Open and Interdisciplinary
Journal of Technology,
Culture and Education

Special issue

From the Teaching machines to the Machine learning. Opportunities and challenges for Artificial Intelligence in education

Edited by
Valentina Grion
& Graziano Cecchinato

Publisher
Progedit, via De Cesare, 15
70122, Bari (Italy)
tel. 080.5230627
fax 080.5237648
info@progedit.com
www.progedit.com

qwerty.ckbg@gmail.com
www.ckbg.org/qwerty

Registrazione del Tribunale di Bari
n. 29 del 18/7/2005

ISSN 2240-2950

Index

Perusall: University learning-teaching innovation employing social annotation and machine learning

Graziano Cecchinato, Laura Carlotta Foschi***

DOI: 10.30557/QW000030

Abstract

This paper presents the learning-teaching innovation process of a University course. The traditional elements of the teaching-learning process (lecture, study, exam) involving students in ongoing activities have changed. The paper focuses on the learning changes introduced by social annotation activities carried out through the Perusall web environment. In particular, Perusall functionalities that assess students' participation were examined. These rely on multiple indicators set by the teacher, and a Machine Learning algorithm, which assesses the quality of annotations. A study was carried out to examine the validity of this process by analysing the relationship between Perusall algorithm's scores and teacher's scores, and how students perceive the automated scoring. The relationship was investigated through the Spearman correlation coefficient and Kendall's coefficient of concordance. Thematic analysis was used to analyse the qualitative data concerning students' perceptions. The results indicate that the Perus-

* University of Padua. Orcid: 0000-0003-3020-4525.
** University of Padua. Orcid: 0000-0001-7511-078X.
Corresponding author: graziano.cecchinato@unipd.it

all algorithm provided scores quite similar to those of the teacher, and that students positively perceived the automated scoring.

Keywords: Perusall; Social Annotation; Machine Learning; Peer Instruction; Learning-Teaching Innovation

Introduction

The innovation of the learning and teaching processes in higher education is an issue that concerns all countries (Armstrong, 2016; Brennan, Broek, Durazzi, Kamphuis, Ranga, & Ryan, 2014). Most prominent Universities have set up research centres on educational innovation[1] and spin-offs thereof to extend their educational offerings[2] with the aim to be more syntonic with strategies to build knowledge in the digital ecosystem.

What seems to be gradually losing its relevance is the lecture (Gibbs, 1982; Hattie, 2008; King, 1993) which registers a widespread and progressive attendance decrease (Kelly, 2012; Kottasz, 2005; Massingham & Herrington, 2006). The growing availability not only of videos but also of textual, multimedial, interactive and virtual digital educational resources makes it possible to share curriculum contents with students outside the classroom (Downes, 2007). According to the "flipped classroom" approach (Baker, 2000; Cecchinato, 2014; Lage, Platt & Treglia, 2000; Mazur, 1997), this shift creates conditions for spending more time in the classroom involving the students in active learning practices aimed at internalizing the contents (Bishop & Verleger, 2013; Jamaludin & Osman, 2014). This perspective entices many teachers because it has the potential to break down an "exam-oriented" habit that encourages rote learning, skipping the lessons or attending them without engaging with course work until a few days before exams (Berry, Cook, Hill, & Stevens, 2010; Burchfield & Sappington, 2000; Nonis & Hudson, 2006).

1. I.e.: www.cwsei.ubc.ca; globaled.gse.harvard.edu; tsl.mit.edu; www.ctl.ox.ac.uk
2. I.e.: www.coursera.org; www.edx.org

Changing this habit is not simple. Digital resources have the potential to make it possible for students to control the time, place and pace of the acquisition of knowledge, but this doesn't guarantee that they will take advantage of this control. Students need to realise that engagement in ongoing activities throughout a course could be really productive to learn useful knowledge, develop good skills and, accordingly, to pass exams. To accomplish this goal, redesigning the overall learning-teaching process is necessary. This requires not only changing the lectures, but also the students' way of studying and the assessment, mutually integrating them and actively involving the students (Biggs & Tang, 2011).

This change appears overwhelming for teachers, particularly in courses with a high number of students. One possible solution is to rely on peer learning processes. Research has highlighted the beneficial effects of these educational methodologies (Boud, Cohen, & Sampson, 1999; Stone, Cooper, & Cant, 2013; Topping, 2005).

Redesign learning-teaching process

Over the last three years, the authors of this contribution carried out a progressive redesign of two undergraduate courses at University of Padua, Italy, setting up peer learning activities. Pursuing the involvement of the students throughout the development of the courses, the course work, the lessons and the assessment practices have been transformed.

Using the Perusall[3] social annotation system, the study habits of students have changed from a pretty solitary experience to a social one (Miller, Lukoff, King, & Mazur, 2018). With Perusall, students can share their questions and replies with each other (and with the teacher) on the subject's topics in a quite easy way. This learning environment has been designed by its developers to anticipate the students' material needs for the following lessons in order to produce a

3. Perusall is a completely free web service at perusall.com.

deep analysis through multiple dialogues on the most controversial concepts. Useful communication tools make this process really productive in fostering participation and comprehension.

Advanced data reports give the teacher useful elements in order to analyse student participation. One of these has been designed to report the areas of most "confusion" for the students, identified by the most highly upvoted and least unanswered questions. Therefore, the teacher can prepare the next class activities specifically targeting the content that students are struggling with the most. These could be tackled through peer learning methodologies and, specifically, with the Peer-Instruction (Mazur, 1997) because the "conceptual questions" of this methodology could derive directly from the students' questions reported by the Perusall "Confusion report". So, part of the class time has been used to actively engage students through the Peer-Instruction methodology.

The changes in the learning-teaching cycle are completed with the integration of ongoing assessment activities. These activities are known to be generally demanding, in particular for teachers with large classes, but a solution could be peer- and self- assessment, which has proven to be very useful not only in reducing teachers' workloads, but mostly to improve learning (Black & Wiliam, 1998; Boud, 2000; Grion, Serbati, & Nicol, 2019; Liu & Carless, 2006; Nicol, 2010). Some digital learning environments make these processes simple, easy-to-manage and productive. One of these is Peergrade (https://www.peergrade.io/), that is specifically designed to enhance the formative dimension of peer- and self- assessment (Foschi & Cecchinato, 2019), promoting practices that derive from educational research on assessment (Cho & MacArthur, 2011; Nicol, 2010; Nicol, Thomson, & Breslin 2014)[4].

In other articles, the authors have presented the overall redesign of their courses (Cecchinato & Foschi, 2018), summarised here. The

4. To carry out these activities the authors have used Peergrade (peergrade.io). This environment promotes meaningful learning with an articulated peer- and self-assessment process.

use of Perusall, and specifically the validity of its grading system, is analysed in this paper.

Perusall

Perusall is a *social annotation environment* specifically designed for undergraduate courses (Miller et al., 2018). Its goal is to foster the comprehension of curriculum contents by involving students in a digital environment where they can share their issues, doubts and questions by helping each other. Its development is grounded in a social constructivist perspective, where knowledge is built by negotiating meanings and reflections through discussions (Vygotsky, 1980). With the development of e-learning, over the last three decades this perspective has been shared through online asynchronous discussion forums (Jonassen, 2008; Mazzolini & Maddison, 2003; Romero, López, Luna, & Ventura, 2013), that are available in all the main Learning Management Systems. Despite the development of different features that have improved the forums' functionalities over time, they still require a certain commitment by the students and often discussions are disorganised, compromising the development of actual conversational modes of learning (Thomas, 2002).

Social annotation systems like Perusall seem to foster a more generative pattern of interaction. The students' dialogues are anchored to specific sentences which students struggle with the most and are focused on promoting a better understanding of those concepts. Some suitable communications tools foster the students' participation by making interactions more productive with the aim to overcome misunderstandings and misconceptions. For instance, with one-click, students can upvote classmates' questions and replies. This avoids repetitions, improving participation and produces a sense of community (Rovai, 2002). Teachers can also upvote questions and replies of the students or pose their own questions and replies.

Perusall makes it easy and productive to carry out social annotation activities for teachers as well. To share materials with the students, teachers can load personal notes from their computers, scien-

tific open papers from the Internet and textbooks from the Perusall catalogue, where, thanks to agreements with the leading publishers, there is an electronic version of almost 70,000 textbooks. Teachers can easily set up assignments defining a range of pages, deadlines, a number of requested annotations, work notes and students' groups. They can contribute to the discussions by making comments and upvoting students' annotations[5].

Generally, a student's commitment to annotation is given a grade that counts towards the final course mark. Research has shown positive learning effects in evaluating the students' ongoing activities (Carless, 2015), but it is necessary to provide this incentive in order to stimulate open and fair participation toward improving learning and not to be an end in itself. Therefore, teachers normally give a low score for each annotation assignment. This can be done manually, by teachers, but also automatically, by the system.

The grading system

To promote the involvement of the students in annotation activities, Perusall provides a system of automated grading which gives a value to their participation. By collecting data from students' activities, Perusall can grade their participation taking into account six different components: The timeliness, quantity, quality, and distribution of the annotations; the amount of reading sessions; the complete reading of the material; the time spent in active reading; the obtained responses; and the upvotes given and received[6]. A dashboard provides the teachers with an advanced set of controls to adjust the grading system to the specific needs of their courses. They can define every single component in a very detailed way and weight them from 0 to 100. In any case, teachers can modify every single evaluation. Moreover, Perusall

5. For a thorough explanation of the functionalities of Perusall, please refer to Miller et al. (2018).

6. For more specific information on the grading algorithm's components, please visit: perusall.com/downloads/scoring-details.pdf.

can also be integrated into the main Learning Management Systems, where grades could be automatically imported.

Although grading students is not the core aim of Perusall, the grading system is very sophisticated and presents an element of real innovation. Five of the six components rely on learning analytics, that are now common in education, but the first one uses a Machine Learning (ML) algorithm to evaluate the quality of the annotations. Using Natural Language Processing (NLP), the algorithm analyses the text of the annotations and grades it following the teacher's settings. By default, an annotation is graded 0 if it is below expectations, 1 if it meets expectations, or 2 if it exceeds expectations. The algorithm has been trained to deal with text in some different languages[7]. The training consists of repeated processes when fine-tuning the algorithm's grade to reflect the teacher's grade on big amounts of annotations[8].

The study presented here aims to provide a contribution on the analysis of the validity of this grading algorithm.

Study

Validity of automated scoring

The validity of automated scoring has been primarily researched in the context of Automated Essay Scoring (AES), especially in the context of language certifications, *e.g.* Test of English as a Foreign Language (e.g., Attali & Burstein, 2006; Attali, 2007), and of standardised tests, *e.g.* Graduate Record Examination, and Graduate Management Admissions Test (e.g., Powers, Burstein, Chodorow, Fowles, & Kukich, 2002a, 2002b). The debate about the validity of automated scoring has been ongoing over time, suggesting that it is necessary to consider the different perspectives and inquiry available on this issue.

7. support.perusall.com
8. www.rug.nl

As early as 1951, Cureton, in the first edition of *Educational Measurement*, defined validity in terms of the relevance of the test to its intended uses. "The essential question of test validity is how well a test does the job it is employed to do. The same test may be used for several different purposes, and its validity may be high for one, moderate for another, and low for a third" (Cureton, 1951, p. 621). Also, more recent definitions of validity highlight the importance of the intended uses in evaluating it, as reflected in the latest Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), which define validity as "the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test" (p. 225).

By defining validity as just outlined, it is appropriate to first clarify the purposes and the uses of AES and, in light of the latter, to define the validation process. Several authors have conceived AES as a replacement for human scoring (Page, 1966; Shermis & Burstein, 2003), and their validation efforts have focused on trying to show that automated ratings are not distinguishable from human ratings. Others have conceived AES as a complement to human scoring, suggesting that using human ratings as the only criterion for judging the success of automated scoring is only a piece of evidence in the validity argument (Bennett & Bejar, 1998; Yang, Buckendahl, Juszkiewicz, & Bhola, 2002) and that the relationship with other measures should be considered. In our study, we do not intend to validate *tout court* Perusall's automated scoring algorithm, but to analyse whether our use of it, *i.e.* as an integral part of a more articulated student evaluation, may be valid. In this perspective, we aim to answer the following question: As an alternative method of scoring annotations, how similar are Perusall scores to teacher scores, and how do the stakeholders, *i.e.* students, perceive the automated scoring? The relation between automated and teacher scores provides important information about the validity of automated scoring. On the one hand, we can speculate that automated scores do not measure the same construct as teacher scores (see paragraph "Procedures"), but, on the other hand, automated scores are used as an alternative method of scoring students' annotations. Because we cannot claim that automated scores are completely inter-

changeable with teacher scores, it is important to estimate the degree of similarity between the two methods of scoring, as well as how students perceive the automated assessment.

Several types of statistics have been used to measure the relation between machine and human scores (see e.g., Attali & Burstein, 2006; Chodorow & Burstein, 2004; Powers et al., 2002a, 2002b). The easiest method is to calculate the percentage of agreement, but this does not take into account the expected agreement due to chance. Another method used, which corrects this shortcoming, is Cohen's *k*, but it requires that the scores be articulated into discrete categories. Therefore, as a measure of the relation between automated and human scores, the correlation coefficients are preferred. Specifically regarding Perusall, currently there seem to be no studies that have analysed the relationship between Perusall and teacher scores. In a study by Liberatore (2017), it is stated that "the machine scoring agrees very well with the professional judgment of the instructor", but the analyses on how this conclusion was reached are not specified. Also, as regards students' perceptions or opinions about the automated assessment process, there seem to be no studies. There are some studies that analysed the perceptions, reactions, and critiques relating, more generally, to Perusall, for instance, the students' experience using Perusall; or investigated aspects related to learning (e.g., Liberatore, 2017; Suhre, Winnips, de Boer, Valdivia, & Beldhuis, 2019; Sun & Smith, 2019).

Research questions

This exploratory study aimed to deepen the validity of the use we made of automated scoring by investigating the following two research questions:

RQ1. Did Perusall provide scores like those of the teacher on the annotations proposed by the students in the social annotation assignment analysed?

RQ2. What considerations do students express regarding the fact that an automated process has assessed their annotations?

Method

Participants

The study involved thirty students attending the University of Padua's "Psychopedagogy of New Media" course of the "Psychological sciences and techniques" three-year degree course in the 2018/2019 academic year. Perusall was used to develop three course modules. In two of these modules, the students expressed themselves in Italian, while in the third they expressed themselves in English. We conducted the analyses on the latter.

Procedures

Research question 1
To answer the first question, we investigated the relationship between the scores provided by Perusall and those expressed by the teacher (the first author) through an index of association between variables. We carried out the analyses on the scores assigned to 30 students. Each student posted an average of 3 annotations (min 1, max 7) (length of annotations: Min 26, max 389 words), for a total of 110. Both Perusall and the teacher scored each of these annotations.

The teacher scoring relied on criteria relevant for the learning process. Originality, appropriateness to the context, and insightfulness were considered for in-depth annotations; clarity and mastery of the content for question annotations; and correctness and capacity of clarification for replies annotations. Nevertheless, the overall scoring of each annotation was carried out without using a rubric but in a holistic way. The criteria by which Perusall algorithm scores the quality of the annotations are not published. However, Perusall reports that the average quality of annotations ranges from 0 (= does not demonstrate any thoughtful reading or interpretation), through 1 (= demonstrates reading, but no – or only superficial – interpretation of the reading), to 2 (= demonstrates thorough and thoughtful reading and insightful interpretation of the reading).

For each of the 30 students, we calculated the average of the scores provided by Perusall and the average of the scores expressed by the teacher. Given the limited sample size and the non-normality of the distributions (p-values of Shapiro-Wilk and Kolmogorov-Smirnov < .001), we measured the similarity between Perusall's and the teacher's average scores by Spearman correlation coefficient (*rho*). We also considered descriptive analyses, and we also calculated Kendall's coefficient of concordance (W) (Kendall & Babington Smith, 1939) to determine the inter-rater agreement between Perusall and the teacher on the scores assigned to the 110 annotations.

Research question 2

To answer the second question, we analysed the answers of 27 students to the following open-ended question: "What considerations do you want to express regarding the fact that your activity has been assessed by an automated Machine-Learning process?". Thematic analysis was used to analyse the qualitative data. The second author conducted the latter in a similar way to the process described by Braun and Clarke (2006). She developed "semantic" themes (ibid.) because we were mainly interested in what students explicitly wrote and not in identifying latent meanings, also considering that the answers to free-text survey questions often tend to be too "thin" to support deeper forms of analysis (LaDonna, Taylor, & Lingard, 2018). In analysing the answers, she proceeded with a circular and recursive process. At first, she familiarised herself with the data by reading and re-reading the students' answers, and she took notes on the overall ideas of the data. Then, she created initial codes using the students' own language. These codes were subsequently interpreted and grouped into potential overarching themes[9]. Finally, the themes were reviewed to find out patterns and to examine the more recurrent themes. Unlike the approach of Braun and Clarke, *i.e.* reflexive thematic analysis, we used an approach in-between a codebook approach and a coding reliability approach (see Braun & Clarke, 2019). To assess the reliability of the coding, the second author

9. In this phase, similarly to the process of template analysis (King & Brooks, 2017), a coding template was also developed. This was revised and refined during the analyses and then applied to the full data set.

sent the coding template, along with different students' answers, to the first author, who proceeded to code them. Inter-rater reliability was then calculated through Cohen's *k* (Cohen, 1960). Finally, the disagreements were resolved by mutual consultation and shared coding was reached.

Results

Research question 1

The calculated rho value was .58 ($p < .001$), which highlights that there is a statistically significant and moderately strong positive relationship, that indicates that Perusall provided quite similar scores to those expressed by the teacher.

Taking the descriptive statistics into consideration, we can note that the mean (1.87) and the median (2) of the scores provided by Perusall were, albeit slightly, greater than those expressed by the teacher (1.72; 1.75). The latter also exhibited greater variability (.1) than those expressed by Perusall (.05). Overall, Perusall provided higher scores than the teacher. In particular, the analysis of the single scores attributed to each annotation revealed 26 discrepancies out of 110 scores (23.64%). In 19 cases (17.27%), the scores of the algorithm were greater (difference of 1) than those of the teacher, while in 5 cases (4.55%) the opposite occurred. In only 2 cases (1.82%) was the discrepancy considerable: In one case, the teacher assessment was 0 while that of the algorithm was 2, while in the other case, the opposite occurred.

Moreover, Kendall's W calculated between Perusall and the teacher on the scores assigned to the 110 annotations was .65 (asymp. sig. < .05). As the interpretation of Kendall's W coefficient can be based on the Cohen's *k* estimation guidelines (Landis & Koch, 1977), the W value highlights a substantial level of agreement between Perusall's and the teacher's scores.

Research question 2

Several interesting themes arose from the thematic analysis of the qualitative data, providing insight into the students' considerations regarding the automated ML-based assessment. Here we focused on three

key dimensions found in the students' answers: Overall perception of the automated assessment, supervision, and assessment preferences.

The initial level of agreement between the two authors, measured by Cohen's *k*, regarding the first key dimension' coding was .83 ($p < .001$), which is considered almost perfect according to estimation guidelines (Landis & Koch, 1977). While for the other two key dimensions, the Cohen's *k* were .80 ($p < .001$) and .78 ($p < .001$), indicating a substantial level of agreement.

Key dimension 1: Overall perception of the automated assessment
The themes within this dimension show the different perceptions that students may have of the experience of having their annotations assessed by an automated ML process. Students showed three different perceptions as presented in Table 1.

Table 1. Different students' overall perceptions about the experience that an automated ML process has assessed their annotations

Theme	Description	No. of students	Quotations
Positive perception	Students describe the automated assessment process as interesting, effective and adequate.	14	S13: "It is a methodology that I found very interesting! [...]" S3: "I believe it is an effective assessment process [...]"
Neutral or hybrid perception	Students who show a neutral perception express neither negative nor positive thoughts about this type of assessment. Students who show a hybrid perception express how the automated assessment can have strengths, but, at the same time, some shortcomings or weaknesses.	7	S24: "I was not affected in a negative way by the presence of an automated correction system. [...]" S9: "In my opinion, it may be an excellent solution with regards to the simplification of the correction procedures, but it could present some gaps in another, more humane profile. [...]"
Negative perception	Students highlight that knowing that they would be assessed by an automated assessment process, at least initially, caused them a little "bitterness" or "sadness", or claim that, if not validated by the teacher, automated assessments would not always be adequate.	3	S25: "[...] The only thing that perplexes me is the fact that knowing to be assessed by a "machine" gives me, in a certain sense, a feeling of "sadness", as if my work was not "worthy" of the attention of one of my peers (intended as a human being) but it was only "one of many" that the machine must dispose of. [...]"

Key dimension 2: Supervision

The two themes within this dimension show the different positions that students may have about the assessments generated by an automated ML process. Students showed two different positions as presented in Table 2.

Table 2. Different students' positions about the assessments received by an automated ML process

Theme	Description	No. of students	Quotations
Need for supervision	Students feel the need for supervision especially in terms of verifying the validity of the assessments provided by the algorithm, but also because the teacher-human is able to grasp, enhance and evaluate individual differences and the nuances that come with being human*.	15	S1: "[...] it would be important to understand how much the assessments expressed by Perusall were actually correct and, therefore, how much they diverged from the human assessments. [...]" S3: "[...] obviously teacher supervision is necessary, guarantor of a qualitative assessment that cannot exclude a minimum of subjectivity, which is useful for giving value to certain aspects of learning. [...]"
No need for supervision	Students find the automated assessment process valid and reliable.	3	S26: "I believe that, since the assessments were given by a system, they are very objective. [...] in this way absolute impartiality is guaranteed. [...]"

* It is interesting to note that the specification of "teacher-human" has its *raison d'être*. Of the fifteen students that expressed a need for supervision, seven refer to the need for teacher supervision, seven to the need for human supervision and one to non-automated supervision. The expressive choice relating to "human" could suggest a man-machine comparison rather than referring to the distinctive traits of a teacher. It would be interesting to deepen these aspects with further studies.

Key dimension 3: Assessment preferences

The three themes within this dimension show the different preferences that students may have about the source of assessment. Students showed three different preferences as presented in Table 3.

Table 3. Different students' preferences regarding the source of assessment

Theme	Description	No. of students	Quotations
Preference for an integrated assessment system	Students believe that the two assessment systems, the automated one and the teacher-human one, must coexist in an integrated assessment system in which both these processes can contribute with their own contribution to the overall evaluation. Students consider this integrated assessment system more valid than the two methods applied individually.	11	S14: "[...] I believe that the two correction systems can coexist, in an integrated assessment system, overall more valid than the two methods applied individually."
Preference for the teacher-human assessment	Students state that, although the automated assessment process can make it efficient and impartial, they personally prefer a human "opinion".	4	S6: "[...] I personally, however, always prefer a more elaborate human opinion, which is also based on personal experiences [...]"
Preference for the automated assessment	Students consider the automated assessment objective, regardless of individual influences, impartial and verifiable.	3	S26: "I imagine that the assessment parameters are standard and therefore verifiable in each intervention, in this way absolute impartiality is guaranteed. That may not always be possible if it is carried out by a human assessor [...]"

Discussion

The results of the first research question indicate that the Perusall algorithm has provided quite similar scores to those of the teacher. This moderate result may also be partially due to language issues. The students are non-native English speakers and it is, therefore, possible that their language skills may be more influential in automated scores than in teacher scores. As regards AES, the literature highlights, for example, how specific groups with different linguistic and cultural backgrounds may have, on average, higher (or lower) automated

scores than human scores (see e.g., Bridgeman, Trapani, & Attali, 2012; Chodorow & Burstein, 2004).

The results of the second research question indicate that the vast majority of students expressed a positive overall perception about the experience of having their annotations assessed by an automated process, at the same time, they also expressed a need for teacher-human supervision. Finally, the vast majority of students believed that the two assessment systems must be integrated. Ultimately, we can claim that our use of Perusall's automated scoring can be considered valid, both because automated scores were like those of the teacher, and because students perceived automated scoring as adequate. As already mentioned, currently there seem to be no studies that have analysed the relationship between Perusall scores and teacher scores, and students' perceptions of this automated assessment process, at least as done here. Similarly, there seems to be limited research that explores the use of Perusall for psychology students. Most research results collected from students refers to natural sciences or engineering subjects (e.g., Lee & Yeong, 2018; Lepek & Coppens, 2016; Lepek & Radl, 2019; Liberatore, 2017; Pejcinovic, 2018), rather than from a social science perspective (e.g., Suhre et al., 2019; Sun & Smith, 2019).

Conclusion

The increasing need to innovate teaching in higher education is leading to the development of new educational technologies. Big Data and Artificial Intelligence are taking place in the learning-teaching processes, promising innovations but also posing new challenges to educational systems. ML algorithms are supporting an increasing amount of processes making it urgent to evaluate their effects on learning-teaching processes.

Our contribution aims to suggest one possible way of analysing systems that integrates these technologies on student assessments. We focused on Perusall because it has the potential to foster socio-constructivist practices and meaningful learning in higher education and schools. Its automated scoring system can promote the adoption of

active learning practices because it saves teachers considerable time that can be spent in more interactive processes with students. We analysed the ML annotations scoring to give an evaluation of its validity in the context and for the use that we made of it in a course.

Our results highlight that the algorithm has provided quite similar scores to those expressed by the teacher. The analyses of the annotations that got different scores seem to show that it could not take into consideration methodological aspects, like, for instance, appropriateness of times, places or modalities of the annotations. Some annotations with good content were given a high score by the algorithm but less by the teacher: This is because the same content had been presented before in other annotations or because there wasn't a clear link between the selected text and the annotation. In other cases, the teacher graded some concise but, at the same time, very incisive annotations higher than the algorithm. One possible reason could be that the grading system relies on NLP algorithms that can interpret the meaning of the text, but is unable to consider other educational dimensions. We point out, however, that no grade differences found in our analyses have produced differences in the overall assessment of any student. This is because we have set the grading system, following Perusall recommendations, giving the students more than one way to reach a high score or the full score as well. Finally, it is worth considering that automated grading is not an essential part of Perusall. It is only one element of extrinsic motivation, which should have little weight on the overall assessment of students. In any case, it is also useful because it gives teachers a way to identify students early on, who, for whatever reason, are struggling with the course and can offer them help.

We also investigated the students' considerations about being assessed by an automated process. The results indicate that generally, students have no problem being assessed by an automated process, especially if there is supervision by a human teacher, and this probably indicates the best way that we can use these technologies at the moment.

The overall functionalities of Perusall that promote peer learning make this environment potentially very useful, particularly with large classes and in the context of online education.

Acknowledgements

This paper, whilst being the result of intense collaboration between the two authors, has been written as follows: The paragraphs and the related subparagraphs "Context", "Perusall", and "Conclusion" are by Graziano Cecchinato; the paragraph "Study" and the related subparagraphs are by Laura Carlotta Foschi.

References

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

Armstrong, L. (2016). *Barriers to Innovation and Change in Higher Education*. TIAA-CREF Institute.

Attali, Y. & Burstein, J. (2006). Automated Essay Scoring With e-rater® Vol. 2. *Journal of Technology, Learning, and Assessment*, *4*(3).

Attali, Y. (2007). *Construct Validity of E-Rater® in Scoring TOEFL® Essays*. Educational Testing Service.

Baker, J. W. (2000). The "Classroom Flip": Using Web course management tools to become the guide by the side. In J. A. Chambers (Ed.), Selected papers from the 11th International Conference on College Teaching and Learning (pp. 9-17). Community College at Jacksonville.

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9-17. doi: 10.1111/j.1745-3992.1998.tb00631.x.

Berry, T., Cook, L., Hill, N., & Stevens, K. (2010). An exploratory analysis of textbook usage and study habits: Misperceptions and barriers to success. *College Teaching*, *59*(1), 31-39. doi: 10.1080/87567555.2010.509376.

Biggs, J., & Tang, C. (2011). *Teaching for Quality Learning at University.* McGraw-Hill.

Bishop, J. L., & Verleger, M. A. (2013). The flipped classroom: A survey of the research. In *Proceedings of the 120th ASEE Annual Conference & Exposition 2013.* American Society for Engineering Education (ASEE). Retrived from http://www.asee.org/file_server/papers/attachment/file/0003/3259/6219.pdf.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74. doi: 10.1080/0969595980050102.

Boud D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151-167. doi: 10.1080/713695728.

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education*, *24*(4), 413-426. doi: 10.1080/0260293990240405.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101. doi: 10.1191/1478088706qp063oa.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589-597. doi: 10.1080/2159676x.2019.1628806.

Brennan, J., Broek, S., Durazzi, N., Kamphuis, B., Ranga, M., & Ryan, S. (2014). S*tudy on Innovation in Higher Education: Final Report*. Publications Office of the European Union.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*(1), 27-40. doi: 10.1080/08957347.2012.635502.

Burchfield, C. M., & Sappington, J. (2000). Compliance with required reading assignments. *Teaching of Psychology*, *27*(1), 58–60.

Carless, D. (2015). *Excellence in University Assessment: Learning from Award-Winning Practice*. Routledge.

Cecchinato, G. & Foschi, L. C. (2018). Involving students in teaching: Analysis of an educational innovation pathway at University. *Form@re – Open Journal per la formazione in rete*, *18*(1), 97-110. doi: 10.13128/formare-22539.

Cecchinato, G. (2014). Flipped classroom: innovare la scuola con le tecnologie digitali. *Italian Journal of Educational Technology*, *22*(1), 11-20.

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology*, *103*(1), 73-84. doi: 10.1037/a0021950.

Chodorow, M., & Burstein, J. (2004). *Beyond Essay Length: Evaluation of E-Rater®'s Performance on TOEFL® Essays*. Educational Testing Service.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. doi: 10.1177/001316446002000104.

Cureton, E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 621-694). American Council on Education.

Downes, S. (2007). Models for sustainable open educational resources. *Interdisciplinary Journal of E-Learning and Learning Objects*, *3*(1), 29-44. doi: 10.28945/384.

Foschi, L. C. & Cecchinato, G. (2019). Validity and reliability of peer-grading in in-service teacher training. *Italian Journal of Educational Research, Special Issue*, 177-194. doi: 10.7346/SIRD-1S2019-P177.

Gibbs, G. (1982). *Twenty Terrible Reasons for Lecturing*. Occasional Paper No. 8. SCED.

Grion, V., Serbati, A., & Nicol, D. (2019). Technologies as assessment change agents. *Italian Journal of Educational Technology*, 27(1), 3-4. doi: 10.17471/2499-4324/1098.

Hattie, J. (2008). *Visible Learning: A Synthesis of over 800 Meta-Analyses Relating to Achievement*. Routledge.

Jamaludin, R., & Osman, S. Z. M. (2014). The use of a flipped classroom to enhance engagement and promote active learning. *Journal of Education and Practice*, *5*(2), 124-131.

Jonassen, D. H. (2008). Instructional design as design problem solving: An iterative process. *Educational Technology*, *48*(3) 21-26.

Kelly, G. E. (2012). Lecture attendance rates at university and related factors. *Journal of Further and Higher Education*, *36*(1), 17-40. doi: 10.1080/0309877x.2011.596196.

Kendall, M. G., & Babington Smith (1939). The Problem of m Rankings. *The Annals of Mathematical Statistics*, *10*(3), 275-287. doi: 10.1214/aoms/1177732186.

King, A. (1993). From sage on the stage to guide on the side. *College teaching*, *41*(1), 30-35.

King, N., & Brooks, J. M. (2017). *Template Analysis for Business and Management Students*. Sage.

Kottasz, R. (2005). Reasons for student non-attendance at lectures and tutorials: An analysis. *Investigations in University Teaching and Learning*, *2*(2), 5-16.

LaDonna, K. A., Taylor, T., & Lingard, L. (2018). Why open-ended survey questions are unlikely to support rigorous qualitative insights. *Academic Medicine*, *93*(3), 347-349. doi: 10.1097/acm.0000000000002088.

Lage, M. J., Platt, G. J., & Treglia, M. (2000). Inverting the classroom: A gateway to creating an inclusive learning environment. *Journal of Economic Education*, *31*(1), 30-43. doi: 10.2307/1183338.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. doi: 10.2307/2529310.

Lee, S. C., & Yeong, F. M. (2018). Fostering student engagement using online, collaborative reading assignments mediated by Perusall. *The Asia Pacific Scholar*, *3*(3), 46-48. doi: 10.29060/taps.2018-3-3/pv2000.

Lepek, D., & Coppens, M. O. (2016). Nature-Inspired chemical engineering: Course development in an emerging research area. In *Proceedings of 123th ASEE Annual Conference & Exposition 2016*. American Society for Engineering Education (ASEE). Retrived from https://www.asee.org/file_server/papers/attachment/file/0006/9794/ASEE2016-FinalPaper-v2.pdf.

Lepek, D., & Radl, S. (2019). Using digital tools for enhanced student learning and engagement in particle technology courses at Graz University of technology. In *Proceedings of the 46th SEFI Annual Conference 2018: Creativity, Innovation and Entrepreneurship for Engineering Education Excellence* (pp. 984-990). European Society for Engineering Education (SEFI).

Liberatore, M. W. (2017). Annotations and discussions of textbooks and papers using a web-based system (work in progress). In *Proceedings of the 2017 ASEE Annual Conference & Exposition*. American Society for Engineering Education (ASEE). Retrived from https://peer.asee.org/board-20-annotations-and-discussions-of-textbooks-and-papers-using-a-web-based-system-work-in-progress.pdf.

Liu, N., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279-290. doi: 10.1080/13562510600680582.

Massingham, P., & Herrington, T. (2006). Does attendance matter? An examination of student attitudes, participation, performance and attendance. *Journal of University Teaching & Learning Practice*, *3*(2), 82-103.

Mazur E. (1997). *Peer instruction: A user's manual.* Prentice Hall.

Mazzolini, M., & Maddison, S. (2003). Sage, guide or ghost? The effect of instructor intervention on student participation in online discussion forums. *Computers & Education*, *40*(3), 237-253. doi: 10.1016/s0360-1315(02)00129-x.

Miller, K., Lukoff, B., King, G., & Mazur, E. (2018). Use of a social annotation Platform for Pre-class reading assignments in a Flipped introductory Physics class. *Frontiers in Education*, *3*(8), 1-12. doi: 10.3389/feduc.2018.00008.

Nicol, D. (2010). From monologue to dialogue: Improving written feedback processes in mass higher education. *Assessment & Evaluation in Higher Education*, *35*(5), 501-517. doi: 10.1080/02602931003786559.

Nicol, D., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: A peer review perspective. *Assessment and Evaluation in Higher Education*, *39*(1), 102-122. doi: 10.1080/02602938.2013.795518.

Nonis, S. A., & Hudson, G. I. (2006). Academic performance of college students: Influence of time spent studying and working. *Journal of Education for Business*, *81*(3), 151-159. doi: 10.3200/joeb.81.3.151-159.

Page, E. (1966). The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, *47*(5), 238-243.

Pejcinovic, B. (2018). Teaching professional skills in microwave circuit design classes. In *Proceedings – 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1400-1404). IEEE.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002a). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research*, *26*(4), 407-425. doi: 10.2190/cx92-7wkv-n7wc-jl0a.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E., & Kukich, K. (2002b). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*(2), 103-134. doi: 10.1016/s0747-5632(01)00052-8.

Romero, C., López, M. I., Luna, J. M., & Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, *68*, 458-472. doi: 10.1016/j.compedu.2013.06.009.

Rovai, A. P. (2002). Building sense of community at a distance. *The International Review of Research in Open and Distributed Learning*, *3*(1), 1-16. doi: 10.19173/irrodl.v3i1.79.

Shermis, M. D., & Burstein, J. (2003). Introduction. In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective* (pp. 147-168). Erlbaum.

Stone, R., Cooper, S., & Cant, R. (2013). The value of peer learning in undergraduate nursing education: A systematic review. *ISRN Nursing*, *2013*, 1-10. doi: 10.1155/2013/930901.

Suhre, C., Winnips, K., de Boer, V., Valdivia, P., & Beldhuis, H. (2019). Students' experiences with the use of a social annotation tool to improve learning in flipped classrooms. In *Proceedings of the 5th International Conference on Higher Education Advance* (pp. 955-962). Editorial Universitat Politècnica de València.

Sun, S., & Smith, M. (2019). Perusall integration framework. In *Proceedings of EDULEARN19 11th International Conference on Education and New Learning Technologies* (pp. 3516-3524). IATED. Retrived from http://lib.uib.kz/edulearn19/files/papers/928.pdf.

Thomas, M. J. (2002). Learning within incoherent structures: The space of online discussion forums. *Journal of Computer Assisted Learning*, *18*(3), 351-366. doi: 10.1046/j.0266-4909.2002.03800.x.

Topping, K. J. (2005). Trends in peer learning. *Educational Psychology*, *25*(6), 631-645. doi:10.1080/01443410500345172.

Vygotsky, L. S. (1980). *Mind in society: The Development of Higher Psychological Processes*. Harvard university press.

Yang, Y., Buckendahl, C., Juszkiewicz, P., & Bhola, D. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education*, *15*, 391-412. doi: 10.1207/s15324818ame1504_04.