Self and Peer Assessment Strategies

Elias de Oliveira[1] and Marcos A. Spalenza[1]

[1]Programa de Pós-Graduação em Informática
Universidade Federal do Espírito Santo
Av Fernando Ferrari, 514, Goiabeiras Vitória, ES 29075-910

elias@lcad.inf.ufes.br

Abstract. *The task of evaluating students in class is a tough job to be carried out even in middle size classrooms without some sort of technologies support. Therefore, lecturer in some Brazilian institutions are adopting varieties forms of self-evaluation strategies to overcome this problem. One drawback is the difficulty of deciding on fair and accurate assessment final marks, which should also take into account the lecturer judgment evaluation. We propose in this work some strategies to deal with this problem by using the Moodle as the interface between the lecturer and our proposed system. The results have two folds. On the one hand, we show that our methods also help to detect responsible students more accurately. On the other, using our methods, we show which students need closer assistance because they are not able to recognize the right answer among some others of their classmates on the second time they are exposed to the same activity.*

1. Introduction

The effort to give the right feedback at the right time to all students while their learning process in class is a tough task to be achieved without some sort of help from modern intelligent processing mining for educational system [Romero et al. 2010]. Besides, there are some initiatives in the literature [Shiba and Sugawara 2014] to overcome the problem of self-assessment in class. Nonetheless, we still also believe that the lecturer/teacher/instructor has a key role to play when it comes to decide on the final fair marks [Wang et al. 2014], when reviewers fails to reach a consensus, or the students are practicing unfair grading. This why in our approach the lecturer's judgment is an important part of our students' peer assessments model.

Instead of giving the students the responsibility of their own individual evaluation, as it is done in some self-evaluation approaches [Nicol and Macfarlane-Dick 2007], our approach introduces the idea that each student *must* also evaluate all the others in class. Thereby each student will have the opportunity to assess what and how their classmates expressed themselves regarding with the same activity. It is important to say that the answers are anonymously shown to them. Therefore, each students does not know which text response s/he is reviewing in class.

In order to deal with all the grades given by each student in class to all the others classmates, we built a system which together with the Moodle[1], a very popular learning management system (LMS), manages all the red-tape work we need. Our system fetches

[1] http://www.moodle.org

all the submitted text answers by the student's class from the Moodle to our computer server. We format the text answers to a nice readable table of responses to be presented back to the students only the answers, without students' identifications, on an other activity created by the lecturer on the Moodle (see Table 2 in Section 4). Having this new activity online, the students now enter their grades for their classmates and including for {him, or her}-self. The rest of calculations are done internally by our system and both the grades for the first and the second activities are set back to the Moodle.

Our system is so far configured to provide the user with two type of evaluations. The first is that when the student's grades *ranking* are similar to that given by the lecturer. This shows that the student is able to assess the importance of some answers over others as quite well as the importance also expressed by the instructor. In this case the value give to grade is not as much important.

Another type of evaluation is that when the student are asked to find the *correct value for grades* given by someone experienced such as the lecturer. In this case the student receives greater grades as much as their own grades are closer to that given by the lecturer. In our opinion, this type of evaluation is harder then the previous one.

Our experiments showed that this type of assessment is also a good strategy to improve student motivation, stimulates students to reflect, discuss, and collaborate in their learning process [Bouzidi and Jaillet 2009, Wang et al. 2014]. We show in the experiments that some students were able to grasp quite well the overall lecturer criteria of evaluation, whereas some others are those who need greater help on the subject. These are usually the students who actually yielded poorer results on the first part of the activity, when from the first time they had to answer by producing the textual answer.

This work is organized as follows. We discuss, in Section 2, some related works show the *dis-* or similarities with the current work conducted in this paper. The architecture of our system is discussed in Section 3, which is responsible to extract the data from the learning management system, process it and send it back to the user interface. In Section 4, we describe how the experiments were performed and the results yielded, and the conclusions are presented in Section 5.

2. Related Works

One of the important part of any learning process is its evaluation [Perrenoud 1998] and the continuous refine the route of corretion along the way. Evaluations serve to the two sides of the process: to the learner and also to the lecturer. For the former to correct their learning procedures in order to achieve increasingly better results. For the later, to rearrange its strategies in order to improve their capacity of precisely helping the learner in their individual needs.

The problem the lecturers are facing these days is to keep pace with the great number of students in class that the educational environment is impinging, specially in the e-learning environments [Breslow et al. 2013]. Thereby some strategies to reduce the lecturer effort, without neglecting the quality, have been proposed in the literature. We discuss in this section only two types of these strategies, for the sake of brevity. The a) self-assessments [McMillan and Hearn 2008] and b) {blind, ou not} peer-assessments [De Grez et al. 2012], whereas we consider to be *traditional* that approach where the assessment is done only by the lecturers themselves. Furthermore, in this work we are considering both the self and peer assessment, based on given grades, used in conjunction with the lecturer grades assessment as the baseline [Boud 1995] for the final results.

Self Evaluation Strategies

In this strategy the students are asked to look into their own activities and, based on the course prior accorded criteria, engage in their own learning, learner responsibility, metacognitive skills and a dialogical, collaborative model of teaching and learning. In other words, this is a powerful way to bring up the student awareness of their active role in the learning process. According to [Boud 1995], self-assessment should include two main elements: a) *making decisions about the standards of performance* expected and then making b) *judgments about the quality of the performance in relation to these standards*. The main problem, besides involve students in both of these aspects, is how to introduced this practice without blowing up the workload the lecturer/instructors already have doing many other day-to-day things [Boud 1995].

Another side of this coin is that self-assessment is also a great tool to reduce the instructors' evaluation effort, as part of this job is handed over to the students themselves. Unfortunately the authors in [Tousignant and DesMarchais 2002, Falchikov and Boud 1989] works show that the students' perception of themselves is not as accurate as their actual performance. In the light of these studies, thus, we need to find a way to balance this strategy with some other approaches to produce the learning desired intent.

The [McMillan and Hearn 2008] claim that when

> *Correctly implemented, student self- assessment can promote intrinsic motivation, internally controlled effort, a mastery goal orientation, and more meaningful learning*

In their work, the authors pointed out a schematic to explain the meaning of self-assessment where three aspects are depicted: a) self-judgment [Zimmerman 2002]; b) learning targets and instructional correctives; c) self-monitoring. All these aspects are the important engine to improve the students' learning.

Peer Evaluation Strategies

Whereas in self-assessment strategies are inward students' journey of their activities, in a peer-assessment strategy they are now impelled to look into not only their own respectively activities, but into their classmates activities as well. Moreover, they can also be asked to express their reasonable ranking of grades among their classmates answers. By doing this kind of activities a great deal of metacognitive skills is worked out by the students.

Peer learning is actually part of our development from the earliest years of life and the centrality of the lecturer makes us lose sight of this. We need to transform our educational environment into a place where one can see the birth of critical thinkers, who can evaluate the pros and cons of different ideas, or point of views, and *etc.* [Spiller 2009].

The authors in [Wang et al. 2014] proposed a strategy to arbitrate what they called the *non-consensus*, *i.e.* when two or more students do not agree with a reasonable evaluation of a particular activity. They say that *non-consensus is a common challenge that makes the reliability of peer assessment a primary concern in practices*. The proposed solution is based on the use of *review deviation* and *radicalness* to identify non-consensus in peer assessment. One of the sugested strategy by the authors was to award that student who gives a score close to the mean value of its review group scores, whereas should be penalized who gives a score far away from the group's mean value.

[Shiba and Sugawara 2014] proposed a trust networks model to assess mutual evaluation students within groups which can be randomly arranged and rearranged during the academic semester. Their model tries successfully to identify irresponsible students whose submit disputable evaluations, degrading thus the effective and accurate final grades. These irresponsible students' evaluations are ignored by the trust networks and students marks are reviewed. Their results showed the effectiveness of the proposed approach, but also some limitations are pointed out in their simulations. Besides, their results showed that their method can actually help marking the individual students in a group of work. Nevertheless, we see as a problem the fact that it is asked to a student to grade each other responses within the same group. It is easy to know who is giving which grade to the other member of the group and, thus, they can all agree, in real-life situations, to give each other the highest grade.

Combining Self and Peer Evaluation Strategies

We claim that a combined approach where the lecturer is part of either self or peer assessment is capable of giving more stability to the results [Shiba and Sugawara 2014]. Most of the previous authors pointed out some problems to blindly accept both self and peer assessment [De Grez et al. 2012]. According to this hypothesis, in our approach, we will always use the lecturer assessments to guide the final results. In this work we will consider two strategies: a) the actual grade value similarity between the student and the lecturer at each activity and b) the ranking grade between the student and the lecturer at each activity. Nonetheless, its is possible to consider many other ways of combining the lecturer given grades with the students', for instance, the mean of students' grades can also be considered when evaluating the student individual assessment [Wang et al. 2014].

In order to implement both of these discussed strategies, we apply the Pearson and Spearman correlation [Bussab and Morettin 2013, Cap. 4]. The first when we want that the student learn to give the right value to the activity, whereas the second when we are satisfied enough with the order of the grade given by the student when compared to the order given by the lecturer.

We claim that the use of a combined approach is already a way to give to the students a fast feedback to how well they are doing in their learning process [Perrenoud 1998]. Besides, they now have the opportunity of analyze their own classmates responses and decide what is good or bad as an answer to the given activity.

3. The System Architecture

Data mining tools are nearly part of our everyday lives. Embedded in many modern system to help to acquire knowledge about the user to provide this user with better experience. In education area this still quite an exception, although some developments can be found in the literature and few others are actually applied in traditional classrooms [Romero et al. 2010, Florian et al. 2011, Oliveira et al. 2013]. A great opportunity is rising with the increase use of learning management system and the popularization of

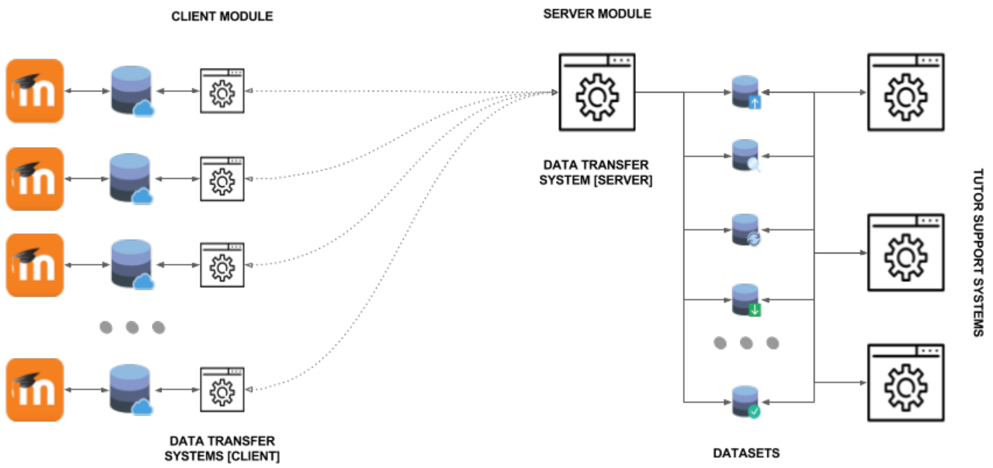

Figure 1. System architecture for LMS data extraction.

open and Mass Online Courses – MOOC's, where a huge amount of data is produced which can review old needs now possible to be quickly discovered by new methodologies [Siemens and Baker 2012].

Research in Educational Data Mining (EDM) [Romero et al. 2010] seeks to count for the data demands which come from LMSs either when in online learning, or in classroom. In order to meet the day-to-day production of data and to provide the suitable support to the instructors of the modules it is necessary to design some intelligent systems that act specifically to acquire enough information from data so that it can be able to act very similar to what would be done by the specialist of that class. Therefore, to this end, we extract the data from the learning platforms and take them to our local servers. This is done by the system module called *Tutor Support Systems* (TSS), ensuring that this transference meets some security requirements.

As an example of the amount of data produced by LMSs, we depict here the statistics yielded by the Moodle. According to its official website [2] there are 72 thousand systems registered in 232 countries around the world. Currently with 95 million users, Moodle is a great example of academic productivity for teaching professionals who individually easily assess all students. Therefore, by extracting the data from a system as such we can process it with modern data mining algorithm and turn back richer knowledge to the lecturer of a course provider of the data.

In order to orchestrate our proposed approach, a *Data Transfer System* (DTS) interacts directly with the LMS platform database to fetch the data uploaded by students. Using the LMS *web service*, the DTS connects clients and servers generating results on demand. Working with these tasks, the required TSS generates and sends report to all participants. The feedback are given by system using a upload methods of DTS. Like download, upload function uses transference services to insert data directly on learning platform. For this application cicle, a developed TSS can become a tailored service which can be activated by the lecturer as wish during the configuration of course assignments. The architecture is presented in Figure 1 in order to keep services in high availability for

[2] https://moodle.net/stats/

all LMS.

In the next section, we show some of the results of our strategy applied to some of our datasets and show that our approach is actually able to resolve the problem in the self- and peer-assessment approaches of being fairness and accurate.

4. Experiments and Results

In this section we present the experiments which was carried out to discuss the proposed approach we describe in this work. As we said in Section 2, we are going to focus here in only two out of many other ways of assessing the grades given by the students when self, or peer evaluating.

The data we use in our experiments were obtained from our own LMS[3]. In one of the activity we have in our LMS, the question was: *According to the author of our text book [Gil 2008], is the positivism a scientific method?*. Thus, we constructed a table with all the students responses, as shown in Table 2. The anonymous ID of the respondent is placed in the first column of this table. The second column, there is a room for the student annotate the intended grade, when on a printed sheet of paper, otherwise they can input their grades straightforward into the used LMS. In the last column, we have the actual text answer, given by the students. Again, this table is made available to the students for this activity within the LMS.

After having examined the list of answers in Table 2, students are then asked to give grade to each of the answer in the list of previous table. This is done in a new activity placed at the LMS by the lecturer. We specified a protocol for entering the values, as follows. Each line written by the student is initiated by a #, followed by the ID of the respondent, shown in Table 2. A semicolon is used to separate the ID and the given grade which follows the semicolon. For instance, four lines of a student peer assessment evaluation is depicted in Table 1:

A Student	
ID	Grade
#169-8700;	50.5
#172-8498;	45.00
#175-8500;	100
⋱	⋱
#544-8855;	0

Table 1. The grades given by a student into the LMS

The grades entering by the students are then paired with that assigned by the lecturer, and the specified correlation for that activity is calculated to be transformed into a new grade to each respective student for the peer assessments.

One possible configuration is to set the activity to transform into grade the value resulted from the *Pearson correlation* between the student grades and the lecturer given grades for the same answers. The greater the correlation with the lecturer grades, the greater the grade this student will be assigned. Another possibility is, alternatively, to set

[3] http://rii.lcad.inf.ufes.br/moodle/

ID	Grade	Textual Answer
169-8700		Sim Segundo Gil é uma concepção científica do saber
172-8498		Não porém para algo ser objeto de questionamento do positivismo precisa atender a muitas características do método cientifico
175-8500		Não o positivismo é uma concepção do saber uma corrente filosófica que surgiu na França no começo do século XIX
⋱	⋱	⋱
448-8502		não porque é uma doutrina filosófica
468-8897		Sim segundo o autor o positivismo é um método científico
483-8802		Sim Porque ele não é empirico ou seja passado de boca a boca atraves dos tempos ele o Positiviso é de cunho academico
544-8855		Sim segundo o autor o positivismo é um método científico

Table 2. Answers to be given back to students for evaluation and grading

the *Spearman correlation* for the activity, in this case the student are asked to correctly rank the grades among the answers. We also provide the user with the possibility of specifying an expression to combine some correlation formulas to produce the final grade for the activity.

Figure 4 shows the correlation results of all students grades together with the lecturer grades, called here as *baseline*. On the top side of the matrix, we have the heat map of these results, whereas on the bottom diagonal side the actual value of the total correlation between the student grades given to each of the responses and the lecturer assigned grades.

In this Figure 4, we can see that the student *X440* is the closest correlated to the lecturer on assigning grades for the all answers. Its Pearson correlation is of 0.74. We can see this on the third line, the *Baseline* – *i.e.*, the lecturer grades – crossed with the second column, the student *X440* column. The second closest correlated student to the lecturer, with correlation of 0.63 is the student *X437*, followed by the student *X323*, with 0.62. Some of the students did not respond the activity: *X317, X441, . . ., X544*, and *X438*. Others are totally uncorrelated with the lecturer way of assigning grades: *X363, X436, . . ., X292*, and *X169*.

One of the common problem when dealing with mutual assessment on online LMS, is the possibility of the students copy somebody else's evaluation. This can be detected by the high level of correlation between the students grading. We noticed that this might have happened in our experiments as shown in Figure 4 with the pair of students: {*(X169,X292); (X320,X294)*}, which yielded surprisingly high level of correlations, both correlation of 1, whereas very poor correlation with the lecturer given grades.

Based on the map of Figure 4, we can now easily grade the students on their correlation achievements. One approach is to assign the correlation straightforward as the

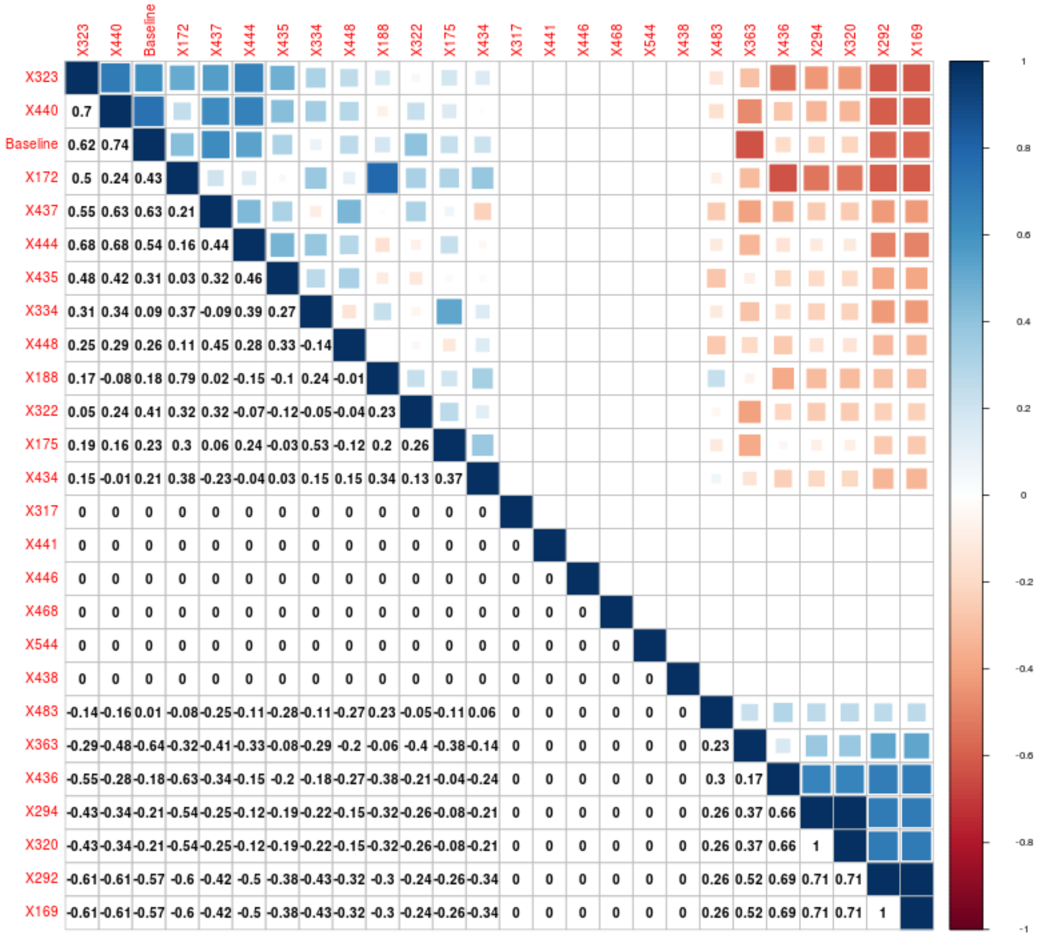

Figure 2. The correlation between the students and the *baseline* grades

grade, adjusting only the correlation value to the range of the LMS grade interval. For those who did not respond, or are uncorrelated to the lecturer, grading them to zero is a possible approach. Besides, the lecturer can now act in order to clarify those situations where pair of students have a high correlation between them on one or more activities.

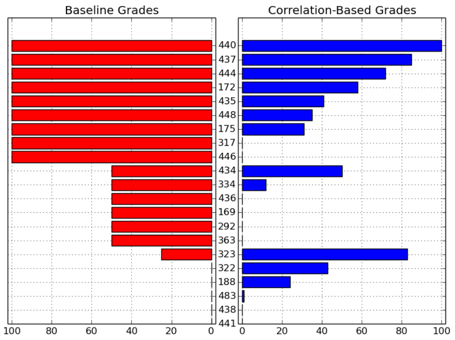

Figure 3. The performance curve of the students on the first and second activities

In addition to helping the lecturers to motivate even more the students in their own learning process, we can extract from this tool the information when the students is actually learning or needs more attention to the topic discussed in class on the subject regards to the evaluated activity. In Figure 4, in the *Baseline Grades* column, on the left side, we show the performance of the students on their given grades in the first activity. In the *Correlated-Based Grades* column, on the right side, the grades received by carrying out the peer evaluation. This figure shows that the majority high grades on the left, also received high grades on the right. Although is still premature to claim this in our experiments, but we noticed that some good students have more chance to receive better grades in the second activity. This was pointed out by [Shiba and Sugawara 2014], when referring to the *responsible* students. On the other hand, this figure also shows the students who need better attention as they were not able to guess, for the second time, what is good or bad answer in the activity.

5. Conclusions

Everybody who teaches have to spend lots of time on inspecting and assessing the activities carried out by every student in order to show their learning progress [Wang et al. 2014]. In this paper we discussed two stategies for weighting students' grades given by themselves to their respectively classmates' open text answers. Our methods are also good strategies to assess the student's cognition when all the answers in class are presented to them. By our proposed method, each student has to assess its own activity, and also give grades to the other answers submitted by their classmates.

In addition, our experiments showed that the implemented tool eased the performance of the automatic mutual assessment, which otherwise this process would be nearly impracticable. Therefore, these tools are ready to use and help lecturers to mediate the students to a better learning. The approach proposed in this work open up the possibility to build a set of reports where we can point out the lack of cognition expressed by some students on certain topics of the learning process.

In future work, we are planning to consider to include different forms of selection of the classmates' answers to send some challenger and more specific set of open text answers to each student for being evaluated. We intuitively observed that it is not necessary to pass all the classmates' answers to actually identify the students' learning needs. Furthermore, we are also planning to improve our architecture to allow the students not only be able to grade one activity at a time, but a set of activities at once.

We believe that student self-assessment, defined as a dynamic process in which students self-monitor, self-evaluate, and identify correctives to learn, is a critical skill that enhances student motivation and achievement.

References

Boud, D. (1995). *Enhancing Learning Through Self Assessment*. Kogan Page.

Bouzidi, L. and Jaillet, A. (2009). Can Online Peer Assessment Be Trusted? *Journal of Educational Technology & Society*, 12(4):257.

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., and Seaton, D. T. (2013). Studying Learning in the Worldwide Classroom: Research into EdX's First MOOC. *Research & Practice in Assessment*, 8.

Bussab, W. O. and Morettin, P. A. (2013). *Estatística Básica*. Saraiva, São Paulo, 8 edition.

De Grez, L., Valcke, M., and Roozen, I. (2012). How Effective are Self-and Peer Assessment of Oral Presentation Skills Compared with Teachers' Assessments? *Active Learning in Higher Education*, 13(2):129–142.

Falchikov, N. and Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of Educational Research*, 59(4):395–430.

Florian, B., Glahn, C., Drachsler, H., Specht, M., and Gesa, R. F. (2011). Activity-Based Learner-Models for Learner Monitoring and Recommendations in Moodle. In *European Conference on Technology Enhanced Learning*, pages 111–124. Springer.

Gil, A. C. (2008). *Métodos e Técnicas de Pesquisa Social*. Editora ATLAS, São Paulo, 6ª edition.

McMillan, J. H. and Hearn, J. (2008). Student Self-Assessment: The Key to Stronger Student Motivation and Higher Achievement. *Educational Horizons*, 87(1):40–49.

Nicol, D. J. and Macfarlane-Dick, D. (2007). Formative Assessment and Self-Regulated Learning: a Model and Seven Principles of Good Feedback Practice. *Studies in Higher Education*, 31(2):199–218.

Oliveira, M. G., Ciarelli, P. M., and Oliveira, E. (2013). Recommendation of Programming Activities by Multi-Label Classification for a Formative Assessment of Students. *Expert System and Applications*, 40:6641 – 6651.

Perrenoud, P. (1998). *Avaliação: Da Excelência à Regulação das Aprendizagens – Entre Duas Lógicas*. Artmed Editora, Porto Alegre, RS.

Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. J. D., editors (2010). *Handbook of Educational Data Mining*. CRC Press, Boca Raton, FL.

Shiba, Y. and Sugawara, T. (2014). Fair Assessment of Group Work by Mutual Evaluation Based on Trust Network. In *2014 IEEE Frontiers in Education Conference (FIE) Proceedings*, pages 1–7.

Siemens, G. and Baker, R. S. J. D. (2012). Learning Analytics and Educational Data Mining: Towards Communication and Collaboration. In *Proceedings of the 2Nd International Conference on Learning Analytics and Knowledge*, LAK '12, pages 252–254, New York, NY, USA. ACM.

Spiller, D. (2009). Assessment Matters: Self-assessment and Peer Assessment. Technical report, Universidade de Waikato, Nova Zelândia.

Tousignant, M. and DesMarchais, J. (2002). Accuracy of Student Self-Assessment Ability Compared to Their Own Performance in a Problem-Based Learning Medical Program: a Correlation Study. *Advances in Health Sciences Education*, 7(1):19–27.

Wang, Y., Liang, Y., Liu, L., and Liu, Y. (2014). A Motivation Model of Peer Assessment in Programming Language Learning. *Computing Research Repository – arXiv.org*, abs/1401.6113.

Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory into Practice*, 41(2):64–70.