Accepted Manuscript

Statistical semi-supervised system for grading multiple peer-reviewed open-ended works

Juan Ramón Rico-Juan, Antonio-Javier Gallego, Jose J. Valero-Mas, Jorge Calvo-Zaragoza

PII: S0360-1315(18)30200-8

DOI: 10.1016/j.compedu.2018.07.017

Reference: CAE 3413

To appear in: *Computers & Education*

Received Date: 16 December 2017

Revised Date: 24 May 2018

Accepted Date: 16 July 2018

Please cite this article as: Rico-Juan Juan.Ramó., Gallego A.-J., Valero-Mas J.J. & Calvo-Zaragoza J., Statistical semi-supervised system for grading multiple peer-reviewed open-ended works, *Computers & Education* (2018), doi: 10.1016/j.compedu.2018.07.017.

Full Title

Statistical semi-supervised system for grading multiple peer-reviewed open-ended works

Authors

Juan Ramón Rico-Juan
Universidad de Alicante, Departamento de Lengua es y Sistemas Informáticos
Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain
juanramonrico@ua.es

Antonio-Javier Gallego
Universidad de Alicante, Departamento de Lengua es y Sistemas Informáticos
Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain
jgallego@dlsi.ua.es

Jose J. Valero-Mas
Universidad de Alicante, Departamento de Lengua es y Sistemas Informáticos
Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain
jjvalero@dlsi.ua.es

Jorge Calvo-Zaragoza (corresponding author)
Universidad de Alicante, Departamento de Lengua es y Sistemas Informáticos
Carretera San Vicente del Raspeig s/n, 03690 Alicante, Spain
jcalvo@dlsi.ua.es

Abstract

In the education context, open-ended works generally entail a series of benefits as the possibility of develop original ideas and a more productive learning process to the student rather than closed-answer activities. Nevertheless, such works suppose a significant correction workload to the teacher in contrast to the latter ones that can be self-corrected. Furthermore, such workload turns to be intractable with large groups of students. In order to maintain the advantages of open-ended works with a reasonable amount of correction effort, this article proposes a novel methodology: students perform the corrections using a rubric (closed Likert scale) as a guideline in a peer-review fashion; then, their markings are automatically analyzed with statistical tools to detect possible biased scorings; finally, in the event the statistical analysis detects a biased case, the teacher is required to intervene to manually correct the assignment. This methodology has been tested on two different assignments with two heterogeneous groups of people to assess the robustness and reliability of the proposal. As a result, we obtain values over 95 % in the confidence of the intra-class correlation test (ICC) between the grades computed by our proposal and those directly resulting from the manual correction of the teacher. These figures confirm that the evaluation obtained with the proposed methodology is statistically similar to that of the manual correction of the teacher with a remarkable decrease in terms of effort.

Keywords

Computer-aided assessment
Automated grading
Open-ended works

Statistical semi-supervised system for grading multiple peer-reviewed open-ended works

Abstract

In the education context, open-ended works generally entail a series of benefits as the possibility of develop original ideas and a more productive learning process to the student rather than closed-answer activities. Nevertheless, such works suppose a significant correction workload to the teacher in contrast to the latter ones that can be self-corrected. Furthermore, such workload turns to be intractable with large groups of students. In order to maintain the advantages of open-ended works with a reasonable amount of correction effort, this article proposes a novel methodology: students perform the corrections using a rubric (closed Likert scale) as a guideline in a peer-review fashion; then, their markings are automatically analyzed with statistical tools to detect possible biased scorings; finally, in the event the statistical analysis detects a biased case, the teacher is required to intervene to manually correct the assignment. This methodology has been tested on two different assignments with two heterogeneous groups of people to assess the robustness and reliability of the proposal. As a result, we obtain values over 95% in the confidence of the intra-class correlation test (ICC) between the grades computed by our proposal and those directly resulting from the manual correction of the teacher. These figures confirm that the evaluation obtained with the proposed methodology is statistically similar to that of the manual correction of the teacher with a remarkable decrease in terms of effort.

Keywords: Computer-aided assessment, Automated grading, Open-ended works

Preprint submitted to Computers & Education *July 18, 2018*

1. Introduction

The progressive improvement in the accessibility to higher education has led to clear benefits for the individual, as for instance larger employment opportunities and higher salaries, as well as to the society, with higher tax incomes and more equity (Ma et al., 2016). However, such accessibility is also pushing current educational models to a limit: university classes are starting to be overcrowded, with large numbers of students for one single teacher (Shin & Teichler, 2014). In addition, in new learning paradigms such as the so-called Massive Open Online Courses (MOOCs), in which no limitation is imposed to the number of enrolled students, this effect is remarkably accused (Kaplan & Haenlein, 2016). Such decompensated student-to-grader ratio implies a remarkable time investment in the proposal and correction of assessment activities. Furthermore, such time-consuming task may also negatively affect the research and management duties to be developed by academics, which are tasks that proved to eventually benefit the student and should be thus maintained (García-Gallego et al., 2015).

A possibility to tackle this situation is to resort to automatic correction tools. Such schemes have been proved to successfully perform in tasks involving close-ended responses, as for instance multiple-choice tests, matching activities or numerical solutions (González-Barbone & Llamas-Nistal, 2008). However, when considering the case of open-ended works, automatic correction turns out to be considerably challenging (Bennett et al., 1997) as it demands a considerable amount of time and resources (Stanley & Porter, 2002). Moreover, while corrections for open-ended works are expected to provide feedback to the students, automatic approaches are generally limited in this sense (Bennett & Bejar, 1998; Hearst, 2000). Thus, human grading is typically required for these cases, which does not solve the issue of time consumption in the correction.

Given this situation, a largely considered alternative is to resort to a peer-review paradigm among the students (Kulkarni et al., 2013): instead of relying on the correction by a single teacher, the students assess the work of their peers. Such paradigm is not only ideal in the sense of reducing the correction workload for the teacher but also because it allows providing feedback to the student by means of checking alternative solutions to the same problem by other students (Nicol et al., 2014).

The main counterpart of this paradigm is that the evaluation criteria for peer review may be ambiguous, both by the lack of expertise of the student performing the correction and because of the subjectivity level of the activity to correct. Hence, the use of grading rubrics is commonly considered for providing a correction guide to the student and thus tackle the aforementioned issues (Jonsson & Svingby, 2007).

Nevertheless, in this educational context, a peer-assessment process for open-ended works still requires that the teacher supervises the process to both guarantee that the task is properly performed and that each student is graded according to the quality of the activity delivered. Therefore, note that the correction workload is now increased: besides assessing the actual activity, the teacher is required to examine the corrections by the peer reviewers to grade the student.

In this paper we present a methodology for tackling the issue of supervising the peer-review process. For that, we propose a semi-supervised correction scheme in which the teacher only intervenes in the correction process when severe discrepancies among the different peer reviewers are noticed. These discrepancies are automatically captured by the system using a statistical procedure designed to detect atypical marking patterns. To verify the goodness of this methodology, a study was carried out with two totally disjoint student groups, in which the activities are also corrected by the teacher to properly validate the assumptions.

In order to put our approach into context, Section 2 describes accurately the situation we aim at tackling; we then present our methodology in Section 3; the case of study is described in Section 4, together with its analysis and discussion; finally, we draw the main conclusions of this work in Section 5, along with some interesting avenues for the future.

2. Contextualization

By definition, close-answer activities generally depict a single correct solution for the task. Such particularity allows to implement automated grading systems in a relatively straight-forward fashion (Wang et al., 2008). Examples of tasks successfully tackled within this paradigm are the assessment of mathematical activities as, for instance, algebra tasks (Pacheco-Venegas et al., 2015), programming courses (Ala-Mutka, 2005), or any evaluation based on multiple-choice tests.

On the contrary, grading open-ended works is remarkably more challenging since these tasks do not generally exhibit a single correct solution. Given this variability, grading open-ended works demands a significant amount of time and effort. In addition, the need of periodically providing correction feedback to the students, makes human grading impossible to scale up when the number of students is relatively elevated (Kulkarni et al., 2013). Automated grading stands as a possible alternative, but the aforementioned variability in open-ended works remarkably complicates the task (Bennett et al., 1997). Research on this topic generally focuses on the use of Natural Language Processing techniques for performing this correction (e.g., Noorbehbahani & Kardan (2011); Xiong et al. (2012)), but their application is still limited.

In this context, peer-review grading has been typically considered as an alternative approach for tackling the aforementioned workload issue in which the proposed open-ended tasks are directly assessed by the students. This peer-review grading not only provides the aforementioned reduction in the correction workload but also entails additional advantages such as the chance for the student to check different solutions for the same problem (Panadero & Brown, 2017) and the provision of useful and timely feedback (Mulder et al., 2014).

As commented, the use of human graders in peer review facilitates the correction of open-ended works. Nevertheless, since the graders are the actual students, they may not have the proper criteria for performing the assessment of the task. In such situations, it is common to consider the use of grading rubrics to provide a correction guide to the grader and ensure certain consistency among correctors as well as a reduction in the time invested in the assessment process (Anglin et al., 2008).

Nevertheless, even in the context of peer-based grading, there is a need for supervision from the teacher: on the one hand, it should be checked that the different corrections and assessments guide the students to the actual goal of the work; on the other hand, given that grading completely relies on the students, the teacher is required to constantly inspect the process and detect possible frauds or errors in the corrections. While this constant supervision leads to satisfactory results for the students, it implies a significant time investment by the teacher, even superior to a classic teacher-to-student grading system.

In this paper we propose a system for reducing the workload of the teacher in a rubric-based peer-reviewing situation for open-ended works as the one described. The idea is that most students do perform properly both the development of the work itself and the peer-reviewing process and thus do not require any particular correction from the teacher, effort which must be invested in the exceptional cases in which these tasks are not properly done. With the use of fundamental statistical tools, the proposed system is able to detect deviations from the expected corrections as outliers and then inform the teacher to intervene in the correction of these atypical cases. The next section details the proposal.

3. Proposed methodology

Due to the nature of open-ended works, it is not possible to impose an objective and unique assessment methodology to a given problem. Nevertheless one may assume that, in certain contexts, the resolution may imply the use of certain resources or methodologies as, for instance, proper bibliography and citations in essays or a correct structure for developing the work. Thus, unambiguous criteria may be established to assess such points in the form of an evaluation rubric.

The use of that rubric allows addressing the correction task in a peer-review fashion: since the students are not required to know the correct solution of the global task but to assess a set of certain points in a collection of works, the evaluations may be trusted. However, in order to reduce the effect of possible incorrect assessments, the same activity must be corrected by different students to eventually produce an aggregate score from all of them.

As commented, the issue of incorrect assessments for producing the final mark may be reduced by considering several corrections of the same activity. This information is also useful in terms of assessing the actual student as a grader. If the assessment from a certain student significantly differs from the ones by the peers, it is unlikely that the student is properly developing the task. In this sense, we include the assessment by the student as part of the global evaluation of the task to encourage proper peer-based assessments.

At first, this scenario seems to even raise the burden of the teacher's workload. That is why the main idea of this work is to assume that those activities that have a relatively consensual score is because the evaluations obtained have been carried out properly. Thus, only those cases in which there is no consensus among the different peer-reviewers must be revised by the teacher. Note that this latter group is, in general, remarkably smaller compared to the total amount of works initially produced.

To carry out this proposal we implemented the scheme shown in Fig. 1. It consists of the following steps:

1. Activity submission from students and randomly assignation of reviewers.
2. Peer-review evaluation from students.
3. Automatic grading of the activities.

In the first step of the process, activities are collected from the students; then these activities are randomly distributed through all the students, avoiding self-evaluations, and assuring that each activity is evaluated by more than one reviewer. Presumably, the robustness of the presented methodology remarkably depends on the number of revisions each activity receives since higher

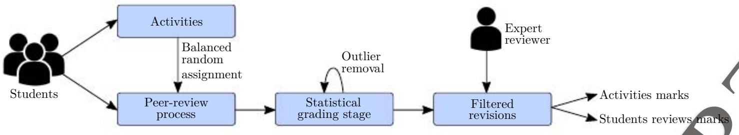


Figure 1: Graphical description of the method proposed. Once the developed activities have been delivered, they are distributed among the students for the peer-review stage; with those initial corrections, the system performs an initial revision based on statistics which additionally discards biased corrections by considering them as outliers; this initial grading is then forwarded to the expert reviewer who only has to manually check the cases for which the system is not confident enough; finally, the entire system provides the evaluation for each student in general as well as the scoring for each activity.

numbers of peer-review evaluations may mitigate the effect of incorrect evaluations. In the second step, each student evaluates the assigned activities. Once all evaluations are gathered (step 3), a final grade is computed for each activity. The system is expected to detect whether there have been atypical evaluations — somehow biased to deliberately raise or lower the eventual mark — so that the teacher supervises those particular cases and imposes the proper criterion for the correction.

The only point left to describe so far is how the final grade is computed out of the different peer evaluations as well as the criteria considered for stating an evaluation as atypical. Next section describes the proposed method through a conceptual example.

3.1. Conceptual example

Let us consider an activity that assigned for its correction to three different students: Alex, Jane, and Victor. The activity itself consists of three different items, and each one must be graded with a value of 0, 1, 2, or 3, which is inspired in the idea of a Likert scale. Note that such type of scale, which is commonly used in research questionnaires for social sciences, is characterized by the fact of being a discrete scale with, typically, 5 grading levels, avoiding the use of large marking ranges which may difficult the grading task. Nevertheless, note

Student	Item 1	Item 2	Item 3
Alex	1	1	1
Jane	1	1	3
Victor	3	3	2

Table 1: Scores given for each activity by the different individual evaluators.

	Alex	Jane	Victor
Alex	0.0	2.0	5.0
Jane	2.0	0.0	5.0
Victor	5.0	5.0	0.0

Table 2: Distance between pairs of evaluations using the L1 norm (Manhattan distance).

that the use of this range in the proposed system is not mandatory but simply recommendable.

Let us suppose that the individual marks given by the three reviewers are the ones depicted in Table 1.

Having evaluated the different individual items for the entire activity, the question now is how to eventually produce a final grading with these scores. As an illustrative example, the *Moodle* platform considers the individual evaluation that is the closest with respect to the others. However, even assuming this criterion, there exist multiple possibilities to estimate this distance. For instance, if the distance is measured amongst the different evaluations considering the L1-distance (Manhattan distance) or the L2-distance (Euclidean distance), we would obtain the pairwise distances amongst the reviewers as respectively observed in Tables **2** and **3**.

From these data we can already obtain what would be considered as the *central individual evaluation*, which somehow represents the consensus score given to the activity by a single reviewer. Let d_{ij} denote the distance of the scores of the reviewer i to the scores of the reviewer j. Then, we would define

	Alex	**Jane**	**Victor**
Alex	0.0	2.0	3.0
Jane	2.0	0.0	3.0
Victor	3.0	3.0	0.0

Table 3: Distance between pairs of evaluations using the L2 norm (Euclidean distance).

Name	**Manhattan**	**Euclidean**
Alex	7.0	5.0
Jane	7.0	5.0
Victor	10.0	6.0

Table 4: Central individual evaluation by the different reviewers considered using the Manhattan and Euclidean distances.

this *central individual evaluation* as:

$$\arg\min_{i} \ {}_{j} d_{ij}$$

Table 4 reports the *central individual evaluation* according to the Manhattan and Euclidean distances computed previously.

It can be observed that the *central individual evaluation* considerably differs depending on the distance measure considered (e.g., the distance between the corrections by Alex and Victor and remarkably lower with the Manhattan distance than with the Euclidean one). Thus, the problem here is that even when an evaluator has been considered as the *central evaluator*, it does not mean that the evaluation of all specific items are proportionate, which could lead to some unfair final grades. Thus, would not it be better to take advantage of all the item-wise evaluation to carry out a more consensual approach?

Our proposal is to consider the concept of item-wise central evaluation, which represents a more fine-grain evaluation than the previous definition of the central evaluation. For this we have to compute the central value of each item from the revision matrix, i.e., the median value of the different evaluations given to a sin-

Student	Item 1	Item 2	Item 3
Alex	1	1	1
Jane	1	1	3
Victor	3	3	2
Median	1	1	2

Table 5: Median value of the scores given by each reviewer to the individual activities of the task.

Student	Manhattan	Euclidean
Alex	1.0	1.0
Jane	1.0	1.0
Victor	4.0	2.8

Table 6: Distances in terms of both the Manhattan and Euclidean dissimilarity function from the median-based centric evaluation to the assessment given by each single reviewer (see Table 5).

gle activity. While other statistical descriptors such as the maximum/minimum or the arithmetic mean could be considered, we resort to the median descriptor due to its robustness against outliers.

The results of applying this item-wise central evaluation to our conceptual example can be seen in Table 5.

In this case, the central item-wise evaluation would be represented by the vector $[1, 1, 2]$. We can then compute its distance to each single reviewer, which may give hints of how close each individual evaluation is to the central one. Table 6 reports the final distance between each individual evaluation and the central item-wise evaluation.

In this case, we consider that L1 distance better reflects the discrepancy, since it is a measure that assumes and yields discrete values such as those requested from the evaluator. The L1 distance, therefore, can be better interpreted than the L2 in this context.

As previously commented, the idea in this case is to use these values for both assessing the actual work of the students and finding atypical correction values (outliers which remarkably differ from the central evaluation), since both influence the eventual mark of the student. Thus, from the prior calculations we proceed to determine the threshold used to discriminate atypical evaluations.

For performing this analysis, we resort to descriptive statistics for sample-based distributions. Given a distribution, a sample S is considered atypical or an outlier when its value is over a threshold U obtained as

$$\mathrm{U} = Q3 + 1.5 \cdot IQR$$

where $Q3$ constitutes the value representing the third quartile of the sample distribution, and IQR the *inter-quartile ratio* which is defined as the difference between the samples in the third and first quartiles (i.e., $IQR = Q3 - Q1$).

In the context of our proposal, our sample distribution is the one formed by the commented distances between the central evaluation and the individual scores by the reviewers. In this context, and as simplistic example, if we obtain a distribution with $Q3 = 2$ and an $IQR = 1$ we can set the threshold (U) for detecting outliers as:

$$\mathrm{U} = Q3 + 1.5 \cdot IQR = 2 + 1.5 \cdot 1 = 3.5$$

Therefore, Victor, who had a distance between his evaluation and the central evaluation of 4 (see Table 6), would surpass this threshold and would be thus considered an outlier. In other words, his evaluation would be considered as an atypical review. In this case, its evaluation would be discarded and the median would be recalculated to obtain the central evaluation without considering Victor's.

If we analyze Victor's evaluation, we observe that he evaluated items 1 and 2 with a deviation of 2 points with respect to the consensual value. This accumulation of deviation makes his evaluation be considered as atypical, in spite of matching the central evaluation for item 3.

The final evaluation of the activity will then be obtained from the corrections that have reached consensus. In addition, the teacher would be warned when an outlier is detected to manually review it. Thus, our proposal would consist of a semi-automatic system that would generally return the final evaluation without human intervention, only requiring expert supervision in the cases where it is necessary due to the disparity in the results. This feature will also allow us to determine how good the students are being evaluating, and grading them accordingly.

4. Case of study

This section presents the case of study carried out to test the validity and reliability of the proposed semi-supervised grading methodology proposed. The study consisted in assessing the work of a university module using two different activities: a first one which deals with intellectual property (see Appendix A) and a second activity consisting on creating a Webquest (see Appendix B). Each activity was assessed with two independent sets of students, i.e., morning and afternoon sessions of the module during the same academic year.

Each activity was solved by groups of students, most of them comprising 2 people. Eventually, each student reviewed an average of 5 activities, and so an average of 10 different evaluations per activity were obtained. The evaluation followed the rubrics described in Appendix A for the intellectual property activity and Appendix B for the Webquest one, being each of them evaluated with a series of individual items with a fixed discrete range for the scores. Table 7 describes the data considered for this study in terms of the number of collected activities, the amount of independent reviewers, and the total obtained peer-review works.

Students were told that part of their final grade (30 %) depended on their performance in the peer-review stage: in order to prevent biased or vague corrections, assessments detected as atypical using the process in Section 3 would be penalized. The rest of the grading (70 %) corresponded to the actual score of the activity itself, which is obtained with the marking from the peer reviewers.

Group	Activity	Collected activities	Number of reviewers	Obtained peer reviews
Group 1	Intell. prop.	44	94	470
	Webquest	42	95	475
Group 2	Intell. prop.	46	79	395
	Webquest	43	83	409
Summary		175	351	1,749

Table 7: Description in terms of number of activities collected, number of independent reviewers, and amount of obtained peer-review works for the study of the proposed methodology.

These proportions (70-30 %) were estimated based on the time and effort the student should have dedicated to each part. Note that the final mark is given in the range 0 to 10 independently of the grading scale used in the peer-review stage.

To prove the validity of the proposed method, all the activities were also evaluated directly by the teacher to analyze the results obtained in these three scenarios:

- Automatic: the system directly computes the final evaluation of the activities and discards the atypical corrections, but it never consults the teacher.
- Semi-automatic: the system automatically computes the evaluation for all the activities in which consensus has been reached, requiring the expert teacher to intervene only those without consensus.
- Expert: all the activities are evaluated by an expert teacher, considering that evaluation as the central one.

4.1. Study conducted

We now present and analyze the results obtained for the three scenarios previously commented for the two works considered: the *Intellectual property* one and the *Webquest*. This analysis shall prove the validity of the proposed methodology by contrasting the results obtained from the expert correction (the typical evaluation in which all work relies on the teacher) against the automatic and semi-automatic correction proposals. Apart from the two commented studies, a third section is included to compare the effectiveness of the proposed correction methodology with another automatic correction strategy based on the Moodle paradigm.

4.1.1. Intellectual property activity

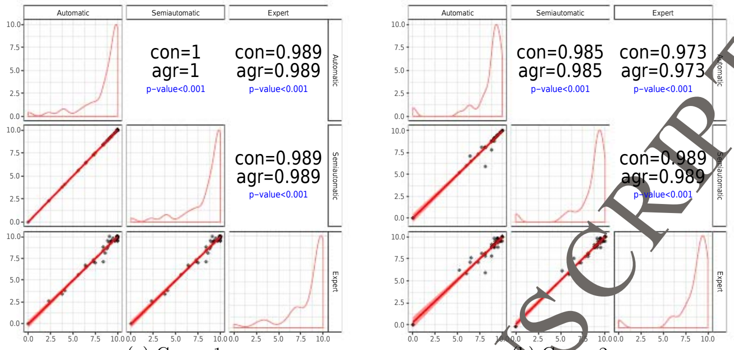
In this first analysis we focus on the assessment of the *Intellectual property* activity, which dealt with the issue of evaluating the proper use of Creative Commons licenses in images.

As a first point we analyze the degree of correlation between the three different scenarios considered. For that we make use of the *Intraclass Correlation Coefficient* (ICC) by Bartko (1966), which measures in terms of the concepts of *agreement* and *consistency* the degree of correlation between two data populations.

Figure 2 shows the results of this ICC correlation analysis for the *Intellectual property* task. Attending to these graphs, both the *automatic* and *semi-automatic* scenarios show a high degree of correlation with the *expert* one in terms of the *agreement* and *consistency* measures. For both groups of students, these values are consistently above 0.97, which is a remarkable correlation indicator. These results suggest that the proposed correction methodology is practically equivalent to manually correcting all single activities, and thus could be used to reduce the correction workload expected to be done by the teacher.

In addition, Figure 3 shows the results of reproducing the previous analysis but discarding the subjective questions within the rubric . The idea is to study whether these points affect the robustness of the approach. As expected,

Subjective points are highlighted in the corresponding appendices. The ratio of these questions in the considered rubrics is around 30 %.



(a) Group 1

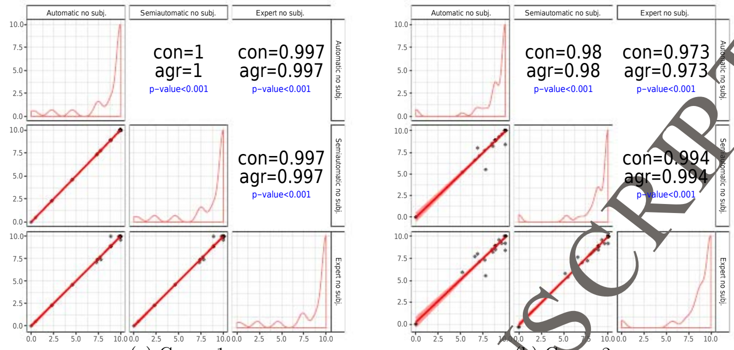
(b) Group 2

Figure 2: Results of the pair-wise comparison of the automatic, semi-automatic, and expert scenarios for the *Intellectual property* activity. The diagonal plots show the density of observations, the lower diagonal shows the correlation among observations, and the upper diagonal shows the numeric results of the intraclass correlation coefficients (consistency and agreement). The Y-axis represents the grading of the activities in a 0-to-10 range.

the conclusion gathered from the previous analysis are maintained with the particularity of being the correlation figures higher than in the previous case due to removing the variability of the subjective points.

Figure 4 shows the results of the pair-wise differences of the grading of the *expert* with respect to both the *automatic* and the *semiautomatic* scenarios. As it may be observed, for all cases the resulting distribution are located around zero, thus pointing out that the two scenarios compared are quite similar. This effect is more noticeable when the subjective questions are removed from the analysis as there is less variability in the answers.

Furthermore, Table 8 shows the mean Pearson correlation values between the *Automatic* and *Semiautomatic* approaches against the expert criterion for the cases in which the subjective questions are both considered and discarded. The obtained correlation values (a figure of 0.98 is obtained in the fully-automated case whereas a correlation value of 0.99 is depicted in the semiautomatic case with the expert intervention) show the similarity between the proposed ap-



(a) Group 1

(b) Group 2

Figure 3: Results of the pair-wise comparison of the automatic, semi-automatic, and expert scenarios for the *Intellectual property* activity without considering subjective questions. The diagonal plots show the density of observations, the lower diagonal shows the correlation among observations, and the upper diagonal shows the numeric results of the intraclass correlation coefficients (consistency and agreement). The Y-axis represents the grading of the activities in a 0-to-10 range.

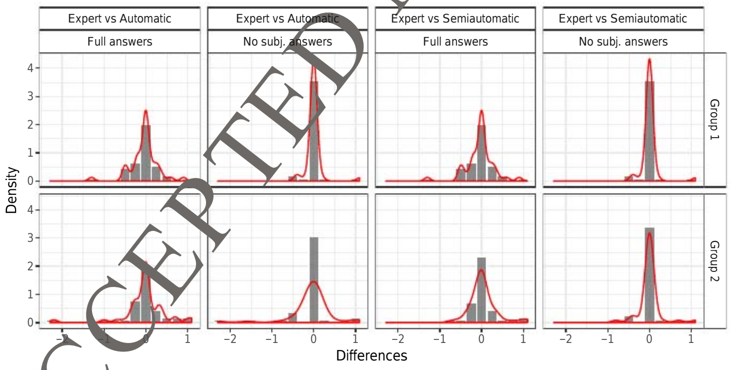


Figure 4: Density histograms of the pair-wise differences of the expert scenario against the automatic and semi-automatic ones for the expert scenarios for the two groups of the *Intellectual property* activity. The *Full answers* label represents the case in which all questions are considered while in *No subj. answers* the subjective points are discarded.

All questions		Removing subjective questions	
Automatic	Semiautomatic	Automatic	Semiautomatic
0.98	0.99	0.98	0.99

Table 8: Results of the mean Pearson correlation value obtained for all the *Intellectual property* activities comparing the *Automatic* and *Semiautomatic* approaches against the expert criterion. The *All questions* column depicts the case in which no questions are discarded for obtaining the indicator, whereas the *Removing subjective questions* one shows the results when the subjective questions are not considered.

proaches and the manual correction, thus suggesting this method as a proper one for such grading tasks.

Finally, the whisker plot in Figure 5 shows the Pearson correlation coefficient between the automatic and expert scenarios as the number of peer reviewers is progressively increased. Such coefficient assesses the degree of correlation/similarity between two statistical distributions, which in this case turn out to be the one obtained by the correction system proposed and the manual correction of the works. Thus, the idea behind this analysis is to assess how similar are the correction results in the two aforementioned scenarios. When checking the results obtained, it can be observed that increasing the number of peer reviewers reduces the dispersion in the correlation coefficient, thus reporting a larger agreement among the correctors. A particular point to remark is that it would be expected that a larger number of correctors implied a higher correlation between the *automatic* and *expert* scenarios. Nevertheless, this effect is only observed for the population of the *Group 1* while for the *Group 2* this correlation achieves its maximum with only 6 reviewers.

4.1.2. Webquest activity

In this second analysis we focus on the assessment of the *Webquest* activity, which dealt with the issue of creating such type of platform for introducing a certain topic at the election of the student.

As in the previous activity, we initially assess the degree of correlation

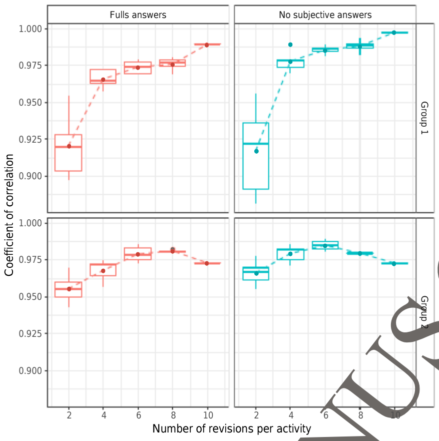
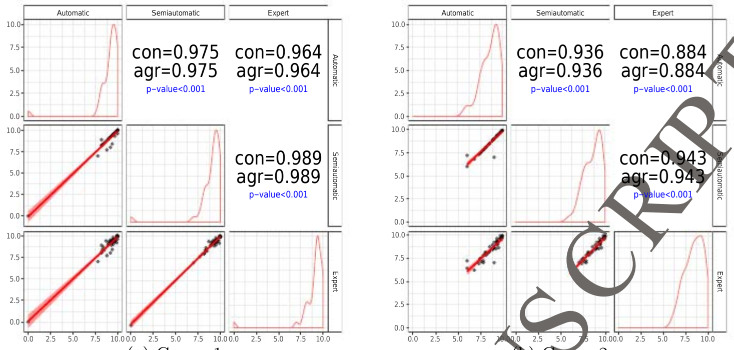


Figure 5: Pearson correlation coefficient between the automatic and expert scenarios for the *Intellectual property* activity as the number of revisions per activity is increased. The *Full answers* label represents the case in which all questions are considered while in *No subj. answers* the subjective points are discarded. Dashed lines link mean values.

amongst the three different scenarios involved. For that, we consider the same *consistency* and *agreement* measures introduced for the *Intellectual property* activity. The results obtained for the two groups of study may be seen in Figure 6 for the case in which all question are considered and Figure 7 for the case in which subjective questions are discarded from the analysis.

The figures obtained for this analysis are consistent with the ones obtained for the previous task, with the particularity of being the correlation values slightly lower. Nevertheless, the conclusions obtained for the previous task may still be valid: on the one hand, these values support the claim of the proposed correction methodology being practically equivalent to manually correcting all single activities; on the other hand, as in the previous case, the removal of the subjective points increases the correlation scores as the variability in the answers is severely reduced.

The results of the pair-wise differences of the grading of the *expert* and the



(a) Group 1

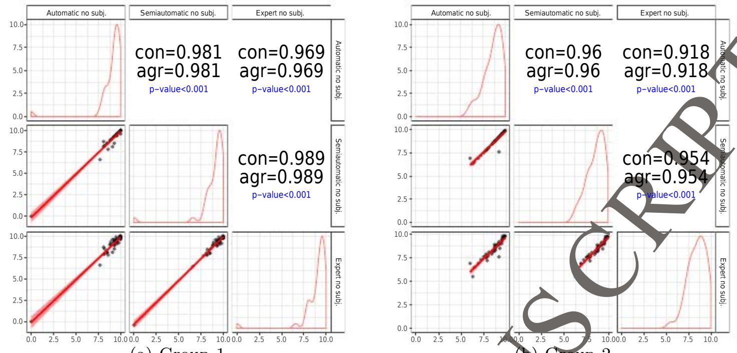
(b) Group 2

Figure 6: Results of the pair-wise comparison of the automatic, semi-automatic, and expert scenarios for the *Webquest* activity. The diagonal plots show the density of observations, the lower diagonal shows the correlation among observations, and the upper diagonal shows the numeric results of the intraclass correlation coefficients (consistency and agreement). The Y-axis represents the grading of the activities in a 0-to-10 range.

automatic and the *semiautomatic* scenarios are shown in Figure 8. As in the previous task, the resulting distributions are consistently around zero, which confirms the initial idea of that the scoring done in the different scenarios is quite similar. Thus, using the semi-automatic, or even the automatic, scenario may be practically equivalent to correcting all assignments manually.

As a last point to address, the whisker plot in Figure 9 shows the Pearson correlation coefficient between the automatic and expert scenarios as the number of peer reviewers is progressively increased. As in the previous task, the main conclusion which can be obtained from this analysis is that increasing the number of peer reviewers reduces the dispersion in the correlation coefficient, which suggests that the larger number of correctors, the better.

Finally, Table 9 shows the mean Pearson correlation values between the *Automatic* and *Semiautomatic* approaches against the expert criterion for the cases in which the subjective questions are both considered and discarded. As it happened in the previous activity, again the obtained correlation values show



(a) Group 1

(b) Group 2

Figure 7: Results of the pair-wise comparison of the automatic, semi-automatic, and expert scenarios for the *Webquest* activity without considering subjective questions. The diagonal plots show the density of observations, the lower diagonal shows the correlation among observations, and the upper diagonal shows the numeric results of the intraclass correlation coefficients (consistency and agreement). The Y-axis represents the grading of the activities in a 0-to-10 range.

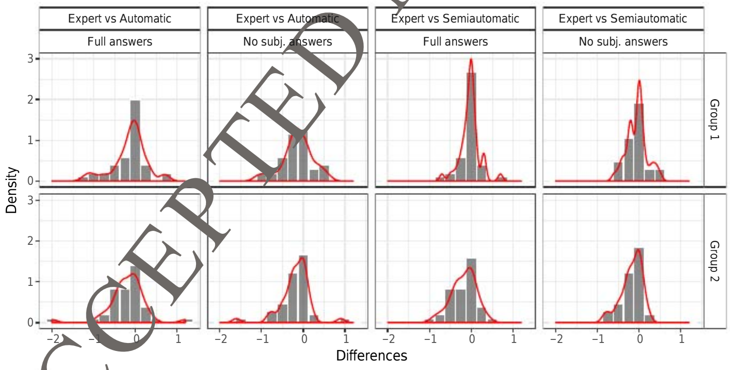


Figure 8: Density histograms of the pair-wise differences of the expert scenario against the automatic and semi-automatic ones for the expert scenarios for the two groups of the *Webquest* activity. The *Full answers* label represents the case in which all questions are considered while in *No subj. answers* the subjective points are discarded.

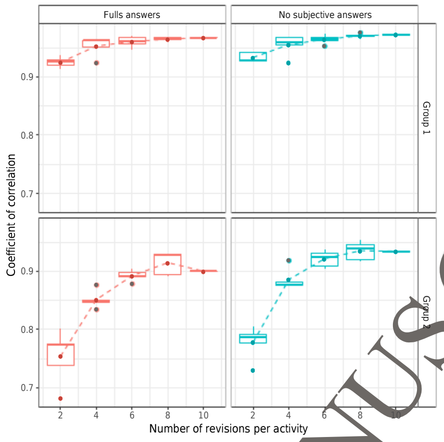


Figure 9: Pearson correlation coefficient between the automatic and expert scenarios for the *Webquest* activity as the number of revisions per activity is increased. The *Full answers* label represents the case in which all questions are considered while in *No subj. answers* the subjective points are discarded. Dashed lines link mean values.

a remarkable similarity between the proposed approaches and the manual correction (correlation values of 0.95 and 0.98 for the fully-automated and the semiautomatic cases, respectively), which suggests that the proposed grading method is capable of performing the corrections in a quite similar level to an expert agent.

4.2. Discussion

Once we have analyzed the two considered activities with the proposed grading methodology, we shall now present a brief discussion section to comment some general conclusions observed in the two experiments performed.

Note that our proposal includes the early detection of erroneous, or even biased, corrections, which are considered to be outliers. Such elements are localized and removed before computing the final grade. Our hypothesis is that computing the grades with outliers is less reliable as they may introduce a bias, reason why the previous sections directly reported the results under

All questions		Removing subjective questions	
Automatic	Semiautomatic	Automatic	Semiautomatic
0.95	0.98	0.95	0.98

Table 9: Results of the mean Pearson correlation value obtained for all the *Webquest* activities comparing the *Automatic* and *Semiautomatic* approaches against the expert criterion. The *All questions* column depicts the case in which no questions are discarded for obtaining the indicator, whereas the *Removing subjective questions* one shows the results when the subjective questions are not considered.

this assumption. However, we still consider that an additional analysis stage is required to assess the impact of such consideration. Table 10 presents the correlation coefficients of the grades proposed by our system, with and without performing the outlier removal procedure, with respect to the correction made entirely by the expert corrector, i.e. the teacher. For the sake of analysis, we also include in this table the number of outliers detected in each activity.
As it can be observed, the number of outliers detected in each case of study is very low, and so the impact they have might be limited. However, even with this small amount, all the correlations increase by the fact of removing the outliers, thus validating our assumption. Actually, there are some remarkable cases like that of the Webquest activity, for which removing the outliers leads to difference of up to 3 %.

As a second point to deal with, we now focus on the number of corrections required for each work and discuss whether it exists a sweet spot in which no more corrections are required and which is universal for any kind of activity. As commented in the previous sections, Figs. 5 and 9 showed the correlation between the automatic and expert correction scenarios for the two activities considered. It can be observed that, in both cases, there is a turning point in which no additional revisions per activity are required to obtain a proper correlation value, which for both the proposed and assessed activities is around 6 corrections/revisions. Nevertheless, note that these conclusions about the optimal number of corrections required are only meant for the two activities

Activity	Group	No. of outliers	Keep outliers		Remove outliers	
			Full	No subj.	Full	No subj.
Intellectual property	Group 1	23/470	98.68	99.75	98.92	99.75
	Group 2	17/402	97.11	96.76	98.34	98.98
	Average	40/872	98.04	98.63	98.67	99.50
Webquest	Group 1	19/484	96.73	97.28	98.94	98.64
	Group 2	14/411	92.43	91.75	95.87	95.97
	Average	33/895	95.52	95.81	97.92	97.84

Table 10: Pearson correlation coefficient (in %) between the expert assessment and the proposed method with or without removing the outliers detected. Results provided are divided in terms of the activity and group, while an average result is facilitated for each activity. Note that *Full* and *No subj.* labels denote the cases in which all questions are considered and the case in which subjective points are discarded, respectively.

considered as this indicator is intrinsic to the design of the activity itself.

Related to this point, it should be born in mind that this system is designed for overcrowded classrooms, with the aim of relieving the teacher's workload. In this context the main idea is that, for each activity, a reasonably large number of corrections may be retrieved. However, for small groups, the situation is dramatically different as the expected number of corrections is lower. In this sense, the application of the grading system to such non-overcrowded classrooms may also make sense because the review activity (i.e., the one carried out by the student itself) is evaluated and reflected on the personal mark of the student itself. However, to avoid conflicts and possible correction biases, the peer-review activities could be assigned to disjoint groups of students (for example, by crossing the assignments between morning and afternoon groups). Nevertheless, note that if the number of samples is very small, the approach becomes meaningless since there would be no reliable statistical significance, which would imply that the teacher is required to correct all the activities manually.

It should be noted that an additional potential weakness of the approach may arise when many students agree to grade similarly, as there is no way to detect it automatically. A possible solution to detect such anomalies is that teachers perform some manual evaluations on randomly selected activities by the students. In addition, if students are aware of this procedure, they would be more reluctant to follow the aforementioned strategy because a wrong peer-review grading, even if agreed with their peers, will have negative repercussions on their mark. In addition, it seems advisable to include the manual verification for those activities whose evaluation is borderline, even if the statistical system does not find any anomaly, as they represent more sensitive cases.

While the use of a correction rubric is clearly necessary for a rather standard correction procedure, it must be pointed out that such rubric may constrain the experience of the student. In general, the use of rubrics may limit the feedback to be learned by the student when reviewing other works as the correction procedure is completely biased towards certain points. While the use of less categorical correction rubrics may palliate the commented effect, the main drawback with respect to our method is that the correction rubric must be quite categorical and restrictive to guarantee the proper functionality of the proposal.

Finally, while the proposed correction method is able to significantly reduce the correction workload for the teacher, we would like to stress again that the system is not expected to be fully automatic but semi-automatic, thus being always required some supervision by an expert, i.e. the teacher.

4.3. Comparative results with the Moodle system

The idea in this final section is to perform a comparative analysis between the proposed methodology and an established automatic correction strategy in which peer reviewing is also considered. For that, we have selected the strategy by *Moodle* due to its extensive use in the education context.

The strategy followed by *Moodle* is relatively similar to the one proposed: once the works are produced, they are uploaded to the platform; after that, the system assigns each work to a certain number of reviewers, who assess the work and provide a score; finally, once all score are obtained, the actual score of the work is computed as the median value of the set of individual scores provided by the peer reviewers.

Note that, unlike the proposed methodology, *Moodle* considers all single reviews as equally important. Such assumption disregards the fact that some scorings might be biased for some reason (outliers of the distribution), which would suppose an uneven evaluation of the student, both in terms of being beneficial or detrimental. While the median descriptor has been typically considered for outlier detection and removal in a large number of disciplines (e.g., noise removal in audio and speech signals as in the work by Kauppinen (2002)), for the cases in which the size of the sample distribution is not large enough, the solely use of this descriptor may not be enough. At this point, the idea of considering human intervention for such doubtful cases takes special relevance.

In order to prove this premise, we consider the two previously analyzed activities (the *Intellectual property* and the *Webquest* ones) and we simulate both correction strategies, the proposed method and the standard *Moodle* procedure, to obtain the scores for the students. The results from these two strategies are correlated with the assessment performed by an expert agent, i.e. the manual correction of the strategy, to check which of them provides a higher correlation degree. As in the previous analyses, we check the case in which all questions are considered as well as the one in which the subjective points are discarded. Table 11 shows the results of such experiment.

As it may be checked, the results obtained show that the proposed methodology is consistently more correlated with the expert results than the *Moodle* method except for one case in which both strategies obtain the same correlation value (the *Intellectual property* activity of Group 1 for the case in which no subjective points are considered). While in some cases the difference in terms of correlation for the two methods is not that remarkable (e.g., in the *Intellectual property* for Group 1, the difference is around the 0.1 %), for the case of the *Webquest* activity for Group 2 in the cases in which no subjective questions are considered, the difference is above the 4 %.

Activity	**Group**	**Proposed method**		**Moodle system**	
		Full	**No subj.**	**Full**	**No subj.**
Intellectual property	**Group 1**	96.63	95.66	96.48	95.66
	Group 2	98.34	98.98	97.11	96.76
	Average	97.11	96.59	96.56	95.99
Webquest	**Group 1**	98.94	98.64	96.74	97.28
	Group 2	95.87	95.97	92.46	91.75
	Average	97.92	97.84	95.56	95.81

Table 11: Pearson correlation coefficient (in %) between the expert assessment and two different corrections methods: the methods proposed and the *Moodle* strategy. Results provided are divided in terms of the activity and group, while an average result is facilitated for each activity. Note that *Full* and *No subj.* labels denote the cases in which all questions are considered and the case in which subjective points are discarded, respectively.

5. Conclusions and future works

We propose a novel semi-supervised correction system to alleviate teachers' workload in the evaluation of open-ended works. Our approach is based on ideas from student-based peer-review methods, corrections rubrics, and statistical analysis for the detection of biased corrections from the students.

The system proposes the teacher as a verification agent rather than an assessment one. The idea is that the teacher assists the students while conducting the peer-review activity during the sessions with a double aim: on the one hand, reducing the time devoted to the individual corrections of all the students; on the other hand, encouraging the students to not only study other possible resolutions to the task by other students but also to develop a critic mentality by forcing them to objectively assess these other resolutions.

Results from the experimentation carried out show high correlation rates between the proposed peer-review methodology and the assessment directly from the teacher, thus proving the robustness and reliability of the proposed approach. Moreover, comparative experiments between the proposed method and a strategy followed by the well-known *Moodle* platform show that with the former method achieves results closer to the manual correction of the works than the latter one, thus proving a superior effectiveness.

As future works, other types of measures to detect biased corrections and their impact on the overall scores could be studied. This system may also be complemented with another anti-plagiarism systems in order to detect equal or very similar activities and/or evaluations. The application of this new methodology to MOOC courses could be interesting with the aim at automatically evaluating open-ended works in such context.

References

Ala-Mutka, K. M. (2005). A survey of automated assessment approaches for programming assignments. *Computer Science Education*, *15*, 83–102.

Anglin, L., Anglin, K., Schumann, P. L., & Kaliski, J. A. (2008). Improving the Efficiency and Effectiveness of Grading Through the Use of Computer-Assisted Grading Rubrics. *Decision Sciences Journal of Innovative Education*, *6*, 51–73.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, *19*, 3–11.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*, 9–17.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an Automatically Scorable, Open-Ended Response Type for Measuring Mathematical Reasoning in Computer-Adaptive Tests. *Journal of Educational Measurement*, *34*, 162–176.

García-Gallego, A., Georgantzís, N., Martín-Montaner, J., & Pérez-Amaral, T. (2015). (How) Do research and administrative duties affect university professors' teaching? *Applied Economics*, *47*, 4868–4883.

Gonzalez-Barbone, V., & Llamas-Nistal, M. (2008). eAssessment of open questions: An educator's perspective. In *38th Annual Frontiers in Education Conference* (pp. F2B–1). IEEE.

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems and their Applications*, *15*, 22–37.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*, 130–144.

Kaplan, A. M., & Haenlein, M. (2016). Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. *Business Horizons*, *59*, 441–450.

Kauppinen, I. (2002). Methods for detecting impulsive noise in speech and audio signals. In *Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on* (pp. 967–970). IEEE volume 2.

Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S. R. (2013). Peer and Self Assessment in Massive Online Classes. *ACM Transactions on Computer-Human Interaction*, *20*, 1–31.

Ma, J., Pender, M., & Welch, M. (2016). Education Pays: The Benefits of Higher Education for Individuals and Society. *College Board*, .

Mulder, R. A., Pearce, J. M., & Baik, C. (2014). Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education*, *15*, 157–171.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, *39*, 102–122.

Noorbehbahani, F., & Kardan, A. A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. *Computers & Education*, *56*, 337–345.

Pacheco-Venegas, N. D., López, G., & Andrade-Aréchiga, M. (2015). Conceptualization, development and implementation of a web-based system for automatic evaluation of mathematical expressions. *Computers & Education*, *88*, 15–28.

Panadero, E., & Brown, G. T. (2017). Teachers' reasons for using peer assessment: positive experience predicts use. *European Journal of Psychology of Education*, *32*, 133–156.

Shin, J. C., & Teichler, U. (2014). The Future of University in the Post-Massification Era: A Conceptual Framework. In *The Future of the Post-Massified University at the Crossroads: Restructuring Systems and Functions* (pp. 1–9). Cham: Springer International Publishing.

Stanley, C. A., & Porter, M. E. (2002). *Engaging Large Classes: Strategies and Techniques for College Faculty*. ERIC.

Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education*, *51*, 1450–1466.

Xiong, W., Litman, D., & Schunn, C. (2012). Natural Language Processing techniques for researching and improving peer feedback. *Journal of Writing Research*, *4*, 155–176.

Appendix A. Intellectual property activity

This practice is devoted to learning how to search and properly use images licensed by Creative Commons (CC). For that, the evaluation focuses on checking whether the images are correctly cited with the corresponding CC license type. The theme of the assessment is free, but all images must be consistent with that topic throughout the work. Presentation and originality of the work is considered in addition to the correctness of those licenses. The task is carried out in pairs.

The following sections introduce the actual description of the task which is facilitated to the student as well as the correction rubric considered for the peer-review process. Subjective points of the correction rubric are marked with an asterisk (*).

Appendix A.1. Assignment description: Presentation and development work on images CC

The aim of this task is dealing with Intellectual property and Creative Common licenses to reinforce the notions of how to cite different sources of information. For that, in this activity you are required to prepared a document with different pictures under Creative Common licenses and properly citing them.

The instructions for preparing the work are the following ones:

- You may use any online tool for preparing the document as, for instance, a presentation or a document from Google Drive, a presentation with Prezi, etc.
- You must include the title chosen for the work, with a brief description of its intent, and 5 pictures on the chosen topic with their respective CC license.
- The characteristics of the licenses of each of the images will have to be:
  - Image 1: Any type of use with the sole requirement of citing the author (2 points).

- Image 2: Commercial use not allowed (2 points).
- Image 3: Modification of the original image is not allowed (2 points).
- Image 4: Modifications of the image as well as in their derivative works are allowed (2 points).
- Image 5: Modifications in the initial image are allowed with the restriction of keeping the same license in derivative works (2 points).

- Additional instructions:
  - The different images must follow the order previously commented according to their licenses. Also, identify the image in the caption following the previous list.
  - Create a visible link/hyperlink to the source where the image was extracted from. The source is required for this work, both in the text and in the caption of the image. This facilitates the process of checking the license used.
  - Always quote the author of the image. If the name of the author is not available, use the '*(no author)*' token.

Appendix A.2. Correction rubric

Image 1: Any type of use with the sole requirement of citing the author. (2 points)

1.1 Are the author and source link cited? (Yes; No)

1.2 Is the Creative Commons license with its particular type clearly indicated? (Yes; No)

1.3 Does the license clearly indicate that the right of Attribution, Recognition or Public Domain is reserved? Does the license included match the one stated in the instructions for this activity? (0 - both licenses are incorrect; 1 - only the work is correct; 2 - only the link is correct; 3 - both licenses are correct)

Note: Only the licenses that indicate the abbreviation [BY], the word [Recognition] or [Attribution], the symbol [one person], or the equivalent of Public Domain are correct. Therefore, the following licenses are valid: CC BY, or CC0.

1.4 (*) Rate the image from 0 to 3 considering the presentation and its originality: 0 - very unoriginal and very poor presentation; 1 - little original and poor presentation; 2 - original and well presented; 3 - very original and very well presented. Note that the figure must belong to the chosen topic of the work.

Image 2: Commercial use not allowed. (2 points)

2.1 Are the author and source link cited? (Yes; No)

2.2 Is the Creative Commons license with its particular type clearly indicated? (Yes; No)

2.3 Does the license clearly indicate that the commercial use of the work is NOT allowed? Does the license included match the one stated in the instructions for this activity? (0 - both licenses are incorrect; 1 - only the work is correct; 2 - only the link is correct; 3 - both licenses are correct)

Note: All licenses indicating the [NC], the word [Non-commercial] or the [euro or crossed-out dollar] symbol, which does not allow commercial use, shall be valid. Therefore, the following licenses are valid: CC BY-NC, CC BY-NC-SA, and CC BY-NC-ND.

2.4 (*) Rate the image from 0 to 3 considering the presentation and its originality: 0 - very unoriginal and very poor presentation; 1 - little original and poor presentation; 2 - original and well presented; 3 - very original and very well presented. Note that the figure must belong to the chosen topic of the work.

Image 3: Modification of the original image is not allowed. (2 points)

3.1 Are the author and the source link cited? (Yes; No)

3.2 Is the Creative Commons license with its particular type clearly indicated? (Yes; No)

3.3 Does the license clearly indicate that NO derived works are allowed? Does the license include match the one stated in the instructions for this activity? (0 - both licenses are incorrect; 1 - only the work is correct; 2 - only the link is correct; 3 - both licenses are correct)

Note: All licenses indicating the [No Derivative Works] or the [Equal Symbol] symbol, which does not allow the derivative work, are indicated by the abbreviation [ND]. Therefore, the following licenses are valid: CC BY-ND and CC BY-NC-ND.

3.4 (*) Rate the image from 0 to 3 considering the presentation and its originality: 0 - very unoriginal and very poor presentation; 1 - little original and poor presentation; 2 - original and well presented; 3 - very original and very well presented. Note that the figure must belong to the chosen topic of the work.

Image 4: Modifications of the image as well as in their derivative works are allowed. (2 points)

4.1 Are the author and source link cited? (Yes; No)

4.2 Is the Creative Commons license with its particular type clearly indicated? (Yes; No)

4.3 Does the license of the image allow its modification and also the change of license? Does the license included match the one stated in the instructions for this activity? (0 - both licenses are incorrect; 1 - only the work is correct; 2 - only the link is correct; 3 - both licenses are correct)

Note: All valid licenses are the ones in which it is NOT indicated that the image cannot be modified and does not require the same license in the derivative work. Therefore, the following licenses are valid: CC BY, CC0 and CC BY-NC.

4.4 (*) Rate the image from 0 to 3 considering the presentation and its originality: 0 - very unoriginal and very poor presentation; 1 - little original and poor presentation; 2 - original and well presented; 3 - very original and very well presented. Note that the figure must belong to the chosen topic of the work.

Image 5: Modifications in the initial image are allowed with the restriction of keeping the same license in derivative works (2 points).

5.1 Are the author and source link cited? (Yes; No)

5.2 Is the Creative Commons license with its particular type clearly indicated? (Yes; No)

5.3 Does the license clearly indicate that the derivative work must use the same license as the initial image? Does the license included match the one stated in the instructions for this activity? (0 - both licenses are incorrect; 1 - only the work is correct; 2 - only the link is correct; 3 - both licenses are correct).

Note: The licenses [SA], the word [Share Alike] or the [circular arrow] symbol indicate that the derivative work must exhibit the same license as the original work. Therefore, the following licenses are valid: CC BY-NC-SA and CC BY-SA.

5.4 (*) Rate the image from 0 to 3 considering the presentation and its originality: 0 - very unoriginal and very poor presentation; 1 - little original and poor presentation; 2 - original and well presented; 3 - very original and very well presented. Note that the figure must belong to the chosen topic of the work.

Appendix B. Webquest

The objective of this activity is to develop a Webquest using the Google Sites tool (http://sites.google.com). This work may be developed in groups of up to three people. The theme of the Webquest is free but must be established in advance.

The following sections introduce the actual description of the task which is facilitated to the student as well as the correction rubric considered for the peer-review process. Subjective points of the rubric are marked with an asterisk (*).

Appendix B.1. Preparation of the Webquest

- First of all, you have to log into Google Sites and create a new Site with public visibility. The name of the site must begin with the current academic year.
- The Site must include seven pages:
  - Introduction.
  - Task.
  - Process.
  - Resources.
  - Evaluation.
  - Conclusion.
  - Credits: This section must contain the name of the authors, the topic developed, related literature and licenses of the images used.
- Each page must be correctly identified with its name (Introduction, Task, etc.), which must be visible on both the title of the page and the menu of the Site.

- Each of these pages must be filled with relevant content that corresponds to the section or passage of the Webquest (see description in the presentation of the work). More precisely, each page must contain at least:
  1. A paragraph of text and an image.
  2. Links to previous and next page (if any).
  3. The Resources page must include an embedded video and provide enough content for the student to complete the Webquest. This content may be indicated by external links or may be included in the actual Resources page (this last option will be better appreciated in the correction).
- All images used must be Creative Commons, Public Domain, Copyleft, free or ownership of the authors. The licenses of these images must be mentioned in the "Credit" section including the author, type of license and link. If the images belong to the authors of the work, it must be clearly stated in this section.
- The appearance of the site is assessed: theme and colors used, header image, appearance of the contents, etc.

Appendix B.2. Rubric

1. Introduction page

1.1 Does the page include text, at least one image and a link to the next page? Rate from 0 to 3 (0 - does not include any element; 1 - includes some element, 2 - includes most of the elements, 3 - includes all elements).

1.2 Is the content of this page consistent with the "Introduction" of a Webquest? This section should only make a brief introduction to some information of the quest as well as motivating and arousing the interest of the reader/player. Rate from 0 to 3 (3 being totally correct).

1.3 (*) Quality of page. Rate from 0 to 3 (3 being the best) the content, appearance, and originality of this page.

2. Task Page

2.1 Does the page include text, at least one image, and link to the next page? Rate from 0 to 3 (0 - does not include any element; 1 - includes some element, 2 - includes most of the elements, 3 - includes all elements).

2.2 Is the content of this page consistent with the task description of a Webquest? This section should contain a formal description of the activity, indicating the contents to be further studied. Rate from 0 to 3 (3 being totally correct).

2.3 (*) Quality of page. Rate from 0 to 3 (3 being the best) the content, appearance, and originality of this page.

3. Process page

3.1 Does the page include text, at least one image, and links to the "Resources" and "Evaluation" pages? Rate from 0 to 3 (0 - does not include any element; 1 - includes some element, 2 - includes most of the elements, 3 - includes all elements).

3.2 Is the content of this page consistent with the process description of a Webquest? This section specifies the steps and exercises to develop the actual task. Rate from 0 to 3 (3 being totally correct).

3.3 (*) Quality of page. Rate from 0 to 3 (3 being the best) the content, appearance, and originality of this page.

4. Resources Page

4.1 Does the page include text, at least one image and an embedded video, and a link to the "Process" page? Rate from 0 to 3 (0 - does not include any element; 1 - includes some element, 2 - includes most of the elements, 3 - includes all elements).

4.2 Is the content of this page consistent with the resources page of a Webquest? This section must include all the necessary materials for the student to complete the exercises. Note that this should be the most complete section in the work. It is preferable, and thus rewarded with better scores, including the materials within the page instead of relying exclusively on external links. Rate from 0 to 3 (3 being totally correct).

4.3 (*) Quality of page. Rate from 0 to 3 (3 being the best) the content, appearance, and originality of this page.

5. Evaluation Page.

5.1 Does the page include text, at least one image, and a link to the next page? Rate from 0 to 3 (0 - does not include any element; 1 - includes some element, 2 - includes most of the elements, 3 - includes all elements).

5.2 Is the content of this page consistent with the evaluation page of a Webquest? This section must provide information on the evaluation system for rating each of the exercises. Rate from 0 to 3 (3 being totally correct).

5.3 (*) Quality of page. Rate from 0 to 3 (3 being the best) the content, appearance, and originality of this page.

6. Conclusion page

6.1 Does the page include text, and at least one image? Rate from 0 to 3 (0 - does not include any element; 1 - includes some element, 2 - includes most of the elements, 3 - includes all elements).

6.2 Is the content of this page consistent with the conclusion page of a Webquest? This section should include a brief conclusion on the key points learned in the Webquest and encourage students for further exploration and learning. Rate from 0 to 3 (3 being totally correct).

6.3 (*) Quality of page. Rate from 0 to 3 (3 being the best) the content, appearance, and originality of this page.

7. Credits page

7.1 Does the page clearly indicate the subject and the level of the Webquest? Rate from 0 to 3 (0 - does not include any element; 1 - includes some element, 2 - includes most of the elements, 3 - includes all items).

7.2 Does the page provide a list of licenses for each image used? Check that you can identify each image with its corresponding license, the author, and the link to the image. Note that the only correct licenses are the ones that allow the reuse of the image, such as Creative Commons, public domain, Copy Left or other free licenses. Rate from 0 to 3 (0 - Does not include any license; 1 - includes any license; 2 - includes most licenses; 3 - includes all licenses).

7.3 Does the page include a list of references to the literature considered? It should include citations to any books, websites, and sources of information used for the work. In case all the work is a genuine contribution of the authors, it should be clearly stated (0 - NOT included; 3 - YES, it is included).

8. Overall evaluation

8.1 (*) Is the chosen theme and content appropriate for primary school students? Rate from 0 to 3 (0 - NOT appropriate; 1 - questionable; 2 - appropriate; 3 - very appropriate).

8.2 (*) Score of the entire Webquest taking into account the content, appearance, and originality. Rate from 0 to 3 (being 3 the best score).

Acknowledgement

This work has been supported by the Vicerrectorado de Calidad e Innovación Educativa-Instituto de Ciencias de la Educación of the Universidad de Alicante (016-17 edition) through the Programa de Redes-I3CE de investigación en docencia universitaria (ref. 3690).

Highlights

- Open-ended works represents a significant teachers' workload with large groups
- We propose a novel methodology for open-ended works peer review
- Analysis with statistical tools is considered to detect possible biased scorings
- We tested the proposal with two different assignments with two groups of students
- The proposed methodology is statistically similar to that of the teachers' correction