase 1 and 3), and the difference between the pre- and post-test (post-test results minus pre-test results within the same group). The data were collected through the web system and the answers were assessed using the marking script that compared the model answer result set against the one retrieved after executing a student's answer.

The motivation and user experience were measured through a survey which consisted of three sections: a standardized questionnaire [60], a list of emotions, and general data questions. The standardized questionnaire presents 27 items organized in three sub-components: motivation, user experience and learning, assessed through a survey (phase 3). The 27 items were measured on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The following is an example of an item: "The way the game works suits my way of learning". The items of this section of the survey are presented in Table 6.

The second section of the survey asked the participants which of the presented emotions they experienced while using QC. They had to choose from a list of 17 different emotions, which can be consulted in Table 7. The last section of the survey consisted of five questions:

- How old are you?
- Have you ever taken this Database course before? If yes, how many times?
- Did you have any SQL knowledge prior to this course? If yes, where did you obtain it?
- Did you use any external resources to solve the queries presented in QC? If yes, what kind of resources did you use?
- Please, provide your opinion of using QC.

The data analysis steps were as follows: first, a descriptive analysis was carried out for the pre- and post-test results (count, mean, median, standard deviation, box-plots and bubble charts); second, a Shapiro-Wilk's normality test was performed; finally, as the data were not normally distributed, a Mann–Whitney–Wilcoxon test to contrast the experimental hypotheses was carried out.

The results were analyzed using R for performing statistical tests, and Excel for calculating basic statistics and creating charts. For the significance of all the statistical tests the alpha level was set at 0.05.

The SQL answers were assessed using a MySQL database and the marking script in PyCharm Professional edition [22]. In addition, the research was approved by the University of Canterbury Educational Research Human Ethics Committee (Reference: HEDN/ERHEC/2019-02).

5 RESULTS

This section presents the results of the study first through descriptive statistics, then through statistical tests run in order to determine statistical significance. In the end, the data collected through the survey are presented and analyzed.

5.1 Analysis of the pre- and post-tests results

Forty-two students out of 65 enrolled did the pre-test in the G group and 50 students out of 74 in the N group. The distribution of correct answers per student is shown in Figure 4, the G group is shown in green while the N group in blue. The median of correct answers for both groups was one. In the G group, the average was 0.86 and the standard deviation was 0.861. On the other hand, the N group showed an average of 0.90 and a standard deviation of 0.700. It can be observed that both groups had a similar level of knowledge.

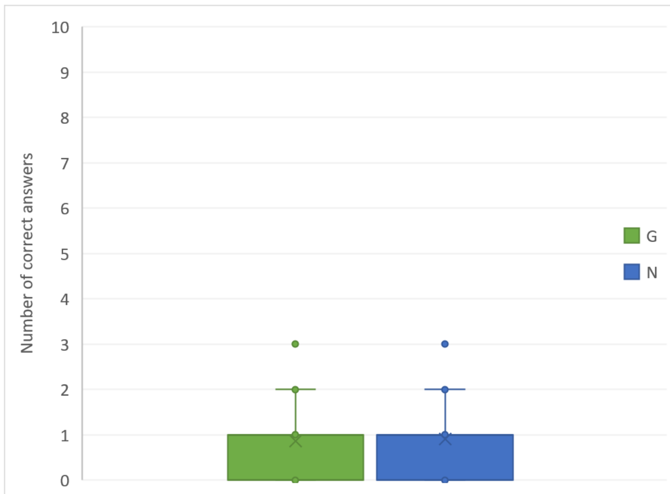

Fig. 4. Data distribution of correct answers per group, pre-test.

The post-test results for the G group showed a higher median (four) than in the N group (three). Also, a higher average and standard deviation (average: 5.17, standard deviation: 2.329) was observed in the G group as compared to the N group (average: 2.60, standard deviation: 1.400). The post-test results distribution is shown in Figure 5.

The descriptive statistics calculated from the post-test results reveal that the grades obtained when using the gamified environment were higher than the non-gamified one. Lastly, in order to calculate the absolute performance improvement on each individual, the difference between the pre- and post-test scores was calculated. Figure 6 shows the difference in performance for each group. A better differential in performance of the G group (median: 4, average: 4.31) than of the N group (median: 2, average: 1.70) was observed.

A summary of the results is presented in Table 3.

Figure 7 shows the correlation between the results of the pre-test and the post-test for each student. The green bubbles represent the G group students while the N group students are shown in blue. Only one student from the G group did not improve obtaining two correct answers in the pre-test and one in the post-test. On the other hand, in the N group, nine students did not improve, seven of them obtained the same number of correct answers in both tests and the rest worsen by one and two correct answers respectively.

In order to determine the normality of the pre-, post-test and the difference in performance of both groups, we performed Shapiro-Wilk's test. The resulting p-values were small leading to the conclusion that the distributions were significantly different from a normal distribution, with the only exception being the difference in performance of the G group, which resulted to be not significantly different from the normal distribution (see Table 4).

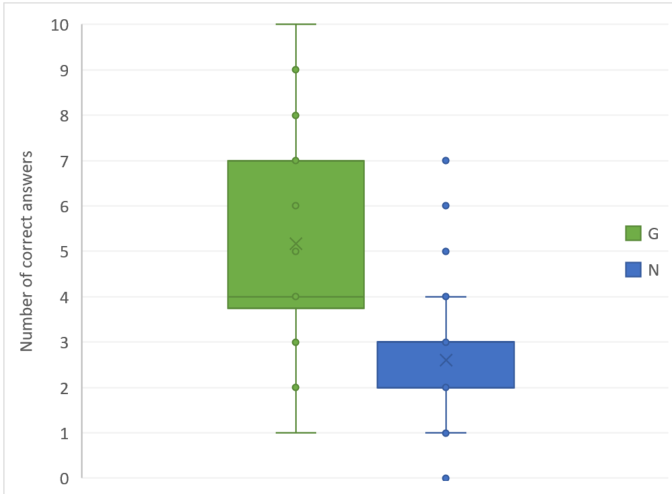

Fig. 5. Data distribution of correct answers per group, post-test.

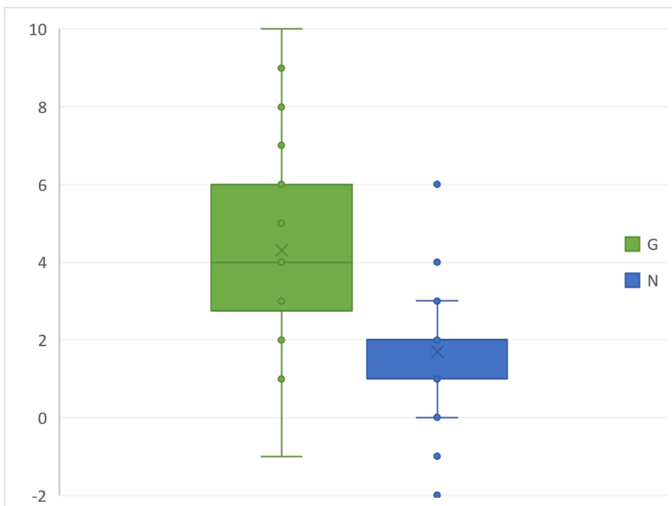

Fig. 6. Data distribution of the absolute performance difference per group.

As the data were not normally distributed, the non-parametric Mann–Whitney–Wilcoxon's test was applied to test the null hypothesis (see Table 5):

H0: The students who used the gamified version of QC will not obtain different marks to those obtained by the students using the non-gamified version of QC in the test (pre, post and the difference).

In the case of the pre-test, the p-value rendered is greater than 0.05 therefore we accept the null hypothesis (*H0*), which in turn leads us to conclusion that no difference between the groups has been found at the beginning of the

Table 3. Descriptive statistics of the study

		G	N
Enrolled students		**65**	**74**
Students that answered the pre- and post-test		**42**	**50**
Students that answered the survey		**34**	**49**
Pre-test	Median	**1.0**	**1.0**
	Average	**0.86**	**0.90**
	Standard deviation	**0.861**	**0.700**
Post-test	Median	**4.0**	**3.0**
	Average	**5.17**	**2.60**
	Standard deviation	**2.329**	**1.400**
Difference	Median	**4.0**	**2.0**
	Average	**4.31**	**1.70**
	Standard deviation	**2.46**	**1.36**

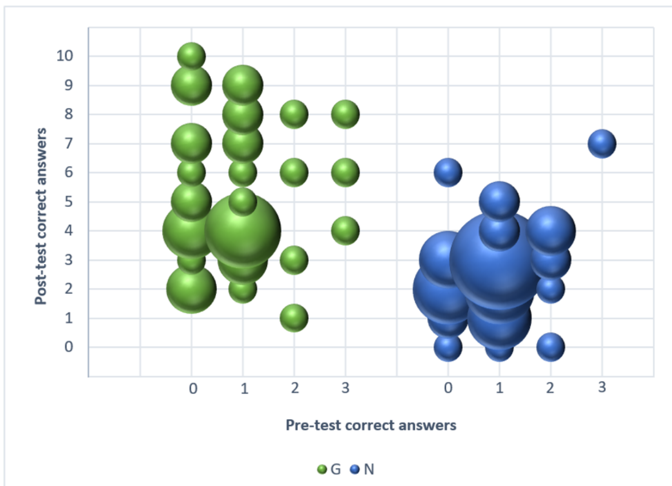

Fig. 7. Comparison of correct answers per group, pre- and post-test.

Table 4. Shapiro-Wilk test results for the pre-, post-test and the difference in performance

	G	N
Pre-test	**<0.001**	**<0.001**
Post-test	**0.0217**	**<0.001**
Difference	**0.1814**	**<0.001**

study. This result also confirms our observation from Figure 4 about both groups possessing similar knowledge when the study began.

Table 5. Mann–Whitney–Wilcoxon test results for the pre-test, post-test and the difference in performance

	W	p-value
Pre-test	**974.5**	**0.5188**
Post-test	**1737.5**	**<0.001**
Difference	**1736.5**	**<0.001**

On the other hand, in the case of the post-test and difference in performance, the p-value turned out to be less than 0.05 significance level, so the null hypothesis was rejected (therefore, the alternative hypothesis, ¬*H0*, is accepted). In conclusion, we can claim that the difference in performance between the two groups is statistically significant.

5.2 Analysis of the survey results

Eighty-three students answered the survey: 34 from the G group and 49 from the N group. The G group obtained an equal or higher median for each of the 27 items from the standardized questionnaire. Table 6 shows the median of each item for each group, highlighted in bold are the items which were evaluated higher by the G group.

According to the results obtained in the survey, the G group participants tended to rate the attention, relevance and confidence, which are the aspects from the Motivation sub-component, higher. The User experience sub-component was evaluated higher in the G group as well, in particular, immersion, social interaction and fun. Regarding the Learning sub-component elements both, short- and long-term learning, were better perceived in the G group. While students had to solve the queries individually and they were not allowed to consult other students, the social interaction component was favoured by the competitive approach followed in the G group, since students tended to discuss the results from the leaderboard after each round.

A reliability test of the survey was conducted using the Cronbach Alpha test [10]. The Cronbach's alpha coefficient of the instrument was 0.9178, which supports the internal consistency of the survey. The Cronbach's alphas for each sub-component were 0.834 (Motivation), 0.848 (User experience) and 0.859 (Learning) thus indicating sufficient reliability of the survey.

The survey used in this research is very similar to the Model for the Evaluation of Educational Games (MEEGA) [41], which is reported as a reliable instrument (Cronbach's alpha: 0.915) and has an acceptable internal consistency in motivation (0.802), user experience (0.862) and learning (0.797). Actually, the main differences between the instrument we used in this research and the MEEGA questionnaire are: first, a slight rewording of nine items and, second, omitting two questions from the digital game category which belong to user experience factor. It is worth mentioning that the wording was slightly adjusted between the G and N groups surveys in order to fit in with their characteristics. Examples of those adjustments are replacing the words 'game' by 'web system', and 'playing' by 'using' accordingly.

The final section of the survey asked students about the emotions they experienced during their involvement in the study. This information was collected by asking participants to tick the emotions from a list. The results are shown in Table 7, positive emotions are preceded by a plus sign (+) while negative emotions are preceded by a minus sign (-). The number in the second and third columns indicates the percentage of participants in each group that chose the respective emotion.

Lastly, we observed that 28 students missed one or more rounds out of the three rounds (phase 2). 14 students from the G group missed one round each while 11 missed one round and 3 missed two rounds from the N group, with round 2 being missed the most: 12 from the G group and 6 from the N group missed it. Regarding the number of dropouts, i.e.

Table 6. Motivation and user experience questions and their median

	Motivation	Median G	Median N
Attention	The game/web system design is attractive.	**4**	**4**
	There was something interesting at the beginning of the game that captured my attention.	**4**	**3**
	The form, content and activities helped me to stay focused on the game.	**4**	**4**
Relevance	The content of the game is relevant to my interests.	**5**	**5**
	The way the game works suits my way of learning.	**5**	**4**
	The game content is connected to other knowledge I already have.	**5**	**4**
Confidence	It was easy to understand the game and start playing/using it as learning material.	**5**	**5**
	While playing the game, I felt confident that I was learning.	**5**	**4**
Satisfaction	I feel positive because I know I will have opportunities to use what I learned playing this game in practice.	**5**	**5**
	It is due to my personal effort that I managed to advance in the game.	**5**	**5**
	User experience	**G**	**N**
Immersion	**Temporarily I forgot about my daily routine; I have been fully concentrated on the game.**	**5**	**4**
	I lost the track of time while playing; when I realized that, the game had already ended.	**4**	**4**
	I felt myself more in the game context than real life, forgetting what was around me.	**4**	**3**
Social interaction	I was able to interact with others during the game.	**4**	**4**
	I had fun playing with other people.	**3**	**3**
	The game promotes moments of cooperation and/or competition among players.	**5**	**4**
Challenge	This game has an adequate level of challenge for me; the tasks are neither too easy nor too difficult.	**4**	**4**
	The game progresses at an adequate pace and does not become monotonous; it offers new obstacles, situations or task variations.	**4**	**4**
Fun	I had fun playing the game.	**4**	**4**
	When interrupted at the end of the activity I was disappointed that the game was over.	**3**	**3**
	I would recommend this game to my peers.	**5**	**4**
	I would like to play this game again.	**5**	**4**
Competence	I achieved the goals of the game applying my knowledge.	**4**	**4**
	I had positive feelings on the efficiency of this game.	**4**	**4**
	Learning	**G**	**N**
Short-term	**How much do you think the game contributed to your learning in this course?**	**5**	**4**
	How efficient was the game for your learning, comparing it with other activities of the course?	**5**	**4**
Long-term	**Do you think the experience with the game will contribute to your professional performance in practice?**	**5**	**4**

students who solved the pre-test but abandoned the study before completing the post-test, six students from the G group and seven from the N group were registered. In other words, 87.5% of students started and completed the study in the G group, and 87.7% in the N group. We could not observe any tendency in terms of students providing more

Table 7. Summary of emotions

Emotion	G	N
- Anxious	**65%**	55%
- Angry	12%	**16%**
- Bored	0%	**18%**
- Lost	21%	**27%**
- Sad	0%	**6%**
- Stressed	**65%**	43%
+ Capable	**53%**	31%
+ Challenged	**74%**	57%
+ Confident	6%	**16%**
+ Encouraged	41%	41%
+ Focused	**65%**	49%
+ Happy	**35%**	27%
+ Immersed	**47%**	24%
+ Interested	**79%**	59%
+ Optimistic	24%	**33%**
+ Relaxed	3%	**22%**
+ Satisfied	**53%**	45%

correct answers by the end of the competition, which can be attributed to the increasing level of complexity of the challenges. Apart from these observations, no other tendencies could be inferred.

6 DISCUSSION

In this section, we examine the hypotheses in the light of the obtained results, we also answer the research question and point out several improvements to QC as well discuss threats to validity and how they were mitigated.

6.1 Performance, motivation and user experience

The discussion in this subsection is guided by the three hypothesis stated in Section 4.1.

H1. Students in the gamified group will have **higher motivation** than those in the non-gamified group. The *motivation* sub-component was positively evaluated by the G group, four out of 10 aspects got a higher median than in the N group, while the rest six aspects got the same median in both groups. In particular *attention*, *relevance* and *confidence* were rated higher, confirming our observations of positive effects the gamification produces in students. Our results are in line with current reviews on the impact of gamification on motivation. For instance, Hamari et al. [20] reported that all studies in education contexts tend to report positive results in terms of both increased motivation and engagement in the learning tasks (as well as enjoyment over them). Similarly, Sailer and Homner [44] concluded that empirical studies showed a statistical significant positive effect of gamification on motivational outcomes. Besides, Dichev and Dicheva [14] found empirical papers reporting the same.

The participation rate in both groups was similar, but slightly higher in the N group (67.5%) as compared to the G group (64.6%). Taking into account the emotions experienced by the participants, the students in the G group felt more interested (79.4% vs. 59.2%), more focused (64.7% vs. 49.0%) and more satisfied (52.9% vs. 44.9%) than the students in the N group. Altogether, the results support that the gamified group shows a higher motivation than the non-gamified group.

Finally, we include several comments related to motivation produced by the students:

- *"Before QC [rounds] started, I pushed myself to study, I learned more"*
- *"I really liked it, it pushed me to study and be more demanding with myself"*
- *"It was a challenging experience, the competition made me put more effort in being faster and better"*
- *"Although some of them were difficult [the challenges], I kept trying until I got it right"*

H2. Students in the gamified group will have a **better user experience** than those in the non-gamified group. The *user experience* sub-component was also evaluated higher in the G group as compared to the N group, five out of 14 aspects of this sub-component were better perceived in the G group than the N group, and the remaining nine had equal medians. Students felt immersed and reported an increased social interaction.

The *fun* factor was positively seen by the participants, they reported they would recommend the game to their peers and would play again. These findings are backed by the emotions the students experienced: immersion (47.1% vs. 24.5%) and fun (35.3% vs. 26.5%).

Challenge and *competence* were two aspects that showed no significant difference in the first section of the survey, however, the emotions section showed clear differences in favor of the G group, participants felt challenged (73.5% vs. 57.1%) and capable (52.9% vs. 30.6%).

On the other hand, we observed several negative effects which are derived from the competitive aspect of QC. Participants in the G group felt more anxious (64.7% vs. 55.1%) and stressed (64.7% vs. 42.9%). This effect can be caused by such features as the leaderboard and the timers (time pressure) that constitute the competition itself.

Also, considerably less participants felt relaxed (2.9% vs. 22.4%). These results were expected since the level of complexity of the challenges and the nature of being immersed in a competitive environment are bound to produce a strong effect on the students. Another interesting finding was that no students in the G group felt sad or bored while in the N group 18.4% felt bored and 6.1% sad, which are also considerably low percentages in general.

We noted that the gamified group had a slightly better user experience. However, such emotions as stress and anxiety must be considered when choosing a gamified activity since not all students feel comfortable in a competitive environment. Therefore, we believe that the findings are not strong enough in order to set a difference between both groups regarding the user experience.

Finally, we include several comments related to user experience produced by the students:

- *"The competitive factor helped me to demonstrate myself that I am really learning"*
- *"I enjoyed it a lot, I could see what I can do with what I learned"*
- *"Amazing, this was the most efficient way of learning SQL"*
- *"It was fun, I enjoyed to compete with my classmates, the time pressure kept me focused and interested"*

H3. Students in the gamified group will **perform better** than those in the non-gamified group. According to the pre-test results, it can be stated that both groups had almost the same level of knowledge and abilities in SQL at the beginning of the study as both groups had the same median of correct answers (1.0) and a close average (0.86 and 0.90). Comparing the results from the post-test, we can observe that the participants in the G group performed considerably better. The G group rendered a higher median of correct answers as compared to the non-gamified group (G: 4.0 vs. N: 3.0) as well a higher average (G: 5.17 vs. N: 2.60).

In addition, the improvement in the performance differential was the double in the G group (median: 4.0, average: 4.31) than in the N group (median: 2.0, average: 1.70). In this case, we can state that the performance of the G group students was considerably better than of the N group students. Regarding the students' perception about their learning, for the three items of the survey the G group got higher medians (5) than the N group (4).

The positive impact on learning performance in this study is consistent with the results of current research on gamification in SE education [2] [20] [44], who report a positive impact of gamification on learning outcomes.

Finally, we include a comment related to student performance produced by a student:

- *"After the pre-test I was very sad, but after the post-test the feeling was completely the opposite, it was so satisfactorily to have the knowledge to solve it correctly"*

Going back to our initial research question *How do challenges, points and leaderboards impact student performance, motivation and user experience in students using QC?*, we observed that points and leaderboards have a positive impact on the student performance and motivation. The results show that the participants exposed to these game elements performed significantly better than the participants who were not exposed to them.

Besides, both groups showed a similar user experience (H2) with the G group revealing a wider range of emotions than the N group. Although they reported being anxious and stressed, they also reported to be more interested and focused, which can be seen as a trade-off between negative (stress and anxiety) and positive emotions (interest and focus). Moreover, the results are in line with previous studies [2] [44] concerning the benefit of using game elements to increase motivation (H1) and student performance (H3) in teaching SE topics. In our work, the contribution is based on studying the impact of gamification on exercises to learn SQL.

However, other topics of interest were discovered and might be worth looking at: which of the game elements had a major effect on the learning performance, which one(s) could be related to user experience and which one(s) has more effect on motivation? For example, we observed that leaderboards kept students motivated. However, a new instrument needs to be designed in order to collect data explicitly about each of the QC game elements in particular.

Another interesting question is whether personal characteristics of the participants, such as gender or age, affect the preference for a gamified learning environment? Anecdotally, in this study, the ratio between female and male participants was close to ½, however, we did not collect gender data.

6.2 Threats to validity

Several threats to validity were identified during the study, and will be discussed in the following paragraphs.

External validity. The results cannot be generalized yet as the evaluation was carried out in one university. However, they are relevant for other courses teaching SQL in CS related degrees. The participants are representative of the population under study: university students in their third or fourth year of studies learning SQL in a Database course.

In order to have a representative experimental material, we prepared the challenges (queries) taking an international standard as a basis and making sure they are representative of basic SQL elements taught in a Database course in a tertiary institution. In addition, the challenges were designed and validated by lecturers with a wide experience in teaching Databases and using SQL in the industry. Regarding the experimental setting, the context and conditions of both groups were as similar as possible: same semester, same course and dates. Moreover, both groups received the same number of lectures and lab/tutorials prior to and during the study: 3 lectures and 2 labs/tutorials per week.

The teaching material might represent a potential source of differences, since it was created by each lecturer. In order to minimize this effect, the lecturers were informed about the study and the expectations from them. Therefore, it was agreed that the SQL-related material would follow a common course outline, the sessions would be scheduled in parallel (same weeks) for both groups and the labs/tutorials would use equivalent handouts.

Internal validity. The participation was voluntary and confidential, so the students who joined the study did it freely and with the assurance that the results of the study would not affect their grade in the course as their marks in the pre- and post-test were not disclosed to their lecturers. Another factor that could potentially affect the results is the students' previous knowledge of SQL. We tried to mitigate it by asking for this data explicitly in the survey. In fact, we identified through the survey that one participant from the G group, who obtained a high number of correct answers in the pre-test, was re-taking the course. That student's results had to be discarded from the data set and were not considered for the analysis. No students from the N group were re-taking the course.

An important threat to the internal validity is the potential demoralization of those students who do not like competitive situations or whose learning styles do not benefit from gamification. We were aware of this threat; as shown by the survey, anxiousness and stress levels were higher in the G group. However, in the experience of the lecturers involved in the study, these emotions did not affect the usual student behavior drastically. In addition, the mortality, i.e. dropouts from the study, was low and very similar in both groups.

Another potential threat is cheating: the results of the pre- and post-tests could be affected by students' cheating. In order to minimize the possibility of cheating, we created five queries per level, which were allocated randomly to the participants at the beginning of each round.

Lastly, the instrumentation for the study was the same. Both groups used the same data collection forms and tools, with the only difference, as mentioned earlier, being the QC versions – either gamified or non-gamified.

Construct validity. In order to clearly identify the impact caused by the game elements and to create a more controlled environment, we decided to include only three game elements: challenges, points and leaderboards. While it could be seen as a limitation or an improvement opportunity, since badges, avatars and other game elements are widely used for a better engagement and user experience, our primary interest lay in validating our previous observations regarding motivation and performance. In addition, it helped us to clearly define what we wanted to measure, i.e. the construct. Nonetheless, the inclusion of more game elements is considered as future work.

Another key aspect of this project was to efficiently determine the correctness of a student's answer to a query, which was assured by a thorough and extensive testing of the marking script conducted prior to the study with a sample of 1,300 answers. During the study, at the beginning of each phase, the setup of the rounds and challenges, and the marked answers at the end were reviewed by the first author of this paper together with the two lecturers that participated in the study.

In order to mitigate the mono-method bias, i.e. using a single type or measure observations, apart from the pre- and post-tests, we also considered the perception of the students regarding their learning, in this case both showed an improvement in the G group. Nevertheless, we believe it is worth considering an addition of another instrument to test the student performance independently of the pre- and post-tests, which would allow for a deeper comparison and reinforcement of our performance improvement claim. For example, a traditional exam within the course regular activities can serve as a reference for measuring the impact of QC on the short- and long-term learning outcomes. Particularly, in both groups, the exams that assessed the SQL knowledge also contained other topics like Security, Transactions and Functions, therefore isolating SQL in one 'regular' assessment item within the course was not feasible due to time constraints of the course schedule. Apart from adding more game elements, an independent exam inclusion is considered for the future studies with QC.

Last but not least, the authors of the paper did not interact with the participants disallowing potential bias or experimenter expectancies, which we consider to be a strength of this study.

Conclusion validity. The data were analyzed using statistical tests recommended in [62]. The reliability of the measures was kept under control by defining a simple metric: correct answers. A potential issue was discovered prior to the start of the study. For those solutions that required ordering (`ORDER BY` clause), if two or more rows of the result set were "tied" by the ordering field, the result sets could be considered different by the script and erroneously marked as incorrect. For example, asking for a list of players ordered by the number of touchdowns they had, the system could return ["Alberto, 5" , "Miguel, 5"], which is equivalent to ["Miguel, 5" , "Alberto, 5"], however, the script will consider these results different and one of them as incorrect. To fix this issue, the queries that involved ordering were tested to ensure the result set would not produce "ties".

Regarding the reliability of the treatment implementation, both lecturers were informed about the study design, their responsibilities and the expected behavior from them. At the beginning of each phase briefing sessions were held through video conferences.

We are aware that the groups had different lecturers and it could have influenced the results in the case of one lecturer explaining better or engaging the students' attention more. However their academic background, database related knowledge, and teaching experience were very similar. The setting was the same for both groups, both ran in parallel under the same conditions within the university. Nevertheless, in order to increase the power of generalization, more empirical studies should be carried out in cohorts from different universities. For this purpose, the following study designs are taken into account: (i) one group – one lecturer, in this format, half of the group would use the gamified system and the other half the non-gamified one; and (ii) two groups – one lecturer, this version of the study would take longer, one year the group would use the gamified system and the next year a new group would use the non-gamified version. These suggested setups are expected to minimize the concerns about having different lecturers and/or teaching resources.

6.3 Limitations and improvements

Several limitations were identified during the study. Some of these limitations are related to the threats to validity previously discussed in section 6.2.

Two groups, two different lecturers. As we have already discussed, the course materials and personal qualities of each lecturer can influence the results. However, given the university constraints of not allowing a lecturer to teach two groups of the same course in the same semester, it was not possible to run the study in two groups with the same lecturer.

Bigger pool of challenges, potentially less cheating. Cheating is an issue that can impact the results in any assessment. For the purposes of this study, we created different queries for each level and randomized their allocation to students; the pool could, however, be bigger in order to minimize the student cheating. As an additional strategy, we carried out a manual process to catch hardcoded answers, considering only those answers that were marked as correct. Only one case was discovered: a student 'hardcoded' the answer to the query 'Find the team that has won the most home games' (`SELECT team_name FROM team WHERE team_name LIKE 'New England'`). An automatic notification feature would be implemented to catch potential dubious cases, particularly when: (i) answers were submitted in extremely short periods of time, which could suggest a 'copy-paste' action; (ii) answers were submitted from the same IP range and almost at the same time; and (iii) answers that do not contain the expected keywords, to avoid 'hardcoding' cases.

The queries were also analyzed looking for tendencies that could reveal other cheating actions, more correct answers while the rounds progressed for example. However, no tendency was observed, which can be attributed to the fact that the queries become more complex in the last rounds.

Not supported by an external assessment item. While having a regular assessment item in the course to evaluate the SQL-knowledge was ideal, this could not be done because of time constraints of the lecturers and how the assessments were distributed through the semester in each course. An independent, and ideally invigilated, assessment item would have given the opportunity to contrast its results against the post-test results, providing more confidence to the study results.

Support and increase of (healthy) social interaction. While the intention of the competition was for the students to solve the challenges individually, the inclusion of 'team challenges' can add another dimension to the social interaction aspect. This addition can improve the quality of the interactions and favour the collaboration and communication among participants.

The following relevant improvements that could be made to QC were collected from the comments provided by the participants: (i) once the answer is submitted, the feedback should be immediately shown to the participant; (ii) an opportunity to resubmit should be allowed for; (iii) hints should be added to the most complex challenges; (iv) interaction between participants should take place, for example by including team challenges; (v) in order to reduce the levels of anxiety and stress, a 'partial' leaderboard can be implemented, showing only the closest participants to a student, i.e., a student only would know the two students above and the two below them. Fully or partial anonymization of participants can be implemented too by allocating random names or letting them choose a personalized username or avatar; and (vi) for the purpose of reducing the probability of cheating, a wider variety of queries should be added to the pool of challenges as well as a feature to detect potential cheating cases should be implemented in the system.

A new version of QC will take these comments into account in order to improve social interaction and learning aspects.

7 CONCLUSION

Gamification has become a topic of interest in the tertiary education domain in the last decade; it has proved to bring multiple benefits to students and has offered numerous alternatives to educators in order to enhance the teaching and learning experience. This study falls within the first research projects that assess effectiveness of gamification on student performance in a Database course. We designed and carried out an empirical study to evaluate to what extent such game elements as challenges, points and leaderboards influence student motivation, user experience and performance while they practice SQL. Two groups from a Mexican university participated in this 3-phase study, where the participants were asked to use QC in a gamified and a non-gamified environment.

The study showed that the participants from the gamified group obtained better results in motivation while user experience was rated only slightly higher than in the non-gamified group. However, the most important improvement observed is the one related to student performance. In fact, students using the gamified version of QC doubled differential of performance compared against the non-gamified group. Negative effects of gamification, particularly those related to the nature of competition, were observed as well; however, we consider they were outweighed by an increase in engagement and motivation when students got public recognition.

As future work, more empirical studies involving the use of QC should be carried out in order to increase the power of generalization of the claims presented in this paper. This can be done by introducing QC in other groups from different universities. As for QC, improving its aesthetical aspects and introducing more game elements are planned for the next release of the web system.

ACKNOWLEDGMENTS

The authors would like to thank lecturers Erick Orlando Matla-Cruz, Gerardo Avilés-Rosas and Marisol Flores-Castro for their collaboration in this study. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] Azim Abdool, Daniel Ringis, Aniel Maharajh, Lynda Sirju, and Hannah Abdool. 2017. DataRPG: Improving student motivation in data science through gaming elements. In *2017 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–5.

[2] Manal M Alhammad and Ana M Moreno. 2018. Gamification in software engineering education: A systematic mapping. *Journal of Systems and Software* 141 (2018), 131–150.

[3] Andrés Francisco Aparicio, Francisco Luis Gutiérrez Vela, José Luis González Sánchez, and José Luis Isla Montes. 2012. Analysis and application of gamification. In *Proceedings of the 13th International Conference on Interacción Persona-Ordenador*. 1–2.

[4] Pierre Bourque, Richard E Fairley, et al. 2014. *Guide to the software engineering body of knowledge (SWEBOK (R)): Version 3.0*. IEEE Computer Society Press.

[5] Juan C Burguillo. 2010. Using game theory and competition-based learning to stimulate student motivation and performance. *Computers & education* 55, 2 (2010), 566–575.

[6] Laurie Butgereit. 2016. Four NoSQLs in Four Fun Fortnights: Exploring NoSQLs in a Corporate IT Environment. In *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*. 1–6.

[7] Cambridge Dictionary – Motivation 2020. *Cambridge University Press*. Retrieved January 20, 2020 from https://dictionary.cambridge.org/dictionary/english/motivation

[8] Cheng-Huan Chen and Chiung-Hui Chiu. 2016. Employing intergroup competition in multitouch design-based learning to foster student engagement, learning achievement, and creativity. *Computers & Education* 103 (2016), 99–113.

[9] Ching-Huei Chen, Jun-Han Liu, and Wen-Chuan Shou. 2018. How competition in a game-based science learning environment influences students' learning achievement, flow experience, and learning behavioral patterns. *Journal of Educational Technology & Society* 21, 2 (2018), 164–176.

[10] Jose M Cortina. 1993. What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology* 78, 1 (1993), 98.

[11] Rodrigo Cursino, Daniel Ferreira, Maria Lencastre, Roberta Fagundes, and João Pimentel. 2018. Gamification in requirements engineering: a systematic review. In *2018 11th International Conference on the Quality of Information and Communications Technology (QUATIC)*. IEEE, 119–125.

[12] Ali Darejeh and Siti Salwah Salim. 2016. Gamification solutions to enhance software user engagement—a systematic review. *International Journal of Human-Computer Interaction* 32, 8 (2016), 613–642.

[13] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From game design elements to gamefulness: defining" gamification". In *Proceedings of the 15th international academic MindTrek conference: Envisioning future media environments*. 9–15.

[14] Christo Dichev and Darina Dicheva. 2017. Gamifying education: what is known, what is believed and what remains uncertain: a critical review. *International journal of educational technology in higher education* 14, 1 (2017), 9.

[15] Darina Dicheva, Christo Dichev, Gennady Agre, and Galia Angelova. 2015. Gamification in education: A systematic mapping study. *Journal of Educational Technology & Society* 18, 3 (2015).

[16] Wilk Oliveira dos Santos, Ig Ibert Bittencourt, and Julita Vassileva. 2018. Design of tailored gamified educational systems based on gamer types. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, Vol. 7. 42.

[17] GalaXQL 2020. *Game was developed by Jari Komppa, Sol*. Retrieved July 30, 2020 from http://sol.gfxile.net/g3/#

[18] Felix Garcia, Oscar Pedreira, Mario Piattini, Ana Cerdeira-Pena, and Miguel Penabad. 2017. A framework for gamification in software engineering. *Journal of Systems and Software* 132 (2017), 21–40.

[19] Gabriel Alberto García-Mireles and Miguel Ehécatl Morales-Trujillo. 2019. Gamification in Software Engineering: A Tertiary Study. In *International Conference on Software Process Improvement*. Springer, 116–128.

[20] Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work?–a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences*. Ieee, 3025–3034.

[21] Miralem Helmefalk. 2019. An interdisciplinary perspective on gamification: Mechanics, psychological mediators and outcomes. *International Journal of Serious Games* 6, 1 (2019), 3–26.

[22] JetBrains 2019. *PyCharm Professional edition*. Retrieved January 20, 2020 from https://www.jetbrains.com/pycharm/

[23] Karl M Kapp. 2012. *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons.

[24] John M Keller. 2009. *Motivational design for learning and performance: The ARCS model approach*. Springer Science & Business Media.

[25] Fitsum Meshesha Kifetew, Denisse Munante, Anna Perini, Angelo Susi, Alberto Siena, Paolo Busetta, and Danilo Valerio. 2017. Gamifying collaborative prioritization: Does pointsification work?. In *2017 IEEE 25th International Requirements Engineering Conference (RE)*. IEEE, 322–331.

[26] Ana Carolina Tomé Klock, Aline Nunes Ogawa, Isabela Gasparini, and Marcelo Soares Pimenta. 2018. Does gamification matter? A systematic mapping about the evaluation of gamification in educational environments. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*.

2006–2012.
[27] Mehmet Kosa, Murat Yilmaz, Rory O'Connor, and Paul Clarke. 2016. Software engineering education and games: a systematic literature review. *Journal of Universal Computer Science* 22, 12 (2016), 1558–1574.
[28] Richard N Landers. 2014. Developing a theory of gamified learning: Linking serious games and gamification of learning. *Simulation & gaming* 45, 6 (2014), 752–768.
[29] Effie L-C Law and Paul Van Schaik. 2010. Modelling user experience–An agenda for research and practice. *Interacting with computers* 22, 5 (2010), 313–322.
[30] Kun Li and John M Keller. 2018. Use of the ARCS model in education: A literature review. *Computers & Education* 122 (2018), 54–62.
[31] Lila A Loos and Martha E Crosby. 2017. Gamification methods in higher education. In *International Conference on Learning and Collaboration Technologies*. Springer, 474–486.
[32] Liliana Machuca-Villegas and Gloria Piedad Gasca-Hurtado. 2018. Gamification for improving software project management processes: a systematic literature review. In *International Conference on Software Process Improvement*. Springer, 41–54.
[33] Beatriz Marín, Jonathan Frez, J Cruz-Lemus, and Marcela Genero. 2018. An empirical investigation on the benefits of gamification in programming courses. *ACM Transactions on Computing Education (TOCE)* 19, 1 (2018), 1–22.
[34] Erick Orlando Matla-Cruz, Miguel Ehécatl Morales-Trujillo, and David Velazquez-Portilla. 2017. QC: un sistema web para practicar consultas en SQL. In *Jornadas sobre la Enseñanza Universitaria de la Informática*. 95–102.
[35] R de A Maurício, Lucas Veado, Renata Teles Moreira, Eduardo Figueiredo, and Heitor Costa. 2018. A systematic mapping study on game-related methods for software engineering education. *Information and software technology* 95 (2018), 201–218.
[36] Antonija Mitrovic. 2003. An intelligent SQL tutor on the web. *International Journal of Artificial Intelligence in Education* 13, 2-4 (2003), 173–197.
[37] Antonija Mitrovic, Faiza Tahir, and Valerie Sotardi. 2019. Using gamification to effect learning behaviors in intelligent tutoring system. *International Conference on Computers in Education* 2 (2019), 781–784.
[38] Miguel Ehécatl Morales-Trujillo, Gabriel Alberto García-Mireles, and Polina Maslova. 2018. What can go wrong in a software project? Have fun solving it. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 995–1004.
[39] Oscar Pedreira, Félix García, Nieves Brisaboa, and Mario Piattini. 2015. Gamification in software engineering–A systematic mapping. *Information and Software Technology* 57 (2015), 157–168.
[40] Giani Petri and Christiane Gresse von Wangenheim. 2017. How games for computing education are evaluated? A systematic literature review. *Computers & education* 107 (2017), 68–90.
[41] Giani Petri, Christiane Gresse von Wangenheim, and Adriano Ferreti Borgatto. 2017. A large-scale evaluation of a model for the evaluation of games for teaching software engineering. In *2017 IEEE/ACM 39th International Conference on Software Engineering: Software Engineering Education and Training Track (ICSE-SEET)*. IEEE, 180–189.
[42] Paul R Pintrich. 2003. A motivational science perspective on the role of student motivation in learning and teaching contexts. *Journal of educational Psychology* 95, 4 (2003), 667.
[43] Rafidah Ab Rahman, Sabrina Ahmad, and Ummi Rabaah Hashim. 2018. The effectiveness of gamification technique for higher education students engagement in polytechnic Muadzam Shah Pahang, Malaysia. *International Journal of Educational Technology in Higher Education* 15, 1 (2018), 41.
[44] Michael Sailer and Lisa Homner. 2019. The gamification of learning: A meta-analysis.
[45] Radhika Santhanam, De Liu, and Wei-Cheng Milton Shen. 2016. Research Note—Gamification of technology-mediated training: Not all competitions are the same. *Information systems research* 27, 2 (2016), 453–465.
[46] Schema Verse 2020. *The Schemaverse*. Retrieved July 30, 2020 from https://schemaverse.com/tw/index.php
[47] Katie Seaborn and Deborah I Fels. 2015. Gamification in theory and action: A survey. *International Journal of human-computer studies* 74 (2015), 14–31.
[48] Mario Soflano, Thomas M Connolly, and Thomas Hainey. 2015. An application of adaptive games-based learning based on learning style to teach SQL. *Computers & Education* 86 (2015), 192–211.
[49] Mario Soflano, Thomas M Connolly, and Thomas Hainey. 2015. Learning style analysis in adaptive GBL application to teach SQL. *Computers & Education* 86 (2015), 105–119.
[50] Mauricio Ronny De Almeida Souza, Lucas Furtini Veado, Renata Teles Moreira, Eduardo Magno Lages Figueiredo, and Heitor Augustus Xavier Costa. 2017. Games for learning: bridging game-related education methods to software engineering knowledge areas. In *2017 IEEE/ACM 39th International Conference on Software Engineering: Software Engineering Education and Training Track (ICSE-SEET)*. IEEE, 170–179.
[51] SQL Island 2020. *Game was developed by Johannes Schildgen at the University of Technology, Kaiserslautern in the Heterogenious Information System group*. Retrieved July 30, 2020 from https://sql-island.informatik.uni-kl.de/
[52] SQL Murder Mystery 2020. *Knight Lab, The Northwestern University*. Retrieved July 30, 2020 from https://mystery.knightlab.com/
[53] SQL Zoo 2020. *Game was developed by Andrew Cumming, Intelligent Growth Solutions*. Retrieved July 30, 2020 from https://sqlzoo.net/
[54] Judith ter Vrugte, Ton de Jong, Sylke Vandercruysse, Pieter Wouters, Herre van Oostendorp, and Jan Elen. 2015. How competition and heterogeneous collaboration interact in prevocational game-based mathematics education. *Computers & education* 89 (2015), 42–52.
[55] Jirapipat Thanyaphongphat and Patcharin Panjaburee. 2017. Effects of Online Learning with Matching Learning Styles and Preferred Digital Media Types on University Students' Perceptions. In *2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. IEEE, 552–555.

[56] Armando M Toda, Pedro HD Valle, and Seiji Isotani. 2017. The dark side of gamification: An overview of negative effects of gamification in education. In *Researcher Links Workshop: Higher Education for All*. Springer, 143–156.
[57] Juan Vargas-Enriquez, Lilia García-Mundo, Marcela Genero, and Mario Piattini. 2014. A systematic mapping study on serious game quality. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*. 1–10.
[58] Juan Vargas-Enriquez, Lilia Garcia-Mundo, Marcela Genero, and Mario Piattini. 2015. A systematic mapping study on gamified software quality. In *2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games)*. IEEE, 1–8.
[59] Julita Vassileva and Lingling Sun. 2007. Using community visualization to stimulate participation in online communities. *E-Service Journal* 6, 1 (2007), 3–39.
[60] Christiane Gresse von Wangenheim, Rafael Savi, and Adriano Ferreti Borgatto. 2012. DELIVER!–An educational game for teaching Earned Value Management in computing courses. *Information and software Technology* 54, 3 (2012), 286–298.
[61] K Werbach and D Hunter. [n.d.]. ve Hunter, D.(2012). For the win: How game thinking can revolutionize your business.
[62] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in software engineering*. Springer Science & Business Media.

A PRE- AND POST-TEST QUERIES

A.1 Pre-test

The purpose of the pre-test was to determine the actual level of the SQL knowledge that the students had before the intervention.

The pre-test queries were as follows:

0. Retrieve all players.

1. Total of cities grouped by climate.

2. Find all stadiums whose ground is 'Natural' and whose seat capacity is between 60,000 and 93,000.

3. List full name and position of the coaches who coach a team with name beginning with 'C', 'W' or 'R' and the foundation year not later than 1980.

4. List full name, university and date of birth of the 'runners' with the least carries.

5. List full names of the players who: are 'receivers' with more than 100 receptions and playing for a team founded after 1960; or are 'runners' with more than 220 carries and exactly 3 fumbles.

6. Retrieve the total pancakes made by 'offensive line' players who were born after 1980. The result must be grouped by position and sorted by number of pancakes in the descending order.

7. Find full names of the 'runners' who have played in more than one team since 1990.

8. Find the team that has won the most home games.

9. Considering the five cities with the largest population, retrieve the cities with the least number of teams playing there.

A.2 Post-test

The post-test queries were as follows:

0. Retrieve all coaches.

1. Total of players grouped by last name.

2. Find all the games played in week 17 at '4:05pm'.

3. List full names and positions of the coaches coaching a team founded in 1960 or whose second letter in the team's name is 'i'.

4. List full name, university and date of birth of the 'kickers' with the most blocked kicks.

5. List full names of the players who: are 'runners' with more than 13 touchdowns and playing for a team founded before 1975; or are 'punters' with less than 30 punts and exactly 3 punts 'inside 20'.

6. Retrieve the total pancakes made by 'offensive line' players who were born before 1990. The result should be grouped by position and sorted by number of pancakes in ascending order.

7. Find full names of the 'quarterbacks' who have played in more than two teams since 1995.

8. Find the team that has won the most road games.

9. Considering the five cities with the largest population, retrieve the cities with the least number of teams playing there.

B COURSE CONTENT AND PHASES OF THE STUDY

The content of the lectures given in both groups and the phases of the study were scheduled as follows:

- Week 1: Introduction to Databases
- Week 2: Database Architectures and Data Models
- Week 3: Entity-Relationship Model
- Week 4: Relational Model
- Week 5: Database design (ER to Relational)
- Week 6: Database design (Applied problems)
- Week 7: Relational Algebra
- Week 8: Data Integrity
  Phase 1: Pre-test
- Week 9: Data Definition Language
  Phase 2: Round 1
- Week 10: Data Manipulation Language
  Phase 2: Round 2
- Week 11: Query Optimization
  Phase 2: Round 3
- Week 12: Functional Dependencies
  Phase 3: Post-test and survey
- Week 13: Normalization
- Week 14: Transactions
- Week 15: Database Security
- Week 16: Selected Topics

C IMPLEMENTATION OF GAME ELEMENTS

The game elements were implemented as follows: the students' answers were stored in a table called `Answer`. This table has relationships with the `Student`, `Query` and `Round` tables. For simplicity, all the attributes required for the marking and the game elements implementation are presented as part of the `Answer` table (see Figure 8).

Each answer has an unique ID (`answer_id`) and is associated to a student (`student_id`), round (`round_id`) and a challenge (`query_id`). The opening and closing dates and times for each round are stored in the `Round` table. The `Student` table only relevant data is the identifier of each student. The `Query` table stores the challenges, their main

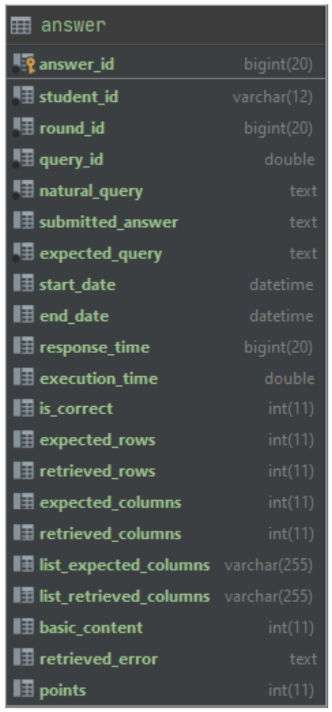

Fig. 8. Answer table.

attributes are `natural_query`, which stores the statement in natural language of the query, and `expected_query`, which stores the expected answer as an SQL statement. The latter is the statement that is executed during the marking process to retrieve the result set which will be compared against the result set obtained from the student's answer (`submitted_answer`).

Answer tuples are created when a participant opens a challenge. At this point a new tuple is inserted with the identifier of the answer (`answer_id`) and the actual time the student clicked on the challenge (`start_date`). Since this process is executed every time that a challenge is opened, it is possible to know how many challenges were seen by the participants.

Once the student clicks 'Submit', the respective tuple is updated with the student's answer, the actual time the student clicked 'Submit' (`end_date`), and the time in seconds that the student spent solving the query (`response_time`). This process is executed as many times as answers are submitted; at this point, information on how many challenges have been solved and how many were seen but not tried (no answer submitted) is available.

When a round is closed, the marking script is executed and each tuple is updated as follows: the answer (`submitted_answer`) is executed and the execution time is calculated in milliseconds (`execution_time`); the result set is calculated and compared against the expected result set, setting it to '1' if they match, otherwise to '0' (`is_correct`). From the execution of the expected and submitted answers other relevant information about the answers' structure is retrieved: the rows (`expected_rows` and `retrieved_rows`), and the columns (`expected_columns` and `retrieved_columns`).

The script also identifies if the query contains the SQL keywords `SELECT` and `FROM` (`basic_content`). This is particularly useful in incorrect answers in order to know if the answer is a genuine attempt or just a random text that the student submitted without any intention to answer the challenge. In the same line, if the execution of the answer causes an error, the error message is stored (`retrieved_error`).

Lastly, the marking process finishes when the average response time and the average execution time are calculated, and the points are allocated (`points`). The leaderboards are created by adding the points up, per round and at the end of the whole competition.