An adaptive interface for computer-assisted rubrics in an e-Marking tool using nearest neighbor

Ignacio Cabrera  
School of Engineering and Science  
Adolfo Ibañez University  
Santiago, Chile  
ignacio.cabrera@uai.cl

Jorge Villalon  
School of Engineering and Science  
Adolfo Ibañez University  
Santiago, Chile  
jorge.villalon@uai.cl

Abstract—With an always growing number of student enrolment in higher education, providing quality feedback in both digital and paper based assessment becomes a heavy burden for teachers and tutors. The use of assessment rubrics can help overcome this burden, defining several criteria including formative feedback within it. However, current computer-assisted rubric systems present some drawbacks that hinder their adoption.

This paper presents the design of an adaptive interface for an e-Marking tool that uses rubrics, that is able to suggest assessors on the next criterion to be evaluated, based on nearest neighbor approach. An experimental setup showed encouraging results that provide evidence that the use of advanced learning technologies for assessment can help improve efficiency even in ill-defined domains.

e-marking; e-assessment; machine learning; evaluations; computer-assisted rubrics; rubrics

I. INTRODUCTION

Quality feedback is agreed among researchers as a critical aspect in the learning process particularly. According to Bloom’s taxonomy of learning [1], there are multiple levels of knowledge, with higher levels indicating a greater level of abstraction. In order to achieve such levels, one of the most important activities is assessment, which includes evaluation [2] and the feedback [3] teachers provide to their students. Both are expected to be personalized and delivered as soon as possible, because a timely efficient retribution of these two aspects will give students the chance to reflect on their own weaknesses [4] and therefore reach higher cognitive levels.

Teachers’ feedback can be formative or summative, and quality feedback must include both. Formative feedback corresponds to annotations and comments with the intention of provoking student reflection. Summative feedback corresponds to a quantitative measure of the overall quality of their work. Both are required to provoke students’ reflection.

However, an increasing number of student enrolments in higher education [5], make assessment a more costly task if one is to provide quality feedback. One of many concerns over this matter is: how to give all students an even and personalized education, considering the fact that the volume of assessment that teachers have to undertake is related to the number of students [2]. This burden is particularly heavy for open ended evaluations which include hand written answers such as free-text answers, hand drawn diagrams or handwritten solutions to mathematical problems.

Rubrics can help provide formative feedback for the assessment of open ended evaluations even for large numbers of students. Open ended evaluations require a certain criteria for assessment, usually because there is no single answer for a question, or answers require professional judgment to be evaluated [7]. “Rubrics are scoring guides that contain evaluation criteria and descriptions of different levels of achievement” [8], in the marking process with rubrics, the assessor identifies a criterion for each question, and depending on how well the student meets the professor’s requirements selects a particular achievement level. When used properly, rubrics can provide many advantages to the regular hand-marking process [9].

The use of computers to provide support on the assessment process using rubrics has been studied and several tools have been proposed, however they still present drawbacks that hinder the adoption of rubrics. Auvinen explains that even though several computer-assisted rubrics systems have been developed in quest for a timely efficient method to assess students and to enhance rubrics’ assessment potential, such systems present several drawbacks from a user-oriented interface which makes them hard to use [8]. Two major problems in such systems are: The separation of object and judgment, and visualizing complex rubrics. The first problem occurs when the visualization of students’ work (the object) and the marking using the rubric (the judgment) are in two separated interfaces, creating the constant need to review the rubric every time the assessor wants to mark and answer. The second problem occurs with complex rubrics, in which there are several criteria with several levels of achievement, making virtually impossible to visualize the whole rubric and the object at the same time. This issue makes the assessment process slower and in times, frustrating.

This paper proposes an intelligent tool for marking using rubrics that includes an integrated interface for rubric based marking and uses an adaptive algorithm to suggest the marker a criterion during the marking process. This interface would help assessors to speed up the marking process by making easier to locate marking criteria on large rubrics, allowing them to reduce time spent on marking and therefore, reducing the cost of assessing exams.

The tool was evaluated in terms of effectiveness, efficiency and reliability of the marking process by comparing two assessors in several assessment scenarios
using rubrics. Results showed promising results that open a whole field for technology enhanced learning tools.

II. PREVIOUS WORK

Computer Assisted Assessment (CAA) [10] has been extensively studied as a way to support the assessment process. CAA tools can be categorized into three major groups: online quiz systems, fully-automated and semi-automated [11].

Online quiz systems (OQS), allow users to create multiple choice questions which have restricted answers. These systems are commonly implemented through Learning Management Systems (LMS) [12]. OQS have been criticized because some say that this type of assessment would not develop the professional skills required for real world situations [11][13].

Fully-automated systems try to solve the whole assessment process by automatically marking students’ exams. This type of tools has been used on essay assessment [14] and faces many doubts from both, teachers and students, arguing that fully-automated assessment would not leave room for creativity in student’s answers. Because this marking process is based on an algorithm constructed from one of many possible solutions and for this reason, cannot have all possible solutions [11]. A second problem for these tools is that they only work on computer generated answers, and considering that paper based exams are still widely used on developing countries [12], it is possible to argue that this type of marking does not fit current evaluation systems’ needs.

Semi-automated marking tools relieve some of the tedious tasks on the assessment processes, but still require human supervision to mark the final grade, allowing hand written evaluations which are proposed as better media for students’ creativity on answers [11][8][7]. An important drawback for these tools is that they do not provide a solution for the need of generating personalized feedback for each student, which becomes a major problem for large courses [8].

It is precisely on this area where computer-assisted rubric systems have been proved to be a real advantage [9], supporting the attachment of personalized feedback and the reuse of common use feedback phrases, saving valuable time in the assessment process.

Auvinen [8] surveys computer-assisted rubric systems and describes their capabilities and limitations. Some of the existing systems are: Aloha, iAsocrates, iRubric, Rubistar, Waypoint Outcomes, GradeMark, Aropa, Trivedi and MarkTool. All of them give support to the assessment process by making fairly easy to submit assignments, create rubrics, assess different types of documents, mark exams, keep records, release results and calculate statistics of students’ performance [15].

Most of the available software requires licenses and all of them are closed sourced, making them impossible to modify with the purpose of adapting them to meet specific requirements.

An e-Marking tool has been recently developed in an attempt to solve both, the problem of the extensive use of paper based exams and the lack of an open source marking software [12]. In this tool the assessment is based on rubrics, but it does not require marking exams directly on it, instead, the assessor can “right click” over the digital copy of the student work and provide a mark using the rubric straight over it through a specific interface. Personalized feedback can also be provided on the digitalized work rather than just having the marked rubric, which helps students identify exactly where their mistake was and reinsures the fact that their work was reviewed completely. A note with the chosen rubric criteria and a customizable feedback phrase for the student to see is placed on the exact location on which the assessor right clicked.

Combining students’ work visualization and the marking process in one display, assessors can easily mark exams without the need of using multiple pages to display large rubrics.

Despite its advantages, the interface still present issues on the marking process due to the need to display the complete rubric, which in cases can have many levels, therefore it takes time for an assessor to get used to the interface and find the criteria that she wanted to use.

With the goal of solving these matters, an algorithm for the automatic suggestion of the next criterion to assess is proposed in the next section.

III. PREDICTION ALGORITHM

The tool’s current interface is shown in Figure 1, in which a diagram made by student is being marked by an assessor. The circle points out the position where the assessor right-clicked, which is used to show all criteria in the rubric in the same order they were originally created.

Once the assessor has right clicked over an exam to mark an answer, the position of her click (X and Y coordinates according to the interface’s width and height) and the page number (P) that she is clicking in, are stored with the selected rubric criterion and the identifier of the question that she is marking; providing information on what page and where in it a determined criterion was used, for each student.

Figure 1. Marking Interface without suggestion algorithm

With every new student that gets assessed, the number of (X,Y,P) positions for every criterion increases, providing the
opportunity to group these positions and predicting the next rubric criterion the assessor will want to mark.

The suggestion algorithm follows a nearest neighbor approach. On every right click the assessor makes over the exam has a given \((X, Y, P)\) position and it can be compared to previously stored positions. As the number of clusters is given by the number of criteria in the rubric, it is possible to calculate each criterion centroid \((X_c, Y_c, P_c)\) for each criterion on every page, and then, calculate the Euclidean distance \(d(x, y)\) of the current click to each centroid existing in that page.

\[
d(x, y) = \sqrt{(X - X_c)^2 + (Y - Y_c)^2}
\]

Using the calculated distances to each centroid, the rubric’s criteria can be sorted from the nearest to the farthest according to the assessor’s choice, and display them accordingly. Figure 2 shows two centroids \(Q_1\) and \(Q_2\) for their corresponding criteria, and as centroid \(Q_2\) is closer to the current click position, criterion 2 in the rubric is shown before criterion 1, as opposed to the normal sort order shown in Figure 1.

![Figure 2. Marking Interface with suggestion algorithm](image)

### IV. Evaluation

The proposed tool has three important aspects that must be evaluated which are required for every learning technology. It must be evaluated from the technological point of view (the accuracy of its predictive algorithm), the human-interaction point of view (how it helps humans fulfilling a particular task) and the learning point of view (the tool should not provoke less reliable assessment). These three aspects can be labeled as: Efficiency, effectiveness and reliability.

Efficiency is related to the work load required for an assessor to provide a mark and feedback for each student. It is measured in seconds and shows how fast an assessor can mark a piece of student work using the tool. In order for the tool to be useful, the inclusion of the algorithm should provoke a reduction on marking work load due to a reduction on marking time.

Effectiveness is related to how well the selected algorithm predicts the assessors’ intention for marking based on previous knowledge. This can be measured if the criterion selection made by the assessor was among the first choices presented by the interface when the algorithm is used. For example if the selected criterion was in the first position, it means that the piece of student work to be assessed, defined by the page number and the position within the page, was related. In case this relation results to be true, the designed interface and suggestion algorithm would be of great help while marking on large rubrics due to the display of the marking criteria that the assessors needs to mark on top of the interface, saving valuable time while looking for the right question to mark.

Reliability is a key aspect for learning technologies, as it provides evidence on the effect that the tool may provoke on the quality of the feedback provided to students. It is important for a tool like the one proposed in this paper to provide an improvement on efficiency without losing quality on the assessment. This aspect can be measured by comparing the grades resulting from the assessment process both with and without the use of the suggestion algorithm, and in which two assessors mark the same set of student work so inter-rater agreement can be calculated. For the tool to be reliable, inter-rater agreement should not have a significant difference.

### V. Experiment setup

The experiment included two assessors marking the final exams of a databases course ran on the second semester of 2012. The exam was proctored and all answers were digitized and uploaded into the e-Marking tool. It included three questions including the drawing of entity-relationship diagrams, hand-written SQL queries and a hand-written explanation on how a specific code should work. A total of 90 exams with 10 pages each were marked using a rubric with 10 criteria related to the questions and its subparts. For every question, there was a statement at the beginning of each page followed by blank pages for answers. Even though students were not explicitly told to answer on the corresponding pages, the rearranged disposition of the questions provoked students to do so.

The time taken to mark each exam by each assessor was recorded separately, and considering the fact that marking is a time and mind consuming task, break times were allowed between exams.

The experiment was performed in two stages with the intention of isolating the effect of the assessors getting used to a specific interface and becoming experts on using the rubric. The first stage presented the interface to the assessors with the algorithm activated randomly between exams, in this way assessors could not know if the interface was using the algorithm and therefore moving through the rubric accordingly. This allowed measuring efficiency gains from the new interface independently of the gains coming from the assessors gaining expertise on the tool. The second stage of the experiment presented the interface with the suggestion algorithm activated at all times. It was expected that at this stage both assessors should gain expertise on the tool and therefore improve efficiency further.
VI. RESULTS

During the first stage assessors were randomly assigned to the suggestion algorithm, and as both marked the same set of exams, all measurements correspond to dependent paired samples. Within these results, two sets can be identified depending on their random assignment of the algorithm: Firstly, a set in which each exam was marked with and without the algorithm, namely the “mixed set”. This set included 33 exams. Secondly, a set in which each exam was marked by both assessors using the same condition, which corresponds to the remaining 29 exams, namely the “non-mixed set”.

Average time taken for both assessors to mark the exams in the mixed set of samples is shown in Table 1. These results are encouraging as the inclusion of the suggestion algorithm improved marking times in almost 2 minutes, which is highly significant. A one factor ANOVA test with α=0.01 between marking times with and without the interface was used to compare means. The ANOVA test showed: \( F = 17.67; \quad p = 8.292^{-9} \) and \( F_{crit} = 7.048 \), showed that the results were also statistically significant when comparing marking times with or without the use of the suggestion algorithm.

<table>
<thead>
<tr>
<th>Stage 1 dependent paired samples</th>
<th>With Algorithm</th>
<th>Without Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Samples</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>Average [minutes]</td>
<td>4.87</td>
<td>6.84</td>
</tr>
<tr>
<td>Variance [minutes]</td>
<td>2.92</td>
<td>1.48</td>
</tr>
</tbody>
</table>

Table 1. Efficiency results: Average marking times with and without the suggestion algorithm.

As mentioned in the previous section, it can be argued that continuous training on the use of a particular interface may affect results. Therefore a second stage in which assessors used only the interface with the suggestion algorithm was performed. Table 2 shows that the average marking time for both assessors using the suggestion algorithm improves further on stage 2 from 4.87 minutes to 3.48, a difference that is also significant.

<table>
<thead>
<tr>
<th>Stages 1 and 2 independent samples</th>
<th>First Stage With Algorithm</th>
<th>Second Stage With Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Samples</td>
<td>33</td>
<td>38</td>
</tr>
<tr>
<td>Average [minutes]</td>
<td>4.87</td>
<td>3.48</td>
</tr>
<tr>
<td>Variance [minutes]</td>
<td>2.92</td>
<td>1.48</td>
</tr>
</tbody>
</table>

Table 2. Efficiency results: Average marking times using algorithm between the two stages.

Another ANOVA test for the two sets: “mixed set with algorithm” and “second stage”, showed the following results: \( F = 20.495; \quad p = 1.844^{-5} \) and \( F_{crit} = 6.928 \), supporting the initial assumption that there is a learning curve on the use of the interface with the algorithm that influence marking times.

Effectiveness is the second aspect to be evaluated, and corresponds to how well the algorithm predicts the assessors’ selection from the rubric. Table 3 shows the percentage of clicks made by the assessors on each criterion within the rubric. As it could be expected, without the use of the algorithm assessors’ selections were evenly distributed, as criteria was sorted by its appearance in the rubric, therefore each one should be selected once per exam (minor differences were caused by blank answers by some students). When the suggestion algorithm was used, assessors increasingly picked the first option offered by the algorithm. Moreover, on the second stage, when the interface included the suggestion algorithm at all times, assessors’ selection of the first option suggested increased further.

<table>
<thead>
<tr>
<th>Stages 1 and 2 Independent Samples</th>
<th>With Algorithm</th>
<th>Without Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interface Row</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stages 1 dependent paired samples</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Samples</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>Average [Grade]</td>
<td>38.46</td>
<td>37.03</td>
</tr>
<tr>
<td>Variance [Grade]</td>
<td>179.55</td>
<td>172.03</td>
</tr>
</tbody>
</table>

Table 3. Effectiveness of suggestion algorithm: Percentage of assessors’ selection on each rubric criterion.

Effectiveness is important while improving efficiency, improvements on effectiveness in every single criterion increases efficiency results, so that not only first criterion enhancements are important.

Even though effectiveness results are encouraging, there is space for improvement, as the distribution of selections still shows a uniform distribution for lower rows.

Reliability is the final aspect to analyze for this tool, as it is important to maintain a high reliability on the marking method. In order to measure reliability, the “mixed set” of dependent paired samples from stage 1 was used. Grades were calculated from 0 to 60 points.

<table>
<thead>
<tr>
<th>Stage 1 dependent paired samples</th>
<th>With Algorithm</th>
<th>Without Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Samples</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>Average [Grade]</td>
<td>38.46</td>
<td>37.03</td>
</tr>
<tr>
<td>Variance [Grade]</td>
<td>179.55</td>
<td>172.03</td>
</tr>
</tbody>
</table>

Table 4. Summary of grades for assessors A and B with and without the algorithm.
Table 4 shows there is no significant difference between grades awarded by assessors when using the algorithm. An ANOVA test was used to compare between grading with and without algorithm with the following results: $F = 0.191$; $p = 0.662$ and $F_{crit} = 7.048$, showing that the results are also statistically significant. These results reinforce the initial assumption that a suggestion algorithm does not affect the summative feedback.

VII. CONCLUSIONS AND FUTURE WORK

This paper presented an adaptive tool to support a computer-assisted rubric system within an e-Marking tool. Using a nearest neighbor approach, the rubric’s criteria were sorted so assessors could select from the first options offered by the interface. The suggestion algorithm used previous marks’ appearance information, particularly the page in which the mark was made and its position within the page.

Experimental results showed that intelligent systems can effectively improve the efficiency of the marking process, with encouraging results indicating that marking time could be cut up to 50% in the best case.

The algorithm’s efficacy was also measured, showing that even though important gains were made using a nearest neighbor approach, there is still space for improvement. Testing the interface using different predictive algorithms should be made in the near future.

Finally, as a learning technology, the tool’s reliability was evaluated so that efficiency gains did not affect the judgment outcome (summative feedback). Results showed that this was not the case.

In summary, this paper presents the design of an e-Marking tool with an enhanced support for computer-assisted rubrics to provide formative feedback, while at the same time is efficient to use. The combination of both e-Marking and enhanced rubrics allow to envision the use of open ended assessment on paper or digital form, that can receive quality feedback with a lower cost for teachers and tutors.

VIII. REFERENCES

[10] Bull, J Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom, Sheffield (United Kingdom). Staff Development Unit, Using technology to assess student learning, United Kingdom, 1993.