Assessment analytics for peer-assessment: a model and implementation

Blaženka Divjak
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2
42000 Varaždin, Croatia
bdivjak@foi.hr

Darko Grabar
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2
42000 Varaždin, Croatia
darko.grabar@foi.hr

Marcel Maretić
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 2
42000 Varaždin, Croatia
mmaretic@foi.hr

ABSTRACT

Learning analytics should go beyond data analysis and include approaches and algorithms that are meaningful for learner performance and that can be interpreted by teacher and related to learning outcomes. Assessment analytics has been lagging behind other research in learning analytics. This holds true especially for peer-assessment analytics.

In this paper we present a mathematical model for peer-assessment based on the use of scoring rubrics for criteria-based assessment. We propose methods for the calculation of the final grade along with reliability measurues of peer-assessment. Modeling is motivated and driven by the identified peer-assessment scenarios.

Use of peer-assessment based on a sound model provides benefits of the deeper learning while addressing the issues of validity and reliability.

Categories and Subject Descriptors

D.2.8 [**Software Engineering**]: Metrics—*complexity measures, performance measures*; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Design, Measurement, Reliability

Keywords

peer-assessment, assessment, analytic tools for learners, assessment learning analytics

1. BACKGROUND ON ASSESSMENT LEARNING ANALYTICS

Learning analytics (LA) is all about usefulness of the data once they have been collected and analyzed [6]. Research in LA is interdisciplinary and it must be emphasized that LA includes the aspects of human judgments and it goes beyond data analysis (business analytics): it has to make sense of information, come to decisions and take action based on data [13]. This is the leitmotiv of the research presented in this paper.

LA has to be useful to a vast majority of students. The so-called average student has to be taken into account when setting the goals of LA, not only the under-performing or over-performing students. Teaching practice shows that a meaningful analysis of assessment results is of interest to all the students.

Assessment is both ubiquitous and very meaningful as far as students and teachers are concerned (Ellis in [6]). It is an essential part of the teaching and learning process, especially in the formal education because assessment guides learning for a vast majority of students. Ellis at the same time claims that assessment analytics are lagging behind other types of learning analytics. There are several reasons for this. Among these, we argue that insufficient granularity of assessment data presents a difficulty for an interpretation of results.

The so called *networked learning* (see [12], e.g. Massive Open Online Courses (MOOCs), social learning platforms, online learning and e-learning in general) presents a completely new playground for learning analytics. In networked learning the number of participants rapidly increases along with the interactions between learners in the form of discussions and mutual learning. We focus here on a special types of assessment: peer-assessment. Use of peer-assessment and self-assessment is appealing and very appropriate for a task leading to a certificate in a MOOC with enrollment measured in tens of thousands. This approach generates a huge amount of assessment data but also asks for sound metrics for the calculation of final grade and for estimates on the reliability of assessment data. Peer-assessment has additional benefits in the learning process, but also additional disadvantages (cf. [4]). Among the disadvantages there are issues of reliability and validity of assessment.

To address validity, we advise the use of the scoring rubrics as they contribute to the quality of assessments and by facilitating valid judgments of complex competencies [10]. Based on the analysis of 75 studies Jonsson and Svigby

in [10] conclude that the use of scoring rubrics enhances the reliability of assessments, especially if the rubrics are analytic, topic-specific, and complemented with examples and/or rater training. Otherwise, the scoring rubrics do not facilitate valid judgment of performance assessments. Besides this, rubrics have a potential to promote learning and/or improve instruction.

Aim of this paper is to model peer-assessment and to discuss issues of final grade calculation and reliability of raters' judgments. Jonsson and Svingby note that variations in raters' judgments can occur either across raters, known as *inter-rater reliability*, or in the consistency of one single rater, called *intra-rater reliability*. Referring to [1] Jonsson and Svingby state that *a major threat to reliability is the lack of consistency of an individual grader*. Reports rarely mention this measure. On the other hand, *inter-rater reliability* is in some form mentioned in more than half of the reports but many of these simply use *percentage* as a measure for agreement. This is in agreement with Sadler and Good's critique in [14] of poor quality of quantitative research regarding self-assessment. Situation has improved since. Nevertheless, majority of current research still uses overly simple statistical measures in order to determine correlations that might indicate reliability.

In the following sections we describe two major peer-assessment scenarios we have recognized and for which we have developed a mathematical model. After that we present and analyze a model for these scenarios.

2. SCENARIOS FOR PEER-ASSESSMENT

Reliability of peer-assessment depends on many factors but consistency of individual evaluator was very early recognized as the most important (see [1]). On the other hand, having more assessments per assignment increases the reliability of peer-assessment with relatively inexperienced evaluators.

From *experienced evaluators* (experts) we presume a high expertize in the domain knowledge and prior experience in evaluation. Similarly, an inexperienced evaluator is an individual with a relatively high level of domain knowledge (high baseline), but lacking experience in evaluation (e.g. peer assessment by senior undergraduates).

We analyze scenarios with respect to the experience of evaluator as is shown in scenario grid (Fig. 1). We have placed a continuum of possible scenarios in a grid with four quadrants. Within four quadrants we recognize two interesting scenarios for peer-assessment and discard the other two as either unrealistic or inappropriate.

In the first scenario, let us call it Scenario A, participants are inexperienced evaluators (for example undergraduate students with introductory domain knowledge and no experience in peer-assessment) whereas in the scenario B evaluators have higher expertize in the evaluated domain (i.e. teachers, graduate students or senior undergraduates) and prior training in assessment. In scenario A, the lack of experience in evaluation must be compensated with a quantity of peer-assessments, i.e. having a larger group size in peer-assessment. On the other hand, setting a group size too large in scenario B is a needless waste of expert's time.

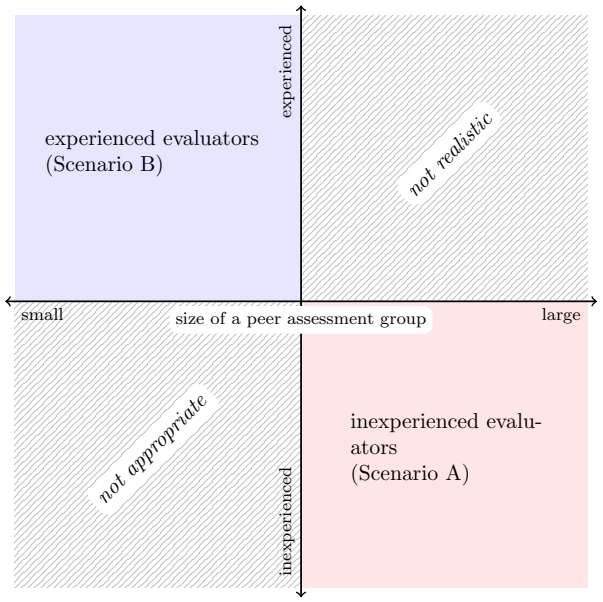

Figure 1: Scenario grid

Detailed analysis is given in the Table 1.

3. OVERVIEW OF THE PEER-ASSESSMENT ACTIVITY

Peer-assessment activity starts after the work on the assignment task has completed. In a general case peer-assessment consists of two phases. We identify following activities in the whole process.

Phase 1: *Assessment of assignments*

i. Learners assess a (predefined) number of assigned assignments

ii. Analysis of peer-assessments (grouped by assignment)

iii. Calculation of the assignment grade

Phase 2: *Assessment of the assessments*

i. Analysis of peer-assessments (grouped by grader)

ii. Calculation of the assessment grade

First phase starts with learners assessing the assignment work of their peers. We assume that each participant grades several assignments (at least 2). At the end of the first phase a reliability check has to be performed and the final grade has to be calculated. Second phase is concerned with the quality of assessments relative to the evaluator. As on outcome of the second phase graders can receive a grade (points) for the quality of their assessments.

Table 1: Scenario table

	Scenario A	Scenario B
Playground – use cases	Networked learning (MOOCs, online learning and e-learning in general, see [12]) Voting for awards where general audience is involved	Multiple graduate/postgraduate assess complex student work [3] Peer assessment of research papers Evaluation of competitive research projects
Evaluators' characteristics	A considerable number of relatively inexperienced evaluators in the area they assess	A few experienced evaluators that are experts in the area they assess
Resources to rely on	Inexpensive evaluators workload in almost unlimited quantities	Expertize of evaluators and their judgment that can be trusted
Reliability thread	Intra-rater and inter-rater inconsistency	Experts don't have equal expertize in all evaluation criteria
Strategy to increase reliability	Quantity of assessment that might be convergent (statistically speaking)	Quality of small number of assessments without outliers

4. MATHEMATICAL MODEL FOR PEER-ASSESSMENT

We recognized three challenges: (1) calculation of the final grade based on different assessment scenarios, (2) measurement of the assessment's reliability and (3) measurement of reliability of each grader (for grading of the graders).

4.1 Overview of the assignment grading

A grading G from the scoring rubric with n criteria is a tuple of numbers $G = (g_1, \ldots, g_n)$. We consider gradings as points in an n-dimensional space endowed with a metric d, i.e. a function that measures the distance between points (i.e. gradings) and satisfies the axioms of a metric space.

In [5] we proposed the use of the non-euclidean taxicab metric d_1, but for the purpose of this paper it is sufficient think of d as any distance metric.

4.2 Calculation of the assignment's final grade

An assignment graded through peer-assessment will receive several peer gradings. These will have to be analyzed. If estimated as reliable these gradings will be use as input for the calculation of the final grade.

A simplest approach is to calculate the final grade of assignment as the mean value of received assessments.

Let $\mathcal{S} = \{S_k^1, \ldots, S_k^m\}$ denote a set of peer gradings for assignment k, then the *mean grade* is

$$M(\mathcal{S}) = (a_1^f, \ldots, a_n^f), \quad \text{where} \quad a_i^f = \frac{1}{m}\left(\sum_{j=1}^{m} c_{k,i}^{(j)}\right) .$$

$M(\mathcal{S})$ is a center of mass of the set $\mathcal{S}$. This method for grade calculation is suitable for scenario A. We can say that $M(\mathcal{S})$ is sensitive to quantity, and less sensitive to outliers (it "*respects the decision of the majority*").

For scenario B, we propose an alternative grade calculation method (see [5]). In scenario B we assume that peers are experienced evaluators. Final grade is calculated as so-called *optimal final grade* $O(\mathcal{S})$ defined by

$$O(\mathcal{S}) = (o_1^f, \ldots, o_n^f), \quad \text{where} \quad o_i^f = \frac{1}{2}\left(W(\mathcal{S}) + B(\mathcal{S})\right) .$$

$W(S)$ and $B(S)$ represent amalgamations of *worst* and *best* received gradings respectively, defined by:

$$W(\mathcal{S}) = (w_1, \ldots, w_n), \qquad w_i = \min_j c_{k,i}^{(j)}$$

$$B(\mathcal{S}) = (b_1, \ldots, b_n), \qquad b_i = \max_j c_{k,i}^{(j)} .$$

This approach is inspired by Hwang and Yoon's TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method of multi-criteria decision making in [9]. When evaluators are trusted experts, we don't expect "wild" gradings (outliers). Here, it is expected that after just a few initial evaluations any additional gradings will have no effect on the final grade $O(\mathcal{S})$. Please consult [5] for additional details.

A summary of our recommendations for two scenarios A and B is given in the Table 2.

Table 2: Grading method recommendations

	Scenario A	Scenario B
Suggested grading method	Mean value grading. Reliability provided by quantity of evaluations.	Optimal value grading. Reliability provided by the quality of evaluators.

With optimal value grading we have the opportunity to allow experts to skip grading for certain criteria. For example this would be reasonable if an expert is not an expert for all the criteria. To be able to calculate $O(\mathcal{S})$ it is sufficient to have every criteria covered by at least one expert.

4.3 Reliability of the peer-assessment

A prerequisite for the calculation of the assignment's final grade is the determination whether a received set of peer-

assessments is (sufficiently) reliable, i.e. acceptable.

For reasoning about reliability it is necessary to have granular data. The importance of granular scoring data is illustrated in the example in Table 3. Gradings S_1 and S_2 agree on the summative level, but seem very distinct at the granular level. This is an example of an unreliable peer-grading set where this incoherence is not visible on the summative level.

Table 3: Highlighting the importance of granular data

	C_1	C_2	C_3	C_4	$\sum$	
S_1	3	0	2	2	7	summative
S_2	0	1	3	3	7	
	granular					

A **diameter** of a set of gradings $\mathcal{S} = \{S_1, \ldots, S_n\}$ is defined as

$$\texttt{diam}\,\mathcal{S} = \max_{i,j} d(S_i, S_j) \ .$$

We consider a set $\mathcal{S}$ of peer gradings as **reliable** if $\operatorname{diam}\mathcal{S}$ (maximal pairwise distance between gradings) is less than $2e$ where e is *acceptable error* given in advance.

Note that the diameter of the set $\mathcal{S}$ is also a diameter of an encompassing sphere. So, we can say that a reliable peer-grading set fits within an encompassing e-sphere.

If a set of peer-assessments is estimated as not acceptable (un-reliable) on the granulated level then the final grade cannot be calculated. A recommendation about acceptability of particular peer-assessment set can be given to teacher or course designer by LA. This can be implemented in the learning management system (LMS, for example Moodle). Practical related issues will be discussed in the section 5.

4.4 Grading process

Assessment set can turn out as unacceptable because of a single outlier grading. As an attempt to eliminate the outlier grading we propose to search for a maximal acceptable subset of the received peer-assessments. If such subset can be found, it is then used as input for the final grade calculation.

As a measure of final resort, an supervisor's intervention is asked for. In a course with a large student enrollment (thousands for a MOOC) this will be avoided as much as possible. However, if present, instructor's assessment becomes a final grade (no need for calculation). This is described in Algorithm 1.

4.5 Normalization

Metric d can be linearly scaled to obtain a normalized metric d_0 with values within the interval $[0, 1]$. Distance of $d_0 = 1$ corresponds to the maximal distance between worst and best possible gradings.

This would facilitate having general recommendations for setting acceptable error e on a normalized scale (setting $e_0 = 0.2$ for example). Additionally, this could facilitate comparison of data from different tasks (within a course, or from different courses).

Algorithm 1: Semi-autonomous Grading Process

input : Set of gradings $\mathcal{S} = \{S^{(1)}, \ldots, S^{(m)}\}$,
acceptable error $e \geq 0$
grading calculation method g
critical size N (i.e. $N = 3$)

output : Final grade or indicate gradings $\mathcal{S}$ as invalid

find a maximal $\mathcal{S}' \subseteq \mathcal{S}$ with acceptable error
if $\#(\mathcal{S}') \geq N$ **then**
  find $\mathcal{S}''$ of size $\#(\mathcal{S}'') = \#(\mathcal{S}')$ of minimal diameter
  return $g(\mathcal{S}'')$ as a proper grade for assignment k
else
  Ask for teacher intervention (grading)

4.6 Evaluation of peer-assessments (awarding the graders)

Goal of the second phase of the peer-assessment process is to reward the graders for their effort. Graders (peers) who have graded consistently and accurately (near the final grade) should be rewarded more than inconsistent and inaccurate graders.

Let us assume that a maximum of A points is awarded for the peer-assessment task. Then grader k can be awarded A_i points for each of the m gradings G_i that he/she was assigned, where A_i is calculated by the following formula

$$A_i(d_i) := \begin{cases} \dfrac{A}{me}\,(e - d_i) \ , & d_i < e \\ 0 \qquad\quad , & d_i \geq e \end{cases} \ ,$$

where $d_i = d(G_i, F)$.

This has the effect that 0 points are awarded for a grading outside of the e-sphere around final grade F. For a grading within this e-sphere A_i is proportional to $(e - d_i)$ where $d_i = d(G_i, F)$.

Finally, grader k is awarded a total of $A(k)$ points for his effort with gradings $G_1, \ldots, G_m$ where $A(k)$ is calculated as a sum of $A_i(d_i)$.

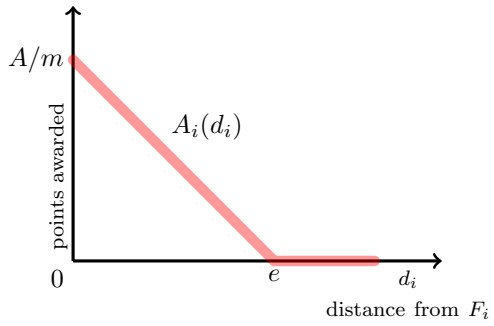

Figure 2: Points awarded to grader for grading G_i

5. IMPLEMENTATION

A support for peer-assessment LA is lacking in assessment analytics in general. We analyze the current implementation in the Moodle LMS where peer-assessment activity is implemented with the *Workshop* plugin.

In a *Workshop* activity, students receive a grade for their work and another grade for the quality of their assessment of other student's assignments.

Each participant in *Workshop* gets a grade for his submission and a grade for her assessments. These grades are visible as separate grade items in student's gradebook.

Current implementation of *Workshop* calculates the assignment grade as a weighted mean of received assessment gradings. Received gradings are not analyzed for reliability. If the teacher wishes to override or influence the calculated assignment grade, he can (a) additional provide his own assessment and set its corresponding weight to a higher value or (b) even completely override the final grade. As we have argued here and in [5] we find this method as inadequate. Therefore, we proposed alternative methods for the calculation of the final grade.

Assessment grade calculation is more complex. The goal is to estimate the quality of each assessment. One assessment is singled out as the best one – it is the assessment closest to the mean value of all assessments. This selected assessment is assigned with highest grade. Other assessments receive grades based on the distance from the selected assessment. Teacher can influence in this process by setting the parameter which determines how quickly a grade should decreased relative to the distance.

We are currently developing a new Moodle plugin for peer-assessment. This plugin will address the identified problems of the current implementation according to our model.

6. CONCLUSION. FURTHER RESEARCH

Peer-assessment has many advantages for students (for example development of metacognitive skills) and for teachers (for example saves teacher's time) but there are several challenges related to their implementation such as calculation of final grade, reliability check and awarding an evaluator for peer-assessment.

In this paper we propose new methods for calculation of the grades in peer-assessment. We propose a measure for reliability and a method for grading peer-evaluations in a peer-assessment exercise. These metrics are based on two distinguished scenario analysis that takes into account a number of possible evaluators and evaluator expertize (domain knowledge and evaluation skills). We pursue an approach to model assessment LA analytics with a geometric model.

In [4] we analyzed a case study based on the master level *Project Management* course at the University of Zagreb. Our analysis has confirmed the need for deeper analysis of reliability in peer-assessment. Further exploring of data related to the peer-assessment learning analytics in MOOCs is expected. Having additional data should result in improvement of the model and recommendations on the applicability of scenarios, parameters and analysis of the acceptable error of the assessment set.

Also, we intend to implement our model (algorithms and the supporting recommendation system) as a peer-assessment plug-in for the Moodle LMS.

Finally, we conclude that a well founded mathematical modeling, based on not just descriptive statistics, should be used more often in learning analytics.

7. REFERENCES

[1] Brown G., Bull J., Pendelbury M., "*Assessing Student Learning in Higher Education*", Psychology Press, 1997.

[2] Divjak, B. "*Implementation of Learning Outcomes in Mathematics for Non-Mathematics Major by Using E-Learning*", in Teaching Mathematics Online: Emergent Technologies and Methodologies, A. A. Juan, M. A. Huertas, S. Trenholm, and C. Steegmann, Eds. IGI Global, 2012, pp. 119–140.

[3] Divjak, B. "*Assessment of Complex, Non-Structured Mathematical Problems*", in IMA International Conference on Barriers and Enablers to Learning Maths, 2015.

[4] Divjak, B., Maretić, M. "*Learning Analytics for e-Assessment: The State of the Art and One Case Study*", CECIIS, 2015.

[5] Divjak, B., Maretić, M. "*Geometry for Learning Analytics*", Scientific and Professional Information Journal of Croatian Society for Constructive Geometry and Computer Graphics, KoG 19, 2015.

[6] Ellis, C. "*Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics*", Br. J. Educ. Technol., vol. 44, no. 4, pp. 662–664, 2013.

[7] Entwistle, N. J. "*Teaching for understanding at university: deep approaches and distinctive ways of thinking*". Basingstoke, Hampshire: Palgrave Macmillan, 2009.

[8] Ferguson, R. "*The state of learning analytics in 2012: a review and future challenges*", Tech. Rep. KMI-12-01, vol. 4, no. March, p. 18, 2012.

[9] Hwang, C.L, Yoon, K.,"*Multiple Attribute Decision Making and Applications*", NY, Springer Verlag, 1981.

[10] Jonnson, A., Svigby, G., "*The use of scoring rubrics: Reliability, validity and educational consequences*", Educational Research Review, 2007.

[11] Moodle LMS (`https://moodle.org/`) Plugins available on January 10th, 2016. at `https://moodle.org/plugins/`

[12] Papamitsiou Z., Economides A.A., "*Learning Analytics and Educational Data Mining in Practice*", A Systematic Literature Review of Empirical Evidence, Educational Technology & Societym 17(5), 49-64., 2014.

[13] Reyes Jacqueleen A., "The Skinny on Big Data in Education: Learning Analytics Simplified", TechTrends: Linking Research and Practice to Improve Learning 59 (April): 75–80. 2015.

[14] Sadler, P., Good, E., "*The impact of self-and peer grading on student learning*", Educ. Assess., vol. 11, no. 1, pp. 37–41, 2006.