Computers in Human Behavior 64 (2016) 247–263

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

A gamified peer assessment model for on-line learning environments in a competitive context

Thyago Tenório [a, *], Ig Ibert Bittencourt [b], Seiji Isotani [c], Alan Pedro [b], Patrícia Ospina [d]

[a] Federal University of Alagoas (UFAL), Campus Arapiraca/Pólo Penedo, Av. Beira Rio, 57200-000 Penedo, AL, Brazil
[b] Computing Institute, Federal University of Alagoas (UFAL), Campus A.C. Simões, Cidade Universitária, 57072-970 Maceió, AL, Brazil
[c] Institute of Mathematics and Computer Science, University of São Paulo (USP), Avenida Trabalhador São-carlense, 400 Centro, 13566-590 São Carlos, SP, Brazil
[d] Center of Exact and Natural Sciences, Federal University of Pernambuco, Av. Prof Moraes Rego, 1235 — Cidade Universitária, Recife, PE, CEP 50670-901, Brazil

ARTICLE INFO

Article history:
Received 5 April 2016
Received in revised form
9 June 2016
Accepted 25 June 2016
Available online 9 July 2016

Keywords:
Gamified peer assessment model
Gamification
Competitive on-line learning environments
Modelling
Regression model
Experiments

ABSTRACT

Peer Assessment (PA) offers a powerful solution that helps teachers in online learning environments to correct essays by distributing the workload among students. Nevertheless, the quality of the results in PA depends on good evaluations of reviewers. Thus, the main drawback for scaling up the use of PA is the presence of inadequate behaviours, such as being too harsh or too soft in the assessment, or even not offering a helpful feedback. This usually occurs due to the lack of motivation and engagement of students in the PA process. To deal with this problem, this paper proposes a gamified peer assessment model, where gamification elements are used to engage students in PA activities. Two experiments using the proposed model within an intelligent tutoring system called MeuTutor shows satisfactory outcomes. We verified that the average grade given by students to an essay are equivalent to those given by experts, but the time and costs to complete the assessments were largely reduced. Furthermore, the use of gamification helped to increase the amount of students' access to the system in 64.28%; increase in 10.53% the number of essays written and submitted; and improve the quantity and quality of assessments for each essay.

1. Introduction

In recent years, several countries have adapted their educational approaches to promote and support the use of technologies on both classroom courses and on-line learning courses. Currently, there are several technologies that support online learning, such as Intelligent Tutoring Systems (ITSs) Sleeman and Brown (1982), Virtual Learning Environments (VLE) (eg Moodle, SOLAR, TelEduc, Blackboard, Sakai), Computer-Supported Collaborative Learning (Miyake, 2007) and more recently, the Massive Open On-line Courses (MOOCs) (Wulf, Blohm, Leimeister, & Brenner, 2014) (eg EDX, Coursera, Udacity).

The recent increase in popularity of MOOC courses has made it accessible for anyone with an internet connection to enroll freely, university level courses (Piech et al., 2013). However, while new web technologies allow for scalable ways to deliver video lecture content, to implement social forums and to track student progress, we remain limited in our ability to evaluate and give feedback for complex and often open-ended student assignments such as mathematical proofs, design problems and essays (Piech et al., 2013). This is because these types of activities require manual correction by the teacher individually, which makes it an over-charging activity. With the increase in the students amount in these environments and, consequently, a greater number of activities being made every time, the correction by teachers would quickly become infeasible.

Facing this issue, Peer Assessment offers a promising solution to scale the grading of complex assignments in courses with tens or even hundreds of thousands of students (Piech et al., 2013). It is an educational arrangement where students judge peers performance quantitatively and/or qualitatively (Van Zundert, Sluijsmans, & Van Merriënboer, 2010). It stimulates students to reflect, discuss, and collaborate in their learning process (Topping, 1998) (Strijbos &

* Corresponding author.
E-mail addresses: thyago.oliveira@penedo.ufal.br (T. Tenório), ig.ibert@ic.ufal.br (I.I. Bittencourt), sisotani@icmc.usp.br (S. Isotani), alanpedro@ic.ufal.br (A. Pedro), patespipa@de.ufpe.br (P. Ospina).

http://dx.doi.org/10.1016/j.chb.2016.06.049
0747-5632/

Sluijsmans, 2010). Peer assessment is a process by which students or their peers attach grades or tests based on predefined benchmarks by professor (Sadler & Good, 2006). In peer assessment, students learn from each other by means of receiving and giving feedback. It is recommended because it reduces teacher's workload (Rubin & Turner, 2012), increases learning outcome (Murakami, Valvona, & Broudy, 2012). Due to its efficiency and active learning nature, peer assessment has been widely used in diverse fields (Falchikov, 1995) (Freeman, 1995).

Peer Assessment models have been used in various ways, improving teaching skills and even providing emotional benefits Sadler and Good (2006). This practice is used to save teachers time and improve students' understanding about the course materials as well as improve their meta cognitive skills (Malehorn, 1994). There are several peer assessment approaches published in the literature, such as formative approaches (Orsmond, Merry, & Callaghan, 2004), probabilistic models (Piech et al., 2013) and even models using Bayesian networks (Wang & Vassileva, 2003).

The effectiveness and quality of an assessment depends on how it is incorporated into the learning process (Schuwirth, 2004). However, students may not have enough knowledge to criticize peers work and conduct a fair evaluation (Wang, Liang, Liu, & Liu, 2014). To alleviate this problem, it was developed a peer assessment process that uses assigning multiple reviewers to an evaluation task, decreasing the bias (Tseng & Tsai, 2007). Thus, Peer Assessment become a collaborative evaluation process, where the quality of the final results depends on good evaluations of their reviewers. It values cooperation over competition and greater respect for the varied experiences and backgrounds of participants can occur (Boud, Cohen, & Sampson, 1999).

However, in some competitive on-line learning environment has the fact that when some students achieved the goal, all other students fail to reach that goal, only the best students will be victorious. As a result, students are not motivated to collaborate with the reviews of the activities. Students are used to being judged in terms of their own efforts and can resent others gaining credit for what they perceive as their own contributions, particularly within the context of a competitive course (Boud et al., 1999), i.e., there is a certain fear to evaluate other students.

The fact is that there is great reluctance to collaborate with those who are competing against. Facing that, this leads to the following question “How can we include Peer Assessment techniques in online educational environments in a competition context?”. This technique applied in this context is especially promising as destabilizing the passive role of the student, such that it take responsibility for their learning, seeking the improvement of the learning process through their active participation. However, a major problem when using it in the competitive context is the presence of inappropriate behaviour by students which decreases the learning and evaluation system (Kapp, 2012). The emergence of this type of behaviour occurs for various cognitive and emotional factors such as boredom, lack of motivation and the need to get results quickly. To solve this problem it is necessary to use techniques and models that have the ability to “engage” positively the emotional state/cognitive of the student, without necessarily increasing the deployment cost.

The paper proposes a gamified peer assessment model that uses gamification elements as a motivational aspect for students inside Peer Assessment process. Gamification is the use of mechanical, ideas and aesthetics games (context, fast feedback, competition, stages, achievements, points, et.), to engage people, motivate actions, promote learning and solving problems (Kapp, 2012). This term is commonly used to express the use of game elements (storyline, score, levels, quests, badges and rankings) in environments that are not games (educational environment) to motivate or influence people to perform a certain activity. The use of gamification applied in education is strongly presented in (de Sousa Borges, Durelli, Reis, & Isotani, 2014), as a motivational aspect to students. Some studies, such as (Pedro, Lopes, Prates, Vassileva, & Isotani, 2015) applied gamification in their environments and results indicate that the gamification implemented contributed to improve student performance in the case of boys.

Andrade, Mizoguchi, and Isotani (2016) points out that several positive effects of using gamification in learning environments has been found to date. The combined use of peer assessment techniques along with gamification makes the process most powerful and complete, avoiding and/or decreasing the presence of inappropriate behaviour by students. The gamification elements applied in the model positively influence the state of the student, encouraging them to participate in the proposed teaching-learning process through the rewards obtained (points, levels, trophies, among other).

The gamified peer assessment model proposed was applied in the MeuTutor educational environment, which is an intelligent tutoring system and aims to monitor the learning of students in a personalized way, ensuring quality in teaching and improving the performance of its members. The version chosen to be used was the MeuTutor-ENEM, which aims to help high school students prepare for the National High School Exam (ENEM). Thus, the environment offers courses related to high school subjects like Portuguese, mathematics, physics, among others. Our goal was evaluate the effectiveness of the use of our Gamified Peer Assessment model in the context of competition in the correction of essays.

We structured this document as follows: in section 2, we present the related works. In section 3, we present our proposal and the concepts created. Section 4 presents the planning and execution of an experiment, where we have applied the proposed model in the MeuTutor. Finally, section 5 summarizes the work, presenting the conclusions we reached, our limitations and some works planned for the future.

2. Related work

This section aims to discuss related work to the proposed work. We have seen that the use of peer assessment is frequently. Peer Assessment can be defined as a learning setting in which individuals evaluate or comment on the amount, level, value, quality, or success of the products or learning outcomes of the peers who learned in a similar context (Topping, 1998). The main goals to use Peer Assessment are “improve the quality of the learning process, sharpen critical abilities in students, and increase student autonomy” (Topping, 1998).

Several studies have reported that learners could receive a great deal of inspiration from peer assessment results (Chen, 2010), which could encourage their learning motivation (Jenkins, 2004), enhance their thinking capability (Prins, Sluijsmans, Kirschner, & Strijbos, 2005), facilitate their self-reflection and communication capabilities (Min, 2006) and improving learning achievements, motivations and problem-solving skills (Hwang, Hung, & Chen, 2014).

We can classify the works that use peer assessment into three groups. The first group includes those who use the peer assessment technique in a specific context, according to the specific needs of the task to be performed. In this case they are highly dependent of the context where was applied. The second group consists in the work that proposes tools that assist in the application of peer assessment in other tasks and contexts. These works proposes generic tools that do all the work necessary to apply peer assessment in a given context and/or task. The third group consists of works that proposes peer assessment models. In that case, they are

usually generic models that can be adapted to the specific needs of the application of peer assessment that are necessary for correct application.

Inside the first group mentioned above, we can cite some works as in (Dominguez, Cruz, Maia, Pedrosa, & Grams, 2012) where the technique is applied for activities evaluations in an engineering course. Chang, Tseng, and Lou (2012) applies it to evaluate portfolios of high school students who study computing. Also in this context, Kawai (2006) uses it for evaluation of voice messages and letters of foreign language students. Another study belonging to this group is the Hwang et al. (2014). In this study, a peer assessment-based game development approach is proposed for improving student's learning achievements, motivations and problem-solving skills. It was concluded that the proposed approach could effectively promote student's learning achievement, learning motivation, problem-solving skills, as well as their perceptions of the use of educational computer games. However, in these studies the use of peer assessment is strongly linked to the context in which it was developed and proposed.

For the second group (tools), the works present some own tools built with the purpose of supporting the implementation of peer assessment. In Sterbini and Temperini (2013) is presented a tool called OpenAnswer, that aims to enable support to written evaluations using peer assessment techniques, something similar to proposed in this work. However, this tool is not integrated with the environment, which would mean that users had to use another platform only for the written evaluations activities. In addition, these tools are inflexible and can not be extended and incremented with new features to handle specific needs of the application of peer assessment that by chance could arise when it is applied in specific context.

In addition to the above work, Cunha and Figueira (2009), Gouli, Gogoulou, and Grigoriadou (2008), Miao and Koper (2007) and Trahasch (2004) also proposes tools. These tools support features beyond the basic functions of sending/comments of activities by students and classification/review of them. They also support features such as individual and collaborative development of activities and evaluating the activities of one or a group of students. Despite these resources that may be useful in specific needs, these tools are still inflexible and have the same problems presented in Sterbini and Temperini (2013).

Trying to solve the problems of the previous groups, the third group of works consists of proposing peer assessment models. In (Kahiigi Kigozi, Vesisenaho, Hansson, Danielson, & Tusubira, 2012), the authors designed and modelled a review process based on peer assessment for collaborative learning. The model is based on student usage, pedagogically supporting their learning. In Tosic and Nejkovic (2010), a new method for peer assessment of the students based on the concept of confidence is proposed. Other works proposed models based on Bayesian networks as in Sterbini and Temperini (2013) and Wang and Vassileva (2003). All these models are implemented in their own way and used according to your needs, but is not intended to integrate with the various educational environments and therefore differ that work in this regard. Finally, we mention the work of the founders of Coursera (Piech et al., 2013). Coursera heavily uses the concept of peer assessment to provide on-line courses to thousands of people around the world. However, in his work, it is presented only parts of the reputation algorithms and calculation of final grades.

Moreover, none of these tools or models presented above are geared to the specific requirements of environments involving competition, which enhances the lack of motivation of the student to participate in the evaluation process. Lack of motivation is one of the main problems that exist when using peer assessment, since students will work more, reflecting their emotional/cognitive state. This lack of motivation leads to inappropriate behaviour by peers of the students, which compromente the quality of evaluations and the process as a whole, as highlighted in the work of (Fermelis, Tucker, & Palmer, 2007), where students reported complete dissatisfaction. This is one of the main problems cited when using peer assessment in Wang et al. (2014), Topping (2009), Tseng and Tsai (2010) and Moccozet, Tardy, Opprecht and Leonard (2013)).

In this paper, we proposed a gamified peer assessment model. The aspect of modelling ensures greater flexibility than tools and allows for greater integration with the educational environments. In our proposal, we want that the use of peer assessment is included directly on environment, without the need of other systems. However, a model requires more time to be developed than these tool. Also, the fact we propose a model allows the solution is extensible and adaptable to certain specific needs of the teacher that perhaps will be necessary. Moreover, the use of gamification techniques as motivational aspect within Peer Assessment is one of the great advantages of this work, allowing greater engagement of students and motivating them to participate in the process.

In Moccozet et al. (2013), it also used an idea similar to gamification. The basic idea consists in adapting the user points approach in order to estimate student's individual contribution to the global effort of the learning community. When gamifying an activity with user points, the objective is usually to engage users to earn as many points as possible and thus create a competition among users. Therefore points are displayed on user's profile, the amount of points each action can bring is displayed, and users are publicly ranked according to the amount of points they have earned. This is where his approach deviates from the traditional gamification. The points are only used to attribute a value to an action. The goal is not that points directly arouse activities but

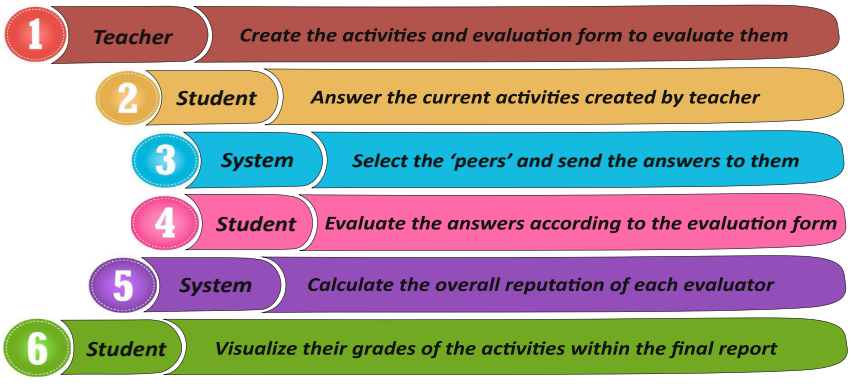

Fig. 1. Model in details.

rather that they estimate the level of activities.

Hamari (2013) gamified a utilitarian peer-to-peer trading service by implementing the game mechanism of badges that users can earn from a variety of tasks. The results show that the mere implementation of gamification mechanisms does not automatically lead to significant increases in use activity in the studied utilitarian service, however, those users who actively monitored their own badges and those of others in the study showed increased user activity.

3. The peer assessment model

This section aims to describe the model. The use of peer assessment model begins after the availability of resources to students in the system and their proper training. The first step is to define the educational resource to be used by the teacher. The resource will then be available to students, and these will do them and send their answers to the system (in the second stage - the submission step). The time devoted to these submissions is fixed and predefined by the professor.

Next, the students receive some activities (from their peers) and will correct them, using the activity's evaluation form created by the teacher. The deadline for corrections is also pre-defined by him. In this third step, the students attached grades to their peers without necessarily knowing the author and the others grades of the activity.

After the third step, the system will manage the grades obtained from each student on each activity and calculates the final grade of the activity (in fourth step). In this step, the final grade will be presented to students in their respective activities. Finally, in the fifth and final step, students visualize the feedbacks and identify their weaknesses and reflect on the mistakes in that activity.

To better understand how the model behaves within the system, Fig. 1 shows the steps of execution that must be incremented in the educational environment to include support for written evaluations.

As can see in Fig. 1, the first step is the creation of discursive activities done by the teacher together with the creation and definition of the evaluation form containing the criteria on which the activities will be evaluated by students. A discursive activity consists of a statement and one or more questions (open alternatives), where the student will answer. On the other hand, an evaluation form is composed of a title and a set of questions for evaluating the activity and each question contains one or more evaluation criteria. These criteria can be:

- Boolean: It is a criterion that only accepts Boolean values, ie, true or false. Useful for eliminating questions;
- Fixed group: Criteria similar to the type Boolean, but accepts a certain group of fixed values. It is a selection criterion values;
- Interval: This type of criteria is used only with numbers. It's just to set a minimum and a maximum value allowed. Thus, any value between the minimum and the maximum shall be permitted;
- Text: Free criteria, accepting any amount of text. Used a lot for comments.

The activities and the forms should be sent (registered) to the system by the teacher on a specific interface for this. Once the activity is available to the student to answer, begins the period of submissions - second step. Students answers the activities in own on-line environment. For this, the student must log in and select the activity, which will be available together with other resources/activities. After the chose of the activity, the system will present it to the user, who will respond with his text.

Then, the student will submit its responses to the environment, witch is responsible for managing the correction of them. In the third step, the system identifies students who will correct the submitted activity and, from that moment, the system will send the activity for these students. This process of choice of users should be done by using a list of users who have made the same activity and also has pending evaluation, in ascending order of assigned activities. Several studies indicates algorithms to accomplish this distribution (Moreira, 2014), Cavalcanti (2008) and (Sung, Chang, Chiou, & Hou, 2005). This process of choice of users was made using the following pseudo-algorithm 1 (created by author):

Algorithm 1 Retrieve Users

```
1:  function RETRIEVEUSERS(minCorrection, safetyMargin, maxPendentCorrection)
2:
3:  */ minCorrection - Minimum number of corrections per student; safetyMargin - Margin of safely below the search
    users; maxPendentCorrection - Maximum number of pending correction activities for a single user /*
4:
5:      List<User> users = retrieveUsersMadeActivity(idActivity, minCorrection + safetyMargin, "Ascending or-
    der");
6:      for i do 1 until size(users)
7:          qtdPendentCorrection = getAmountPendentCorrectionByUser(users.get(i));
8:          if qtdPendentCorrection > maxPendentCorrection then
9:              users.remove(user.get(i));
10:         end if
11:     end for
12:     if size(users) >= minCorrection then
13:         return users.sublist(0,minCorrection-1);                    ▷ Return the minCorrection first list elements
14:     else
15:         List<User> moreUsers = retrieveUsersHighProbabilityCorrect(minCorrection - size(users));
16:         users.addAll(moreUsers);
17:         return users;
18:     end if
19: end function
```

It is noteworthy that the minimum number of necessary corrections (numMin from Step 1) for an activity is very dependent of the type of activity being performed. Theoretically, greater number becomes more reliable to the evaluation results. On the other hand, greater number generates more work to students, which can be prejudicial to the final result of the model. So there is a trade-off in choosing this number. The number of safety margin (numFolg from Step 2) serves as a store of users who will be retrieved for analysis.

The Higher the number of students, the higher the accuracy and the slower the algorithm. In turn, the maximum number of pending correction activities to a single user (numMax from Step 3) should also be chosen carefully, since it can overload the involved students.

The minimum number of selected students was 2. The reason was that with two corrections it is possible to evaluate the results if they are similar. If there is disagreement, a third student is associated. The final grade is the average of the two closer students grades. On the other hand, for the model to work properly the maximum number of pending activities was chosen as 3 because we expect a user to correct at least three essays, with a minimum recommended of two essays.

Then, in the fourth step, the students, according to the distribution in the previous step, will evaluate the answers of their peers based on pre-defined evaluation form by the teacher. The deadline for the review (correction) of the activities in the system is called evaluation period. This is when students evaluate their peers and will send the grades to the system. To assign his grade, the student will receive the answers of a particular student and the evaluation form with pre-defined criteria. Then, it will analyze the activity from the point of view of the criteria (as a teacher would do) and then simply fill out the form with their answers. This process will be repeated as long as there are existing activities to correct for that student. According to the presented algorithm, this number is expected to be an average of the students, to avoid overloading any of them. However, there may be slight variations.

In the fifth step, in order to have greater confidence in the final grade, it should be evaluated the overall reputation and the level of competence of each evaluator for each criterion. The calculation of the overall reputation of an evaluator should consider the following activities performed by a student (where + is a positive aspect and - means a negative aspect):

- + Frequently access to the system (A);
- + High number of discursive activities performed (QP);
- + High number of activities corrections performed (QC);
- − Low frequency of access to the system;
- − High number of pending corrections (PC)
- − High number of activities performed with low number of corrections

This process of calculating the overall reputation (OR) of the student X happens with the following equations:

$$FA = \frac{A(X) - AVG(A)}{MAX(A)} \quad FQP = \frac{QP(X) - AVG(QP)}{MAX(QP)} \quad FQC = \frac{QC(X) - AVG(QC)}{MAX(QC)}$$

$$OR = \frac{FA*1 + FQP*2 + FQC*2}{5}$$

The interval for the functions FA, FQP and FQC is $-1 \leq F \leq 1$. A higher access number, amount of discursive activities performed and amount of corrections performed imply closest 1 shall be the functions values. Similarly, lower values imply values close to −1. If the student is on average, the functions values will be close to 0. Note that the overall reputation (OR) function has a behaviour similar and it is formed by a weighted average of the previous functions. Then, to calculate his competence's level in a particular criteria, the activities history made should be looked. If the student have no data, their level will be considered neutral (0). If they have data in history, this should be used to determine if they has a positive and/or negative aspect.

In addition to the general reputation in the system it is necessary to calculate the reliability in the ratings of the users of the system. This calculation is of great significance in the process. So we decided that we would use the calculation proposed by Piech et al. (2013), based on probabilistic models. We selected this model calculation because it was designed with purpose similar to that of our model and is currently used in Coursera, one of the on-line education environments with thousands of users.

In summary, we used the calculation PG1 to calculate of "Grader bias" and reliability. In this case, a probability distribution is assumed a priori to the latent variables and uses the premise that the bias of an individual student may be other than 0, but the mean of many raters always tends to zero. In summary, the formulas used are:

$$(Reliability)\tau_v \sim \mathscr{G}(\alpha_0, \beta_0) \text{ for each evaluator } v,$$

$$(Bias)b_v \sim \mathscr{N}(0, 1/\eta_0) \text{ for each evaluator } v,$$

$$(RealGrade)s_u \sim \mathscr{N}(\mu_0, 1/\gamma_0) \text{ for each user } u \text{ and}$$

$$(ObservedGrade)z_u^v \sim \mathscr{N}(s_u + b_v, 1/\tau_v) \text{ for each pair of score observed},$$

where $\mathscr{G}$ refers to gamma distribution with hyperparameters α_0, β_0 fixed, while η_0 e γ_0 are hyperparameters to priori distribution of biases and real grade, respectively.[1]

From the defined calculations, the system calculates the reliability of each assessment, the competence's level of evaluator in each criteria and their reliability and can then calculate the final result as a weighted average.

Finally, in the sixth step, the system will generate a report containing the results of the evaluation to be presented to the student who submitted the activity. This report should be detailed, with the grades of each criteria as well as the final result of the activity. It's possible the including of comments and possible changes on the grades from the teacher before your presentation to the student.

From the information presented above regarding the execution flow, we can summarize the model into a set of necessary steps to the peer assessment process works properly. These steps are illustrated in Fig. 2.

It can see that the model takes as input the set of answers of student activities. The first step is to determine, according to the conditions above, to who every activity response will be given to correct (distribution of activities). This will generate, as a partial result, a partial list of grades for each answer given by students. This list will serve as input to the next stage of the model, which aims to

[1] Information on how these formulas were obtained can be found in Article Site, available in http://web.stanford.edu/~cpiech/bio/papers/appendices/edm13_appendix.pdf.

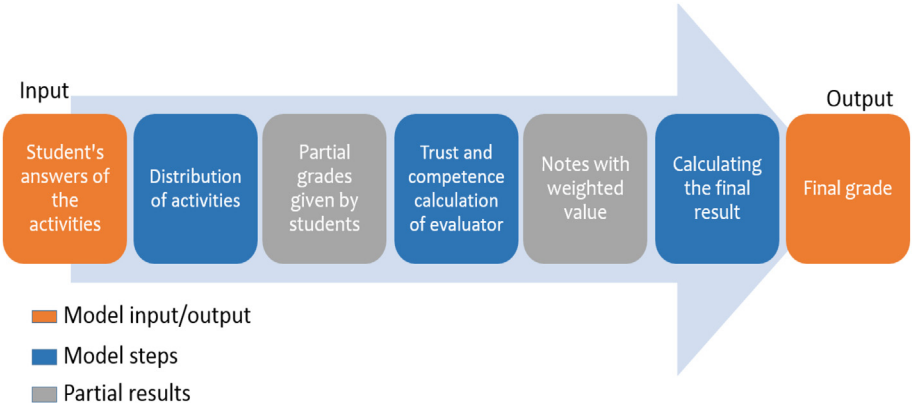

Fig. 2. Model's steps.

calculate the confidence and the competence of the each evaluator, assigning a weight to the grade. In this sense, we as a second partial result the list of notes with weighted value. In sequence, the last stage of the model is to calculate the final result for each activity response given using their weighted scores, thus generating model output with these final results.

This entire process is kept in functioning through the use of gamification techniques, which aims to increase the motivation of users.

3.1. The gamification into the model

This section aims to describe how gamification is included in the model. The creation of a gamification plan requires you to think of many details in order to have a success in its implementation. These details should be well planned and well founded. The set of all of them are included in a document called gamification plan. There are several ways of how to create this plan, but this work will be presented only the framework of gamification 6D, available in a course of the Coursera, by Kevin Werbach of the University of Pennsylvania. This framework was chosen for its simplicity and high coverage of the main required points to describe a gamification model. Also, it is characterized as an agile approach. The 6D framework is a design process with the next six elements, steps or premises: Define Business Objectives; Delineate target behaviour; Describe yours players; Devise activity loops; Don't forget the fun; Deploy appropriate tools.

In the first time we must define the business goals of gamification, that is, what is our final goal of including gamification in the proposed model? The main objective of gamification is to ensure students to feel motivated to participate on both performing activities and correcting them. In this sense, we defined the following targets behaviours:

- Encouraging students to participate more often in activities;
- Challenging students to get badges;
- Providing medals to students who participate in activities;
- Provide badges or points to students for answer/correct activities;
- Promote a ranking parametrized for students;
- Encourage competition for students, with rewards to the winners;
- Rewarding students for their unexpectedly actions in the system.

The act of participating more often of the activities to be performed indicates a greater capacity for student to learn and consequently to achieve the goal in question. To challenge such students to get some trophies, especially those related to learning activities, will encourage students to participate. Through the promotion of medals to students, trophies and list the ones Which to make more activities in rankings, will serve as a motivational aspect for those students who become demotivated, including even those who for some reason gave up of performing such activities.

The tendency in the system of evaluations is that students always competes with their friends. In this sense, we have as a target behaviour to stimulate competition, considering that a healthy competition leads to better outcomes among students. Thus, those students who come out best in the competition will receive rewards. In addition, unexpected actions in the system can also reward students, which will further motivate their participation. All these behaviours lead to students feel challenged by the system as well as by their peers, which leads to realization of the proposed tasks.

Users will be students of on-line courses. Generally, they do not have any kind of relationship with each other. The main players of our model would be explorers and especially the collaborators. Explorers because they tend to perform all actions in the system, since they like to use all the resources involved. Collaborators, in turn, because they have the instinct to help their friends, which is a very strong and necessary point for a peer evaluation system to function adequately. However, we also adapted the model for competitors.

Then, we defined the loops to be performed by users. Basically an activity loop has three main components: motivation, action and reward. The motivation is the goal of the loop, that is, its purpose to be created. The action is the desired behaviour for the user. And finally, the reward is what they will win if he has the desired behaviour. Table 1 presents the loops defined in our gamification model to achieve the goal.

The gamification elements that were used are points, badges, missions, rankings and medals. All information of how to get them are detailed in Table 1. These elements are highlighted as those can help and motivate students enrolled in online learning environment (de Santana et al., 2016).

4. Controlled experiment

Just as all scientific research, it is needed a way to validate what is being proposed in order to present their positive results as well as possible negative outcomes. In order to assess the effectiveness, we designed and performed a controlled experiment in MeuTutor environment context.

The model was implemented and integrated with the

Table 1
Activity Loops defined in our gamification model.

Number	Type	Action	Reward
1	LP	To Perform 3 activities	Points
2	LP	To correct 3 activities	Points
3	LP	Log in for 3 consecutive days	Points
4	LE	Reply activity	Points
5	LE	To correct activity	Points
6	LE	Log in	Points
7	LE	Unexpected action	Points
8	LP	Reply 5 activities	Badge "The Writer"
9	LP	To correct 5 activities	Badge "The Evaluator"
10	LE	Staying in the top 10 ranking	Badge "On top"
11	LE	Stay in the top three of an activity	Gold, Silver or Bronze Medal

Fig. 3. Initial home screen to written evaluations to a student in MeuTutor.

educational environment MeuTutor. It was chosen as a good scenario because of its deployment flexibility and integration of new components in the system and have an environment already prepared for gamification, making easy to integrate the model on the system. In addition, this environment had a specific requirement (with written evaluations), which will be better detailed below.

The educational environment MeuTutor is an intelligent tutoring system, which aims to monitor the learning of students in a personalized way, ensuring quality in teaching and improving the performance of their students. The version chosen to be used was the MeuTutor-ENEM, which aims to help high school students to prepare for the National High School Exam (ENEM). Thus, the environment offers courses related to high school disciplines such as Portuguese, Mathematics, Physics, among others.

Initially, MeuTutor did not support the written evaluations. For this reason, the environmental creators included our model, allowing that their students could practice essays within the system. From the perspective of the student, this when entering on the system and click on writing on the top menu bar, he will see whether there is any writing activity available in the system to be answered by him, as illustrated in Fig. 3.

Fig. 3 shows the student who is in the first step of the evaluation process (submission stage), where he will perform (answer) the proposed activity. As we can see, above is the title of the essay that is available. Students starts their activities clicking on "Start Writing" button. After clicked, the student will see the entire proposal of the written activity and must to write their responses directly on the screen. In addition, the system presents some instructions that the student must comply when he is creating their essay. The restrictions are that the student must answer at least six lines and the inability to copy texts in the present proposal. Additionally, he receives score zero if flee the theme or not to write argumentative text, contrary to human rights and/or deliberately write topics disconnected from the presented topic text.

After student to submit his answer to the system, he must wait the deadline for the start of the second step - the correction. While the process does not start, it will appear to the student a timer indicating how much time is left for the start of corrections. When

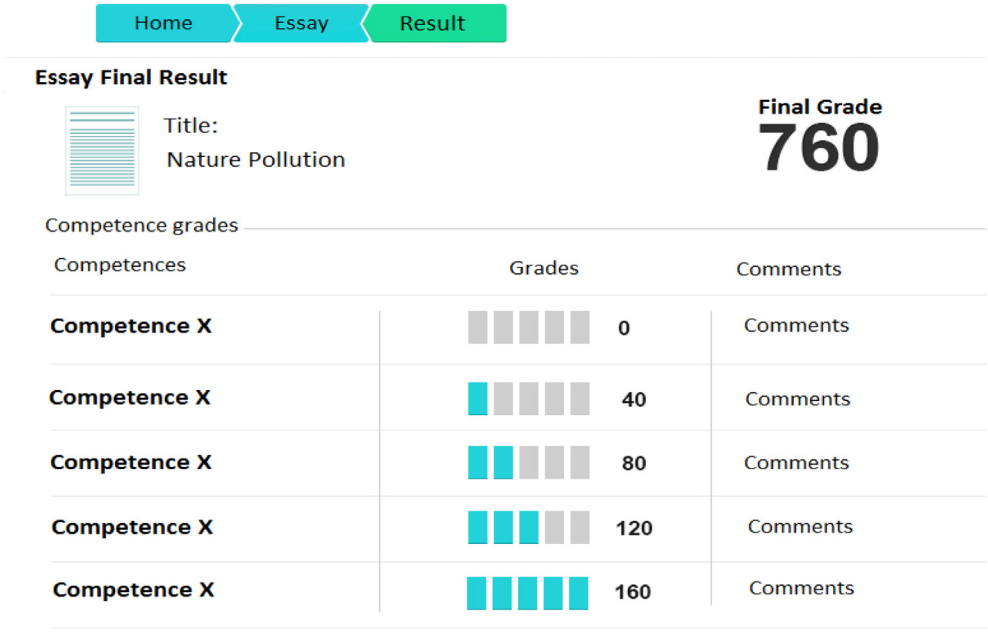

Fig. 4. Report screen to a student in MeuTutor.

the submission deadline is over, the second step of correction of essays begins, where the student will see there are other essays to review (usually two or three others). In this case, for each one, the student should review it along with the evaluation form created by the teacher. Students should then calmly and critically read the writing of his partner, judging it under the predefined criteria by the teacher and then must fill out the answers by clicking in the end button "Submit feedback".

Before sending the answers, the system asks for confirmation of the student. This process must be repeated by the student until he was completed the correction of essays attributed to him. In the end, there will be nothing to be done in the second stage of correction. Finally, after the deadline of the second stage, all essays will be with the corrections completed. It will be initialized the third stage - which is the presentation of the final results. The student can then view his final result report, with the grades on each criterion assessed and the final result, according Fig. 4.

The environmental business problem was that the use of Peer Assessment as a correction model of the discursive activities (essays) within MeuTutor ENEM was being ineffectual by lack of student motivation, considering that, because it is a competitive environment, students feared to contribute to the process. Thus, we included the proposed model in this work within the MeuTutor environment with the main objective of supporting essays among students.

When applying the model proposed in the context of MeuTutor, we have to evaluate two fundamental aspects. The first aspect is the own effectiveness of using Peer Assessment. Thus, questions like "How can we evaluate the quality of the corrections made by the students?" and especially "These assessments can be compared with an assessment of an expert (teacher)?" must be answered. We also inquired about the "Does the model applied in this context really reducing the cost and overwork of the teacher? ". In addition, the second fundamental point that needs to be evaluated is about gamification. So, questions like "What is the influence of gamification in the proposed model?" and "The gamification really influence and motivates students within the peer assessment process in this competitive environment?" must be answered.

Thus, two experiments were performed at separate. The first experiment evaluates the effectiveness of the peer assessment model, answering the questions of the first fundamental aspect and it is presented in Subsection 4.1. The second experiment evaluates the influence of gamification into peer assessment model. It answers the questions of second fundamental aspect and it is presented in Subsection 4.2.

4.1. Evaluating the use of peer assessment

This experiment aims to evaluate the proposed model with regards to the cost, reducing overwork for teacher and effectiveness if compared with the traditional way (without using any peer assessment model, where the activities are corrected by specialist teachers). In this sense, the first research question (QP1) aims to answer if the proposed model really is equivalent to the traditional method within MeuTutor-ENEM environment.

QP1 - Does the use of the peer assessment model proposed have the same efficiency, i.e., have similar results if compared to the traditional method?

Which brings us to the following hypotheses:

H1-0: The use of the proposed peer assessment model is equivalent to the traditional method.

H1-1: The use of the proposed peer assessment model is not equivalent to the traditional method.

The first question attempts to answer if the grades quality given by our model can be compared with the grades given by experts (teachers). If the null hypothesis is not refuted - indicating that there is an equivalence in the results - we need to identify what/which features the use of a proposed model has significant

Table 2
Formal definition of research hypotheses.

Hypotheses	Null hypotheses	Alternative hypothesis
H1	H1-0: N(M1,E1) = N(M2,E1)	H1-1: N(M1,E1) ≠ N(M2,E1)
H2	H2-0: T(M1,E1) = T(M2,E1)	H2-1: T(M1,E1) ≠ T(M2,E1)
H3	H3-0: C(M1,E1) = C(M2,E1)	H3-1: C(M1,E1) ≠ C(M2,E1)

Table 3
Factors levels.

Factor	Level	Description
Model	M1	Traditional Model (Experts)
	M2	Proposed Peer Assessment Model
Enviroment	E1	MeuTutor Enviroment Applied
	E2	Without Educational Environment

Table 4
Definition of treatments.

Number of treatment	Name	Used model	Environment
1	T1	M1	E1
2	T2	M2	E1
3	T3	M1	E2

differences advantageous with respect to the traditional method. In this sense, it is necessary to evaluate the cost needed to keep the model on the environment and the time spent by teachers and students in the process. So, some secondary research questions were defined with respect to these metrics.

QP2 - Does the use of the proposed peer assessment model present differences in time metrics in relation to the traditional method?

Which brings us to the following hypotheses:

H2-0: The use of the proposed peer assessment model brings no time differences over the traditional method.

H2-1: The use of the proposed peer assessment model brings time differences over the traditional method.

QP3 - Does the use of the proposed peer assessment model present differences in cost metrics in relation to the traditional method?

Which brings us to the following hypotheses:

H3-0: The use of the proposed peer assessment model brings no cost differences over the traditional method.

H3-1: The use of the proposed peer assessment model brings cost differences over the traditional method.

Formally, the hypotheses described above can be defined as shown in Table 2. The functions N, T and C, shown in the table, return respectively the value of the final grade, the time spent and the cost involved. They are regarding the use of the traditional model (M1) or use of the proposed model (M2), when applied in the educational environment MeuTutor (E1) or when applied without educational environment (E2). (N, T and C are the metrics and M and E are factors of the our experiment), according Table 3.

In our scenario, the experimental unit is very specific. The experimental research unit is a certain essay and an evaluation form, pre-defined by teacher. This is a sample that is used in which the "treatments" of the experiment (set of factor) is applied to obtain the response variables (dependent) mentioned above. It is through this unit that will be possible to obtain statistical variation in the analysis of research results. The unit consists of essay on the subject of artificial intelligence. This theme was chosen randomly and were asked to create an essay by an expert.

As we have factors with only two levels each, we can use a factorial experiment 2^k without repetition, removing a combination that is impossible to be performed (that is the use of the model M2 in none educational environment E2). Each execution of the model has relatively high cost and require lots of time to prepare the environment, students and other efforts required for its completion. For this reason, a factorial design was chosen without repetition. In this sense we have only three possible tests, running each only once, with three possible treatments. Table 4 describes each of the treatments.

After configuring MeuTutor environment, the next step was the selection of students who would participate in the experiment. For this, we selected 30 users of MeuTutor who would like to participate of this experiment. Users were recently admitted on higher level courses - 20 students in math course and 10 students in the accounting course at Federal University of Alagoas. Then, we invited two experts in discursive questions, indicated here by specialist 1 and 2. Both experts have specific training in higher education.

The group accounting students performed the treatment 3, that is, made manually writing (without the use of environment E2) and the correction of their essays were made by expert 1. On the other hand, math students made their essays on MeuTutor environment

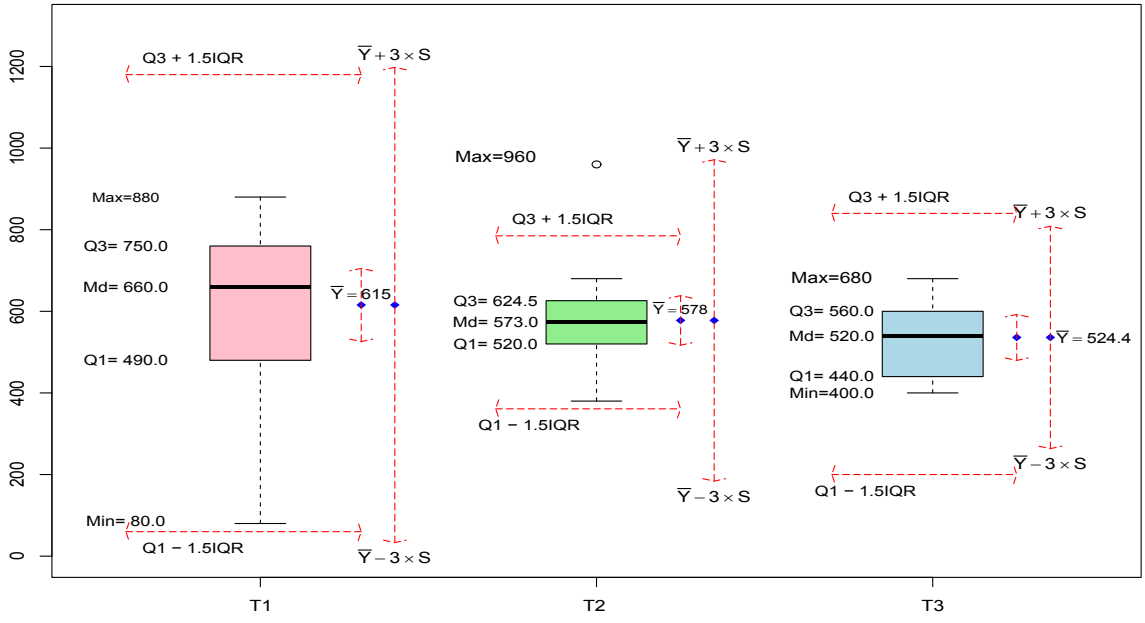

Fig. 5. Box diagram with comparative metric final grade for the treatments T1, T2 and T3.

Table 5
Summarization of the data of variable final grade (N).

Treatment	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
T1	80.0	490.0	660.0	615.6	750.0	880.0
T2	380.0	520.0	573.0	577.7	624.5	960.0
T3	400.0	460.0	540.0	536.0	590.0	680.0
T1−T2	−400.00	−79.75	40.00	37.89	110.00	480.00

Table 6
Results of Shapiro-Wilk test to data from metric final result (N).

Treatment	$W_{calculated}$	W_{α}	P_{value}	α
T1	0.92223	0.897	0.1416	0.05
T2	0.90344	0.897	0.06599	0.05
T3	0.96201	0.842	0.8086	0.05

Table 7
Results of *t*-test test to data from metric final result (N).

Treatment	t	df	p-value
T1 × T2	0.85383	17	0.4051
T1 × T3	1.4728	25.546	0.153
T2 × T3	0.98704	24.523	0.3333

(E1) through the available link. The correction of their essays were made both using the model in this work (treatment 2) and the traditional model (treatment 1), which was carried out by specialist 2.

In the case of treatment 3, students made their essays and delivered to the mediator the answer sheet who delivered to the expert 1 for evaluation. The specialist 1 reviews the essays and submitting their answers into an online form available through the GoogleForms, since he has no access to environment int this treatment.

4.1.1. Data analysis

As mentioned earlier, the variables and defined levels in experiment planning were: evaluation final grade (N) (from 0 to 1000); Time for correction (T) (in minutes); Cost for correction (C) (em R$). The first variable analyzed is the final grade (N). Fig. 5 presents the box diagram for the metric of the final grade N with respect to the treatments T1, T2 and T3.

It can see from the figure that the mean (AVG(T1) = 615, AVG(T2) = 578 and AVG(T3) = 524.4) and median (median(T1) = 660, median(T2) = 573 and median(T3) = 520) are similar, considering the scale. This figure suggests that the grades obtained with the application of these treatments have similar statistics variations, which indicates a certain similarity in the treatments. However, statistical evidence not has been generated yet.

As one of our research questions is to evaluate the grades obtained by the proposed model compared to the grades obtained by the traditional model, then we can subtract the grades by T1 and T2. Thus, the difference of the grades by the treatment T1 and T2 in the mean (AVG(T1−T2) = 37.8) and median (median(T1−T2) = 40) is close to 0 (zero), indicating which are well equivalents. In general, the difference between T1 and T2 grades varied from −79.75 to 110. If we consider the rating scale between 0 and 1000, we have differences between the grades range from 7.9% to 11%.

In summary, the data for the metric final Grade (N) obtained with the execution of treatments (T1, T2 and T3) are showed in Table 5, for better viewing.

In this sense, carrying out further analysis in order to verify the validity of its hypotheses of research is needed. We can emphasize that for each hypothesis, we must indicate which treatment was better. To perform such verification, we will take several statistical tests on each defined evaluation metric.

The first task to conduct the verification of a specific research hypothesis is to analyze if the data involved are a normal distribution. The normality of the data is important because it determines which statistical test must be used in the analysis. There are some statistical tests to verify the normality of the data, but the most recommended by statisticians is the Shapiro-Wilk test (Shapiro & Francia, 1972). For this reason, it will be used in calculations.

The first hypothesis check will be made with respect to final grade metric. In this case we applied the Shapiro-Wilk normality test. The results are in Table 6.

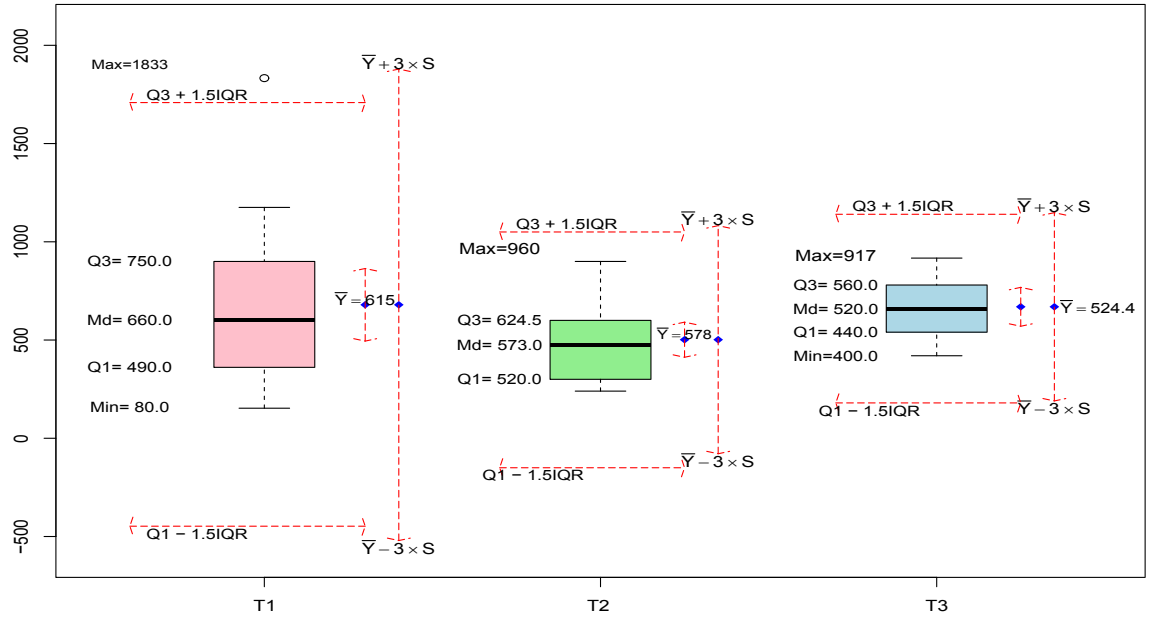

Fig. 6. Box diagram comparing the metric Time for the treatments T1, T2 and T3.

Table 8
Summarization of the data of variable time (T).

Treatment	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
T1	153.0	382.8	600.0	679.3	889.5	1833.0
T2	240.0	337.5	472.5	501.7	588.8	900.0
T3	420.0	555.0	660.0	669.1	780.0	917.0

Table 9
Results of Shapiro-Wilk test to data from metric Time (T).

Treatment	$W_{calculated}$	W_{α}	P_{value}	α
T1	0.89637	0.897	0.04969	0.05
T2	0.91724	0.897	0.1155	0.05
T3	0.97249	0.842	0.9129	0.05

Table 10
Results of statistical tests to the data of metric Time (T).

Treatment	Used test	v/w	t	df	p-value
T1 × T2	Wilcoxon signed rank test	122	–	–	0.1169
T1 × T3	Wilcoxon rank sum test	81.5	–	–	0.7007
T2 × T3	Welch Two Sample *t*-test	–	−2.4605	21.938	0.02222

In this case, we see that the data are normal distributions, since $W_{calculated} > W_{\alpha}$ and $P_{value} > \alpha$ in all cases. Thus, as the data distribution from all treatments are normal distributions, we apply the T-Test (Welch, 1938), comparing the treatments. The T-test results are presented in Table 7.

We can see that the p-value(T1 × T2) = 0.4051, p-value(T1 × T3) = 0.153 and p-value(T2 × T3) = 0.3333, obtained with the execution of the T-test, are all greater than $\alpha = 0.05$. So with 95% confidence statistically it can not be said the values of the notes obtained between treatments have significant differences, i.e., there is no statistical evidence that show the non-equivalence of the final grades.

The fact that these coefficients are larger than the α indicates the results of the grades are equivalent (equal), not generating enough statistical evidence to refute the null hypothesis H1-0, implying that **the use of the evaluation proposed model is equivalent to the traditional method**.

The next metric to be analyzed is time to correction (T). Fig. 6 presents a set of box plots for the metric of time (T) with respect to the treatments T1, T2 and T3.

When analyzing the average and median of the data of the treatments where there is a specialist teacher involved, (AVG(T1) = 679.3/AVG(T3) = 669.1 and median(T1) = 600/median(T3) = 660), we can verify the data is very similar. Comparing these data with the data of the treatment T2 (AVG(T2) = 501.6/median(T2) = 472.5), we note the time in T2 is significantly shorter than the time in the other two treatments. However, as the diagrams intersect, it is necessary to check the confidence interval of the data and execute the hypothesis tests to get a valid statistical conclusion.

Similarly, the data for measuring time (T) obtained with the execution of treatments (T1, T2 and T3) are summarized in Table 8.

Like was done in metric note, it is necessary to perform the statistical tests to validate or invalidate the research hypotheses involving the correction time. So, the next hypothesis verification is made with respect to the metric time (T). In this case, we applied the Shapiro-Wilk test with the data from these treatments to analyze the normality of the data and the results can be seen in Table 9.

Under these conditions, only data T1 is not normal, since $W_{calculated} < W_{\alpha}$ and $P_{value} < \alpha$ in this case. Thus, as the data of the treatment T1 has a non-normal distribution, so when comparing data involved T1, we used a non-parametric test - the Wilcoxon test (Wilcoxon, 1945). In this case, only the T2 × T3 combination involved the T-test, since both are normal distributions. The data of execution these tests are shown in Table 10.

We observe that as p-value (T1 × T2) = 0.1169 > α = 0.05, then it is not possible to conclude that there are significant differences between these treatments. However, when we analyze p-value (T2 × T3) = 0.02222 < α = 0.05, we can conclude that there are statistically significant differences and we can conclude with 95% confidence that the time in T2 is less than the time in T3. If we analyze the combination of T1 and T3 treatment (which both use experts) we have a p-value (T1 × T3) = 0.7007 >α=0.05 and

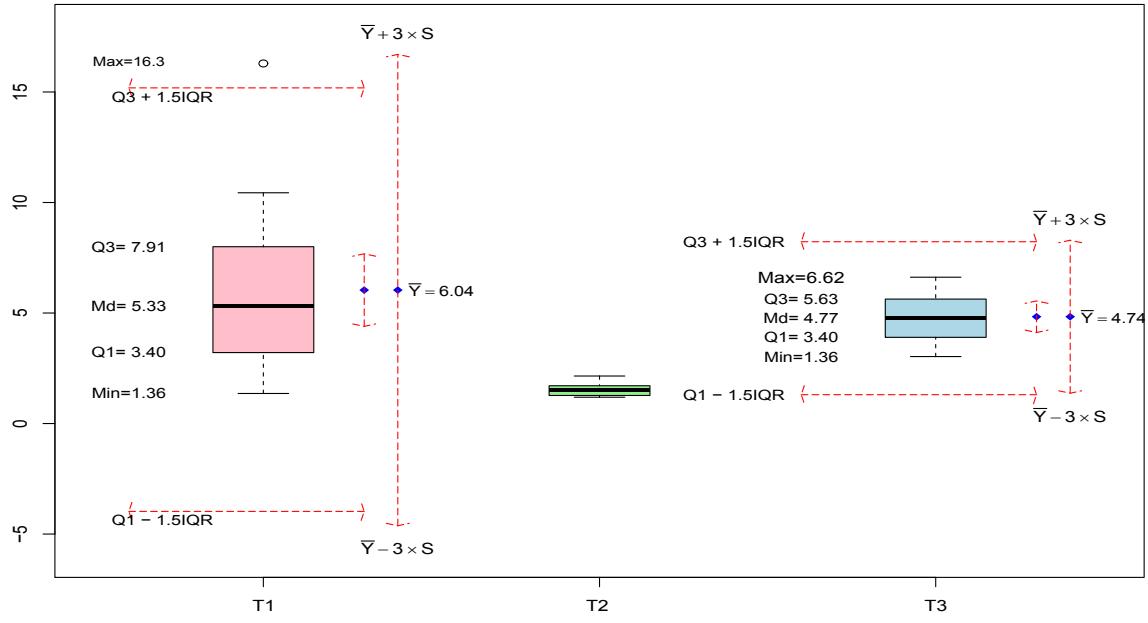

Fig. 7. Box diagram with comparative metric cost between treatments T1, T2 and T3.

Table 11
Summarization of the data of variable final cost (C).

Treatment	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
T1	1.36	3.40	5.33	6.03	7.90	16.29
T2	1.19	1.27	1.52	1.56	1.71	2.15
T3	3.03	4.01	4.77	4.83	5.63	6.62

Table 12
Results of Shapiro-Wilk test to data from metric Cost (C).

Treatment	$W_{calculated}$	W_{α}	P_{value}	α
T1	0.89635	0.897	0.04965	0.05
T2	0.25269	0.897	1.057e-08	0.05
T3	0.97276	0.842	0.9152	0.05

therefore there is no statistical difference with the desired confidence.

Since our evaluation is based on the behaviour of treatment T2 data (where our model is applied), we have statistically that the time involved in it is lower than one of treatment T3 (only specialist) with 95% confidence level, but we do not have enough statistical evidence to show that the time in T2 is different from time in T1 (expert in the environment). Anyway, there is enough statistical evidence that the use of the proposed model brings differences in time compared to the traditional method. Thus, we refute the null hypothesis H2-0 (there would be no difference) and accept the alternative hypothesis H2-1 with 95% confidence, implying that **the use of the evaluation proposed model brings significant differences in time compared to the traditional method**.

Finally, the last metric to be analyzed is the cost of correction (C). Fig. 7 shows the box diagram for the metric of cost involved (C) with respect to the executed treatments.

The mean and median in T1 and T3 (AVG(T1) = R$6.03/ AVG(T3) = R$ 4.83) are similar, because both treatments used experts in corrections. On the other hand, the cost of treatment T2 appears fixed, and as observed in Fig. 7, it is significantly below than the other two treatments, which can reveal the cost by using the proposed model in environment (T2) is lesser than the cost of correction by experts using the environment or not (T1 and T3).

The data for the cost metric (C) obtained with the execution of treatments (T1, T2 and T3) are summarized in Table 11. It is noteworthy that the data have been rounded to two decimal places.

Finally, the last hypothesis test analyze the cost metric (C). The results of the application of the Shapiro-Wilk test with the data from the treatments T1, T2 and T3 are shown in Table 12.

Due to the fact that the data T1 and T2 are not normal, since $W_{calculated} < W_{\alpha}$ and $P_{value} < \alpha$ in both cases, then we must use the Wilcoxon test. The results of applying the Wilcoxon test with the data from these treatments are shown in Table 13.

We observe that the p-value (T1 × T2) = 0.0002522 < *alpha* = 0.05 and p-value (T2 × T3) = 1.111e-06 < *alpha* = 0.05, implying that there are significant statistical differences between the cost of treatments T2 compared to T1 and T2 compared to T3. Therefore, we can conclude the cost of treatment T2 is lesser than the cost of such treatment with a confidence level of 95%. On the other hand, the p-value (T1 × T3) = 0.5486 is greater than $\alpha = 0.05$, which means there is no significant difference between times T1 and T3.

Thus, we have enough statistical evidence today the cost is different and we refute the null hypothesis H3-0 (where there would be no difference) and accept the alternative hypothesis H3-1, with a confidence level of 95%, implying that **the use of the evaluation proposed model brings significant differences in cost compared to the traditional method**.

With the above conclusion we know the cost of applying the model is lower than using traditional methods, however, there is a need to specify how this cost is lower. For this we created a regression model (Espinheira, da Silva, & Silva, 2015) (Ferrari & Cribari-Neto, 2004). Consider the following linear model:

$$y = X\beta + \varepsilon = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \underbrace{\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}}_{(x_1 \quad x_2 \quad \dots \quad x_k)} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}. \tag{1}$$

We see that y is a vector of n observations of the random variable dependent, or, $y_1,\dots,y_n$ is a sample of the population of interest. X is a matrix $n \times k$ composed of the covariates. Note that each column X is a set of n observations of the covariate x_t, $t = 1,\dots,k$, thus, we have k covariates. X is not a random variable, it is observated and fixed. We still have β is a vector of k unknown fixed params (not random variables). Finally, ε is a vector of n random errors with zero average $(E)(\varepsilon_i) = 0$ and constant variance over the observations, that is, $\text{var}(\varepsilon_i) = \sigma^2$ for each $i = 1,\dots,n$.

One of the major assumptions of linear regression models is that:

$$E(\varepsilon) = \mu_{\varepsilon} = 0 \tag{2}$$

Thus,

$$E(y) = E(X\beta) + E(\varepsilon) \Leftrightarrow E(y) = X\beta \Leftrightarrow \mu = X\beta. \tag{3}$$

So, **in practice** our final model is

$$\mu = X\beta \tag{4}$$

$$\text{var}(y) = \underbrace{\text{var}(X\beta)}_{0} + \underbrace{\text{var}(\varepsilon)}_{\sigma^2} \Leftrightarrow \text{var}(y) = \sigma^2 \tag{5}$$

We can represent our model considering the ith observation as

$$\mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \dots + \beta_k x_{ik}, \quad i = 1, \dots, n.$$

To really know the above model we need to estimate $\beta_1,\beta_2,\dots,\beta_k$. Typically we do this by using the method of maximum likelihood. Thus, we get $\widehat{m}u_i$ when obtained $\widehat{\beta}_1, \widehat{\beta}_2, \dots, \widehat{\beta}_k$, such that

$$\widehat{\mu}_i = \widehat{\beta}_1 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3} + \dots + \widehat{\beta}_k x_{ik}, \quad i = 1, \dots, n.$$

Now we can obtain estimates for μ_i.

If our dependent variable belongs to positive real, we can use

Table 13
Results of Wilcoxon test to data from metric cost (C).

Treatment	Used test	v/w	p-value
T1 × T2	Wilcoxon signed rank test with continuity correction	170	0.0002522
T1 × T3	Wilcoxon rank sum test with continuity correction	103	0.5486
T2 × T3	Wilcoxon rank sum test with continuity correction	0	1.111e-06

the gamma distribution, so a suitable regression model would be:

$$\log(\mu_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \ldots + \beta_k x_{ik}, i = 1, \ldots, n.$$

This is because the log can only be calculated for positive real values and its outcome has values in all real, which releases the $\widehat{\beta}$ s. Note that $\log(u) \in (-\infty,+\infty)$ for each $u \in (0,+\infty)IR^{+}$.

In the case of cost in the experiment, we tested the model

$$\log(\mu_i) = \beta_1 + \beta_2 CostIndicatorT2$$

in which μ_i is the average cost of considering all three groups and CostIndicatorT2 is a variable that takes the value one when the model is used and zero for others. The calculation was done with the following commands in R:

```
> fit <- glm(t(Custo)~t(Ind_T2), family=Gamma(link=log))
#Function for generalized linear model (glm) using gama distribution
> summary(fit)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.7240     0.0809   21.31  < 2e-16 ***
t(Ind_T2)    -1.2751     0.1293   -9.86 1.03e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Gamma family taken to be 0.1832411)
Null deviance: 22.5610  on 45  degrees of freedom
Residual deviance:  7.1323  on 44  degrees of freedom
AIC: 157.56
```

Then, based on the above results, we have the indicator covariate of the model is highly significant, with a level about 0.00001. In other words, the proposed method has a different value for cost from the other two groups, which is a lower cost, which is confirmed by the negative sign of the estimate $\widehat{\beta}_2 = -1.2751$. Also, we have additional information based on the regression model. We have to $\{1 - exp(\widehat{\beta}_2)\} \times 100$ represents **the percentage reduction on implanting the correction system by student peer, with value equal a** 72.4%.

4.2. Evaluating the gamification of the model

This section aims to present the project and execution of the second experiment which aims to evaluate the motivational impact of the inclusion of gamification in the proposed model in this work. Taking this into consideration, we set our research question, which aims to respond if the model of peer assessment with gamification really differentiates the model without gamification. Thus, our research question is:

QP - The use of gamification elements in the peer assessment model brings significant differences of the motivational of the students involved?

To answer this question requires some knowledge of how to assess students' motivation. Given that motivation is something that can not be measured easily, our assessment will be based on expected behaviours by students within the system, evaluating whether gamification elements encouraged students to participate more (register, access the platform, perform and correct essays, for example).

In this sense the main question can be divided into secondary questions. Considering the amount of access of students in the system, we can formulate the following research hypothesis:

QP1 - Does the use of gamification elements in the peer assessment model present significant differences in the number of accesses (logins) of the students?

Which brings us to the following hypotheses:

H1-0: The use of gamification elements in the peer assessment model brings no differences in the number of accesses (logins) of the students on the platform.

H1-1: The use of gamification elements in the peer assessment model brings differences in the number of accesses (logins) of the students on the platform.

Considering the realization of essays by students in the system, we can formulate the following research hypothesis:

QP2 - The use of gamification elements in the peer assessment model present significant differences in the amount of essays made by students?

Which brings us to the following hypotheses:

H2-0: The use of gamification elements in the peer assessment model brings no differences in the amount of essays made by students

H2-1: The use of gamification elements in the peer assessment model brings differences in the amount of essays made by students

Finally, considering the corrections of essays made by students in the system, we can formulate the following research hypothesis:

QP3 - The use of gamification elements in the peer assessment model present significant differences in the amount of essays corrected by students?

Which brings us to the following hypotheses:

H3-0: The use of gamification elements in the peer assessment model brings no differences in the amount of essays corrected by students

H3-1: The use of gamification elements in the peer assessment model brings differences in the amount of essays corrected by students

Formally, the hypotheses described above can be defined according Table 14. The functions A, RF and RC presented in the table, return respectively, the amount of access, number of essays made

Table 14
Formal definition of research hypotheses of gamification experiment.

Hypotheses	Null hypotheses	Alternative hypothesis
H1	H1-0: A(M1) = A(M2)	H1-1: A(M1) ≠ A(M2)
H2	H2-0: RF(M1) = RF(M2)	H2-1: RF(M1) ≠ RF(M2)
H3	H3-0: RC(M1) = RC(M2)	H3-1: RC(M1) ≠ RC(M2)

Table 15
Definition of treatments of the gamification experiment.

Number of treatment	Reference	Used model	Description
1	T1	M1	No gamification
2	T2	M2	With gamification

and amount of corrected essays, regarding the use of the model without gamification elements (M1) or the use of model with gamification elements (M2).

In the case of this experiment specifically, as we have only one variant factor (model used) with only 2 levels, we can use the same previous experiment design (factor 2^k without repetition), even because each repetition of the experiment has a high cost relatively and requires a long time to preparing the environment.

Table 15 describes each of the two treatments.

The experiment was conducted in a private school with high school classes. After configuring of the MeuTutor environment, a training was performed on day 07.20.2015 at school where the environment was presented to directors, coordinators, teachers and all the students of high school classes (separately). Almost 100 students who were present were interested and made themselves available to participate. Randomly, it was distributed to each student a group (1 or 2) representing the control group (GC) and experimental group (GT), for each class.

Students in the control group (CG) performed the treatment T1 (without gamification) while students of the experimental group (GT) performed the treatment T2 (with gamification). It is noteworthy that both were using the peer assessment model proposed, differed only with the presence or not of gamification elements.

The access the environment varies according to the chosen group. All the instructions of access, configuration, and the processes involved were given to students. Based on the date of this training, the final date of the experiment was set on 03.08.2015 (15 days later). Students had the option to register on the platform and do the activities freely. All their interactions were stored in the database, obtained directly by the system. Analysis of these data will be presented in the following section.

4.2.1. Data analysis

This section aims to briefly present some of the main results obtained in the experiment, in a direct comparison between the application without gamification model (for the treatment T1) with the application of the model with gamification (in the case of treatment T2). Set of data observed are presented in Table 16.

We can see in Table 16 that there were two more registers in the group of the treatment T2 (+11.76%). The number of pendent essays (non corrected essay) was four, regardless of treatment. This may have been caused by the short deadline for realization and correction of essays, implying that some students failed to correct enough essays.

Regarding the average of the grades of essays, there was a significant difference, where the grades of treatment with gamification T2 were lower (−105.41 or −15.75%) than treatment without gamification T1. Is worth mentioning that the students in each group are different and consequently their essays will be different, which leads to the difference of the grades. Furthermore, the group with gamification had more essays performed, which can enhance this difference in absolute note. Anyway, to effectively compare the results of one more correct way, we asked a specialist to correct the essays of both groups in order to compare the grades. Thus, we had a parameter to compare the grade obtained by the model with and without gamification, based on rating assigned by the teacher to essays.

Table 16
Summarization of data on the variables analyzed.

Data observed	T1	T2	Difference	% Dif.
Number of student records	17	19	+2	+11.76%
Total Pending Essays	4	4	0	0%
Average grades	669.33	563.92	−105.41	−15.75%
Number of logins (A)	42	69	+27	+64.28%
Total Essays Made (RF)	19	21	+2	+10.53%
Total Corrected Essays (RC)	45	54	+9	+20.0%

The average grade given by the expert for the group 1 (without gamification) was 602.47, about 66.86 lesser than the average of the grades given by the students, which is equivalent to −9.98%. On the other hand, the average grade given by the specialist for treatment 2 (with gamification) was 613.28, about 49.36 higher than the average of the grades given by the students, which is equivalent to +8.75%. Thus, we can conclude that there is a proximity in percentage terms (ignoring the sign) in the grades of both treatments (with and without gamification) compared the grades given by the specialist (9.98 % × 8.75 %). So, in percentage terms, we can conclude that gamification did not affect the final results (final grade) of the model compared to the expert results, keeping the results presented in experiment 1. However, there is a curious fact. Although the notes are close, students with gamification have tended to give lower results than the grades of the expert while students without gamification have tended to give higher results than the grades of the expert.

On the other hand, by analyzing the metrics defined in the experiment, we can see the amount of access (logins) of treatment with gamification (T2) was 64.28% higher than without gamification, which indicates the strong influence of gamification in student access on the platform. The number of essays made was also higher in treatment T2, with two essays performed more than the treatment T1. Despite the small number, it represents about 10.53% of more essays. The same is true with respect to metric corrected essays. In treatment T2, we have 9 corrections more than in treatment T1, which represents an increase of about 20%.

Similarly to the previous experiment we statistically analyses the variables of this experiment, starting with the access number (A). Fig. 8 shows the box diagram for this metric with respect to treatment T1 (without gamification) and T2 (with gamification).

We can see from the figure that the mean (AVG(T1) = 2.471 versus AVG(T2) = 3.632) and median (median(T1) = 2 and median(T2) = 3), present certain differences in behaviour. Analyzing only those measures, it may be noted that there is a tendency that the amount of access (logins) of treatment T2 is greater than the number of logins of the treatment T1. The figure suggests that this metric data obtained with the application of these treatments have statistically significant variations, which indicates in a certain disparity in treatment.

We can also observe that the amount of *outliers* (Nonstandard points) is much higher in T1 than in treatment T2, which indicates to us that few students accessed more than twice (median value) in T1, while it was much more often in the treatment T2 (considering the presented dispersion and the median(T2) = 3 > median(T1) = 2).

However, statistical evidence has not yet been generated to assert these findings. Such statements can only be made when the statistical tests are performed and executed to check hypothesis regarding access metric (A). In this case, in the first moment we must apply the Shapiro-Wilk test for find normality whether or not data of each treatment. The results of applying the Shapiro-Wilk test when executed on the data of the metric access for treatments are presented in Table 17.

By analyzing this data, it can be noted that the values of

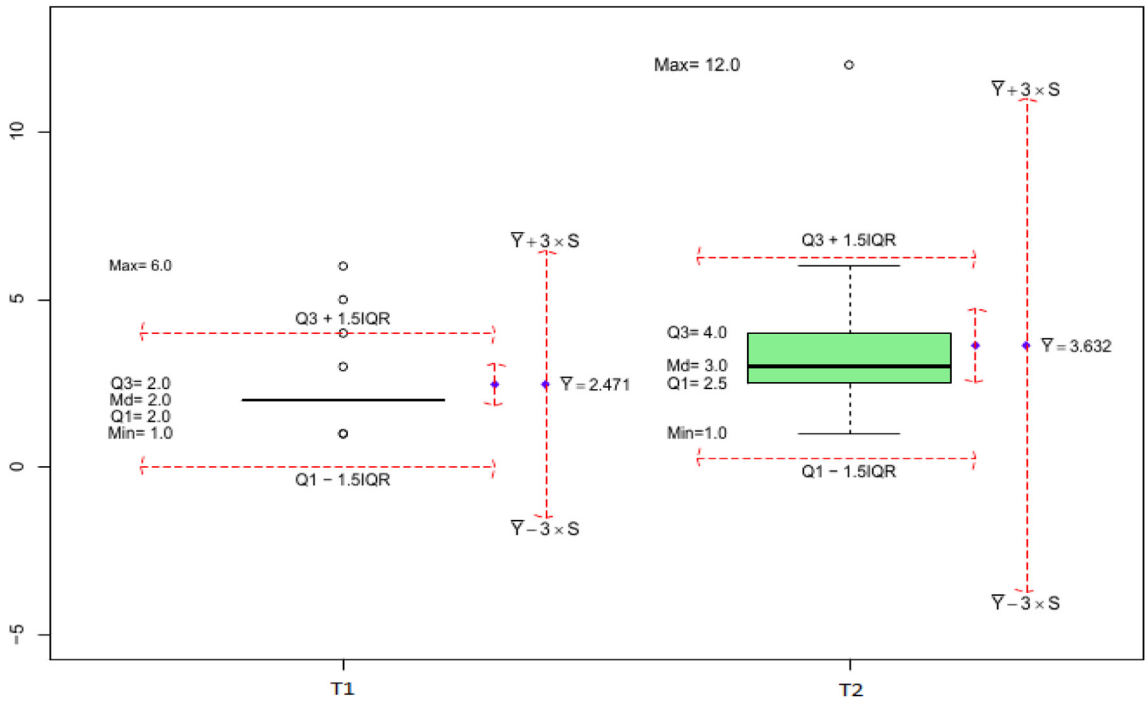

Fig. 8. Box diagram comparing the metric access for the treatments T1 and T2.

Table 17
Results of Shapiro-Wilk test to data from metric access (A).

Treatment	$W_{calculated}$	W_{α}	P_{value}	α
T1	0.71721	0.892	0.0001823	0.05
T2	0.74726	0.901	0.0002154	0.05

$W_{calculado}(T1) = 0.71721 < W_{\alpha}(T1) = 0.892$ and $P_{valor}(T1) = 0.0001823 < \alpha = 0.05$. So it is possible to refute the null hypothesis of Shapiro and therefore we can affirm with a significance level of 5% that the sample T1 does not come from a normal population.

Similarly, if we analyze the data from the T2 treatment, we have to $W_{calculado}(T2) = 0.74726 < W_{\alpha}(T2) = 0.901$ and $P_{valor}(T2) = 0.0002154 < \alpha = 0.05$. Just as in T1, the null hypothesis Shapiro can be refuted and T2 distribution data is also not normal.

Thus, since the distribution data of the two treatments are not normal distributions, we apply the Wilcoxon test comparing the treatments. The data of the execution of the Wilcoxon test are shown in Table 18.

We can see that the value of p-value(T1 × T2) = 0.03209, obtained by running the Wilcoxon test is lesser than $\alpha = 0.05$. Thus, with 95% confidence statistically can be refute the null hypothesis, affirming that the values obtained of the amount of access between treatments have significant differences, i.e., there is statistical evidence demonstrating the non-equivalence of data.

The fact that this coefficient be lesser than α generates enough statistical evidence to refute the null hypothesis H1-0 (where access would be equal) and accept the alternative hypothesis H1-1, implying that **the use of gamification techniques in peer assessment model brings differences in the number of accesses of the students on the platform.**

4.3. Discussion

Two experiments in MeuTutor educational environment showed satisfactory results, given that it has been proven statistically that the grades of the model are equivalent to the traditional model (correction by experts), having advantages with respect to time (lower time) and a cost about 72% less, according to the regression model created. Moreover, it was shown in the experiment that gamification positively influenced on the overall context of the peer assessment model. The results indicate the gamification encouraged students to use the platform. In our analysis, the increased amount of access was 64.28%. Moreover, our results indicate that there were approximately 10.53% more essays performed and about 20% more essays corrected.

To analyze these results we will simulate a real scenario of application. Suppose that a single school with about 100 students resolves promote tests for creation of essays in their on-line environment. Suppose that will be made one essay per week per student, which leads to four essays monthly per student. Suppose an annual period of 10 months of work in this project, we will have at the end a total of 4000 essays made (100 × 4 × 10).

Considering the number of essays and the average time generated in our experiments, we have an expert would take about 754 h (4000 × 11.31 min) to correct all essays in total. On the other hand, the simple use of the model even without gamification decrease this time to an average of 8.36 min for redaction, which would give us a total of 557 (4000 × 8.36 min) hours spent on corrections (a

Table 18
Results of Wilcoxon test to data from metric access (A).

Treatment	Used test	v/w	p-value
T1 × T2	Wilcoxon signed rank test with continuity correction	96	0.03209

gain of 197 h, or 26.12%). The inclusion of gamification in the model reduces the correction time for 6.68 min. So, with the application of the complete model the time total to correct the essays would be 445 h (4000 × 6.68 min), a reduction of 309 h or 40.98% compared to the traditional model or reduction of 112 h or 20.10% compared to the model without gamification.

Considering the same number of essays and analyzing the average of the costs involved in the experiments, we have that on average, an essay correction by a specialist would cost about R$6.04. Thus the total to correct all essays would be about R$ 24.160 (4000 × 6.04). On the other hand, the simple use of the model even without gamification decrease this cost to an average of R$1.68 for redaction, which would give us a total of R$ 6.720 (4000 × 1.68) (about R$ 17.440 lesser or 72.18%). So, with the application of the complete model the cost total to correct the essays would be R$ 5.360 (4000 × 1.34), a reduction of R$ 18.800 or 77,81% compared to the traditional model or reduction of R$ 1.360 or 20.23% compared to the model without gamification.

Is worth emphasizing that the model proposed in this paper can be deployed/integrated into other online learning environments. For this, it is first necessary analyze the objectives of the application of the model and the feasibility of adaptation/modification of the environment to use the proposed model. It is inevitable that minor adjustments are required in the implementation of the model, even this has been deployed in flexible and adaptable way. It is necessary, however, an effective implementation in other environments. With the implementation/integration into other environments, the difficulties that perhaps may appear serve as input for possible improvements in the implementation of the model, leaving the most complete, flexible and compatible with most educational environments.

4.4. Threats to validity

This section describes concerns that must be improved in future replications of this study and other aspects that must be taken into account in order to generalize the results of this study.

First of all, the implementation of the model, even though it was designed and constructed in an abstract and independent environment, also suffered a strong influence of technological MeuTutor educational environment. Thus, it is possible that emerge adjustment needs in the application with other educational environments. This way, conducting an environmental study is needed before applying the peer assessment model proposed in this work.

The experiments design and executed in this work to evaluate the peer assessment model was incorporated into Meututor environment. Although it was planned impartially, the fact that the model was proposed under the Meututor needs, and the technology of its implementation is compatible with the Meututor, may have influenced positively in the implementation of model of and its results. In addition, all students wanted to participate in the experiment.

5. Conclusions and future work

The work presented a gamified peer assessment model that aims to provide a solution to the inclusion of written evaluations in competitive on-line educational environments. The need for the creation of this model arose from the large number of students who were present in these environments and the great difficulty in providing discursive activities on them. On the other hand, without Peer Assessment, it generates a high cost and a large overwork on teachers involved. On the other hand, with Peer Assessment, the problem was the lack of motivation of students, especially by the competitive aspect of the environment.

The proposed model has reached the proposed goal, allowing the inclusion of discursive evaluations in on-line environments in a viable way, as can be seen in its integration with MeuTutor educational environment and its use in practice with about 130 students. The experiments also gave statistical evidences that the results with the proposed model are similar to those of the traditional model. At first time, we evaluated the results by our model with the results by the traditional model (correction by experts) under the metrics final grade, correction time and associated cost.

The results were good, considering that it was statistically demonstrated that the grades obtained for both models are equivalent. Similarly, the results of metric time indicated a possibility of correction time be lower using our model, but in some cases it is similar. Finally, the results for metric cost were very favourable when applying our model, considering that it is not necessary corrections by experts. Through regression model has been proven that the cost is 72% lower.

In the second experiment, we evaluate the impact of gamification elements as a motivational aspect to student. There were significant statistical variations in the amount of access (logins) of students between the model with gamification and the model without gamification. This indicates the gamification encouraged students to use the platform. In our analysis, the increased amount of logins was 64.28%. Moreover, our results indicate there were approximately 10.53% more essays made in the treatment T2 and there were about 20% more essays corrected in this treatment.

However, there are some limitations and possibilities for future work. The implementation of the model, even though it was designed and constructed in an abstract and independent environment, also suffered a strong influence of technological MeuTutor educational environment. In this case, we plan to integrate the model into other environments.

The model experiment was incorporated into Meututor environment. Although it was planned impartially, the fact that the model was proposed under the Meututor needs, and the technology of its implementation is compatible with the Meututor, may have influenced positively in the implementation of model of and its results. Thus, we have as a future work planning perform new experiments in different conditions, both with respect to the environment used, as well as users involved and the form of evaluation of the model.

Moreover, we must evaluate new metrics comparison of the model with the traditional model that may be relevant, such as the level of student learning, improvements in its judges skills, among other benefits who a peer assessment model can bring us. Thus, we intend to make further experiments in order to evaluate such metric.

Acknowledgements

This work has been supported by the Brazilian institutions: Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) (grant number: 20130603-002-0040). We also thank MeuTutor company.

References

Andrade, F. R., Mizoguchi, R., & Isotani, S. (2016). The bright and dark sides of gamification. In *International conference on intelligent tutoring systems* (pp. 176–186). Springer.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education, 24*(4), 413–426.

Cavalcanti, S. R. (2008). *Veer: Um algoritmo de seleção de pares em redes ad hoc veiculares* (PhD thesis). Universidade Federal do Rio de Janeiro.

Chang, C.-C., Tseng, K.-H., & Lou, S.-J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment

and peer-assessment in a web-based portfolio assessment environment for high school students. *Computers & Education, 58*(1), 303–320.
Chen, C.-h. (2010). The implementation and evaluation of a mobile self-and peer-assessment system. *Computers & Education, 55*(1), 229–236.
Cunha, E., & Figueira, Á. (2009). A web-based tool for assessing online peer-reviews. In *8th European conference on e-learning, University of Bari, Italy, 29-30 October 2009* (p. 132). Academic Conferences Limited.
Dominguez, C., Cruz, G., Maia, A., Pedrosa, D., & Grams, G. (2012). Online peer assessment: An exploratory case study in a higher education civil engineering course. In *Interactive collaborative learning (ICL), 2012 15th international conference on* (pp. 1–8). IEEE.
Espinheira, P. L., da Silva, L. C. M., & Silva, A. d. O. (2015). *Prediction measures in beta regression models*. arXiv preprint arXiv:1501.04830.
Falchikov, N. (1995). Peer feedback marking: Developing peer assessment. *Programmed Learning, 32*(2), 175–187.
Fermelis, J., Tucker, R., & Palmer, S. (2007). Online self and peer assessment in large, multi-campus, multi-cohort contexts. In *Ascilite 2007: ICT: Providing choices for learners and learning* (pp. 271–281). Ascilite.
Ferrari, S., & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics, 31*(7), 799–815.
Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education, 20*(3), 289–300.
Gouli, E., Gogoulou, A., & Grigoriadou, M. (2008). Supporting self-, peer-, and collaborative-assessment in e-learning: The case of the peer and collaborative assessment environment (pecasse). *Journal of Interactive Learning Research, 19*(4), 615–647.
Hamari, J. (2013). Transforming homo economicus into homo ludens: A field experiment on gamification in a utilitarian peer-to-peer trading service. *Electronic commerce research and applications, 12*(4), 236–245.
Hwang, G.-J., Hung, C.-M., & Chen, N.-S. (2014). Improving learning achievements, motivations and problem-solving skills through a peer assessment-based game development approach. *Educational Technology Research and Development, 62*(2), 129–145.
Jenkins, M. (2004). Unfulfilled promise: Formative assessment using computer-aided assessment. *Learning and Teaching in Higher Education, 1*(1), 67–80.
Kahiigi Kigozi, E., Vesisenaho, M., Hansson, H., Danielson, M., & Tusubira, F. (2012). Modelling a peer assignment review process for collaborative e-learning. *Journal of Interactive Online Learning, 11*(2), 67–79.
Kapp, K. M. (2012). *The gamification of learning and instruction: Game-based methods and strategies for training and education*. John Wiley & Sons.
Kawai, G. (2006). Collaborative peer-based language learning in unsupervised asynchronous online environments. In *Creating, connecting and collaborating through computing, 2006. C5'06. The fourth international conference on* (pp. 35–41). IEEE.
Malehorn, H. (1994). Ten measures better than grading. *The Clearing House, 67*(6), 323–324.
Miao, Y., & Koper, R. (2007). An efficient and flexible technical approach to develop and deliver online peer assessment. In *Proceedings of the 8th iternational conference on Computer supported collaborative learning* (pp. 506–515). International Society of the Learning Sciences.
Min, H.-T. (2006). The effects of trained peer review on efl students revision types and writing quality. *Journal of Second Language Writing, 15*(2), 118–141.
Miyake, N. (2007). Computer supported collaborative learning. In *The SAGE handbook of e-learning research* (pp. 248–265).
Moccozet, L., Tardy, C., Opprecht, W., & Leonard, M. (2013). Gamification-based assessment of group work. In *Interactive collaborative learning (ICL), 2013 international conference on* (pp. 171–179). IEEE.
Moreira, B. G. (2014). Desenvolvimento de uma ferramenta de avaliação por pares para disciplinas de algoritmos e programação. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação* (Vol. 3).
Murakami, C., Valvona, C., & Broudy, D. (2012). Turning apathy into activeness in oral communication classes: Regular self-and peer-assessment in a tblt programme. *System, 40*(3), 407–420.
Orsmond, P., Merry, S., & Callaghan, A. (2004). Implementation of a formative assessment model incorporating peer and self-assessment. *Innovations in Education and Teaching International, 41*(3), 273–290.
Pedro, L. Z., Lopes, A. M., Prates, B. G., Vassileva, J., & Isotani, S. (2015). Does gamification work for boys and girls?: An exploratory study with a virtual learning environment. In *Proceedings of the 30th annual ACM symposium on applied computing* (pp. 214–219). ACM.
Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). *Tuned models of peer assessment in moocs*. arXiv preprint arXiv:1307.2579.
Prins, F. J., Sluijsmans, D. M., Kirschner, P. A., & Strijbos, J.-W. (2005). Formative peer assessment in a cscl environment: A case study. *Assessment & Evaluation in Higher Education, 30*(4), 417–444.
Rubin, R. F., & Turner, T. (2012). Student performance on and attitudes toward peer assessments on advanced pharmacy practice experience assignments. *Currents in Pharmacy Teaching and Learning, 4*(2), 113–121.
Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational assessment, 11*(1), 1–31.
de Santana, S. J., Souza, H. A., Florentin, V. A., Paiva, R., Bittencourt, I. I., & Isotani, S. (2016). A quantitative analysis of the most relevant gamification elements in an online learning environment. In *Proceedings of the 25th international conference companion on world wide web* (pp. 911–916). International World Wide Web Conferences Steering Committee.
Schuwirth, L. (2004). Optimising new modes of assessment: In search of qualities and standards. *Tijdschrift voor Medisch Onderwijs, 5*(23), 250–251.
Shapiro, S. S., & Francia, R. (1972). An approximate analysis of variance test for normality. *Journal of the American Statistical Association, 67*(337), 215–216.
Sleeman, D., & Brown, J. S. (1982). *Intelligent tutoring systems*. London: Academic Press.
de Sousa Borges, S., Durelli, V. H., Reis, H. M., & Isotani, S. (2014). A systematic mapping on gamification applied to education. In *Proceedings of the 29th annual ACM symposium on applied computing* (pp. 216–222). ACM.
Sterbini, A., & Temperini, M. (2013). Openanswer, a framework to support teacher's management of open answers through peer assessment. In *Frontiers in education conference, 2013 IEEE* (pp. 164–170). IEEE.
Strijbos, J.-W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*(4), 265–269.
Sung, Y.-T., Chang, K.-E., Chiou, S.-K., & Hou, H.-T. (2005). The design and application of a web-based self-and peer-assessment system. *Computers & Education, 45*(2), 187–202.
Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of educational Research, 68*(3), 249–276.
Topping, K. J. (2009). Peer assessment. *Theory into practice, 48*(1), 20–27.
Tosic, M., & Nejkovic, V. (2010). *Trust-based peer assessment for virtual learning systems*. Springer.
Trahasch, S. (2004). From peer assessment towards collaborative learning. In *Frontiers in education, 2004. FIE 2004. 34th annual* (pp. F3F–F16). IEEE.
Tseng, S.-C., & Tsai, C.-C. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education, 49*(4), 1161–1174.
Tseng, S.-C., & Tsai, C.-C. (2010). Taiwan college students' self-efficacy and motivation of learning in online peer assessment environments. *The Internet and Higher Education, 13*(3), 164–169.
Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279.
Wang, Y., Liang, Y., Liu, L., & Liu, Y. (2014). *A motivation model of peer assessment in programming language learning*. arXiv preprint arXiv:1401.6113.
Wang, Y., & Vassileva, J. (2003). Trust and reputation model in peer-to-peer networks. In *Peer-to-Peer computing, 2003. (P2P 2003). Proceedings. Third international conference on* (pp. 150–157). IEEE.
Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 350–362.
Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 80–83.
Wulf, J., Blohm, I., Leimeister, J. M., & Brenner, W. (2014). Massive open online courses. *Business & Information Systems Engineering, 6*(2), 111–114.