Learning Analytics for e-Assessment: The State of the Art and One Case Study

Blaženka Divjak, Darko Grabar and Marcel Maretić
Faculty of Organization and Informatics
University of Zagreb
Pavlinska 3, 42000 Varaždin, Croatia
{blazenka.divjak, darko.grabar, marcel.maretic}@foi.hr

Abstract. Learning analytics deals with the data that occurs from students’ interaction with ICT: collecting data, analyzing and reporting that can influence learning and teaching. Analysis of validity and reliability of assessment lags behind other applications of learning analytics. We present here mathematical modeling of learning analytics for assessment, especially for peer-assessment. In addition, we analyze and categorize students’ recognition of advantages and disadvantages of peer-assessment. Finally implementations of reliability check of peer-assessment in Moodle Workshop module are explained.

Keywords. Learning analytics, assessment, peer-assessment, metrics, reliability and validity of peer-assessment

1 Introduction

Society today is characterized by a rapid social and economic change. From accelerating evolution of ICT arise needs for new competences such as self-regulated and peer learning, evaluation of peer work and metacognitive skills. The usual critique toward online tasks is that they rarely meet the requirements for development of higher order skills and higher order knowledge. Entwistle states that “Some of these advances [in e-learning], however, have done little more that move information around in more efficient ways.” (cf. [11], p. 138). Their development is enabled by deep learning (cf. [10]) and assessment has a clear connection with learning outcomes (cf. [1]) that comprise key competences. Our research is based on the Embedded Assessment Paradigm (cf. [23]), where learning analytics are used in order to interpret data about students’ learning, to assess their academic progress, to predict future performance and to personalize educational process. The 2015 edition of the Horizon reports learning analytics [12–14] as a midterm trend in education on a 3–5 year horizon, and just a year later in Horizon report 2016 learning analytics and adaptive learning along with Bring Your Own Device (BYOD) are expected to be increasingly adopted by higher education institutions.
in one year time [15].

We have conducted action research during the three year period in the course Project Management at the Master Level of Entrepreneurship study programme at the Faculty of Organization and Informatics (FOI) at University of Zagreb in which 131—107 students were enrolled — in academic year 2013/2014 and 2014/2015. Assessment and learning tasks were carefully prepared in the blended learning environment and clearly connected with intended learning outcomes of the course and the study programme (cf. [4]).

This paper is organized as follows. After briefing the current state of the art, we investigate the possibilities of combining peer assessment with learning analytics to enhance deeper learning approach by students. Specifically, we propose a new metric to measure peer assessment and self assessment reliability and discuss validity of peer reliability of peer-assessment and self-assessment.

For initial prototyping we present the results on the test data gathered in the last three years of the Project Management course.

Constructive alignment on the PM course

Students work individually on essay writing and peer evaluation 15—20h = 0.6ECTS (approx. 15% of the course’s 4 ECTS)

Students work in teams on task of project application writing and self—and peer evaluation 30—40h = 1.5ECTS (approx. 30% of the course’s 4 ECTS)

Finally, based on literature, students’ feedback and authors own experience, we analyze advantages and disadvantages of peer-assessment.

2 Learning Analytics for Assessment: State of the Art

Learning analytics (LA) as a research field is quite new — but already very propulsive and influential. Still, the research arena is just shaping and its research methods are still under construction. Learning analytics deals with analysis of data: LA analyzes data primarily, but not exclusively, produced by student’s interactions with information and communication technology (ICT) and especially with Learning Management System (LMS) where huge quantity of data is stored. The following definition of learning analytics is the least contested: “Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for the purposes of understanding and optimizing, learning and the environment in which it occurs.” This definition, according to [12], originated at the first international Conference on Learning Analytics and Knowledge (LAK2011) and was adopted by the Society for Learning Analytics Research (SoLAR) — [17].
LA is, as an interdisciplinary field, positioned at the intersection of several disciplines: business intelligence, web analytics, educational data mining and recommender/recommendation systems (cf. [12]). LA’s motivations and research ideas come from education science and mathematics, specifically geometry and metric spaces. Application area of LA is certainly in formal and informal education but also in non-formal learning. Basically LA is all about learning. Gašević and Dawson in [13] stress: “That is, instructors expressed their preferences of learning analytics features that offer insights into learning processes and identify student gaps in understanding over simple performance measures. With such insights, instructors can identify weak points in the learning activities performed by their students; topics the students have struggled with, and provide instructive and process related feedback on how to improve their learning.” Further, Ellis and Ferguson in [9] discuss definition of Learning Analytics and Knowledge and point out two limitations: (1) limited usefulness from both practical and pedagogical perspective; and (2) limited focus where only a portion of the student body is considered with too often students that are neither at risk nor the best forming an “overlooked middle”. Further, the author argues that in “...the scholarship on learning analytics, assessment data are almost never considered or referred to as part of the available data sets that can inform learning analytics.” The reason behind this, she argues, is most likely “...a direct product of the fact that, until relatively recently, the possibility of collecting and collating assessment data at a level of granularity that is meaningful and useful has simply been unthinkable.” Finally, among several sets of assessment data, [9] mentions “achievement mapped against explicit learning outcomes or assessment criteria (e.g., rubrics results)”.

This paper argues for the need and opportunity of utilizing results from granular assessment criteria (rubrics) in order to have insights into students learning as well as to evaluate the reliability and validity of student peer-assessment. Our motivations and research ideas come from education practice and mathematics, specifically geometry and metric spaces.

At the same time we are using e-assessment embedded in Moodle Learning Management System (LMS). It is possible to implement e-assessment for LMS Moodle for assessment of complex problems and authentic tasks (cf. [5]). In the area of e-assessment shift has been made [4, 5]. A shift from computer-based assessment towards embedded assessment is happening in the area of e-assessment (cf. [23]).

In our approach we are less inclined to conform to the paradigm of Explicit Testing. We are much closer to the Embedded Assessment paradigm which does away with tests and instead, via Learning Analytics, uses the data produced during the learning process as a basis for providing feedback and guidance to both learners and teachers (see [23]).
3 **Assessment and Peer-Assessment**

**Skills**

3 **Peer-Assessment**

There is a considerable pool of research on employability and 21st century skills. Skills identified by these researchers as most wanted and most important for long-term employability are ability for lifelong and peer learning, the ability for managing own learning and peer-earning, the ability to successfully work in groups, making judgments about peer work, ability of objective judgment of peer-work (critical thinking) as well as metacognitive skill of reflecting on her/his reflection about own learning and performance [22]. Consequently, we should strive to enhance and develop exactly these skills through formal and informal, informal and non-formal learning.

Formative assessment and feedback can help students take control of their own learning, i.e. become self-regulated learners ([21]). According to [24], peer assessment and self-assessment have offer the following four advantages:

1. **Logistical** – because it saves teachers time.

2. **Pedagogical** – because judging the other students work is an additional opportunity for students to deepen their understanding about a topic.

3. **Metacognitive** – because grading can help to demystify testing and students become more aware of their own strengths, progress and gaps in knowledge and skills.

4. **Affective** – because these types of assessment can make students more productive and cooperative, and thus can build a greater sense of shared ownership for the learning process.

In general this means that students are Peer-assessment engages students to become more active learners, more responsible, take responsibility for their learning, to apply deeper learning strategies and have to gain a better understanding of their own subjectivity and judgment. At the same time, we (the authors) recognize some possible disadvantages of peer assessment which we classify in the following four groups and add possible strategies to mitigate them based on the literature review and our own experience.

1. **Logistical** – because students need additional briefing time and teacher has to plan extra time for discussion of assessment criteria, goals, write some instructions in LMS, implement scoring rubrics etc.
(2) *Reliability risk*—because students are assessing their own peers. Some of their peers can be their friends and others can be members of other cliques in the classroom. Therefore teacher must be aware of it and if necessary anonymize assessment tasks.

(3) *Equalizing*—because of a tendency to award everyone the same mark. Learning analytics can help, especially with bigger groups, to discover assessment patterns and aid the discovery of assessment patterns (especially in large groups).

(4) *Metacognitive*—because not all students are well equipped to undertake peer assessment and they have not developed metacognitive skills so far. Therefore, teacher should start with the self assessment tasks that have lower stakes to train the students and use LA analysis to analyze reliability of peer assessment whenever necessary (big groups, high stakes assessments).

Finally, students’ peer assessment can only be considered a satisfactory substitute for teacher assessment if the grading results are comparable to the teachers’ assessment. If students’ grades are not reliable, the teacher must override the assessment. Further, we must be aware that peer assessment of simple tasks (for example determining whether a claim is correct) is much easier than grading a complex task such as essay, problem solving or a project. In the later case students must be guided in their assessment tasks by discussing and explaining grading criteria and their weights (cf. [5]).

Assessment packages for LMSs have been developed to integrate self-assessment, peer assessment and summative assessment. These packages also often integrate the automatic analysis of learner data. In our case study we have used a package Workshop in the Moodle LMS as assessment support and data collection. Students are able to submit their work during the Workshop activity. Submissions can be assessed by teachers, self-assessed, or assessed by peers (students). The Workshop also allows multi-criteria assessment based on scoring rubrics. Students can obtain two grades in a single Workshop activity—one grade for their submission (that is how good their submitted work is) and another grade for their assessment (that is how well they assessed their peers).

4 **Case Study: Project Management Course**

We have conducted action research in a period of three years in the course Project Management (PM) at the Master Level of Entrepreneurship in which 131 students were enrolled. Assessment and learning tasks were carefully prepared in the...
blended learning environment and clearly connected with intended learning outcomes of the course and the study programme (details in [4]).

Constructive alignment (cf. [1]) has been prepared to pair learning outcomes (LOs) of the study program with the course LOs, and also to connect course LOs with teaching and learning methods, assessment tasks and student workload. The problems of specific LOs of PM were described in [2].

For the research presented in this paper we considered only LOs relevant for the peer-assessment. The constructive alignment for two LOs of this study programme is presented in the Table ?? . Careful preparation of constructive alignment is essential for validity of assessment.

Table ?? lists assessment tasks at PM course along with their percentage value relative to the total course grade-

3.1 Final grade calculation

Final grade for a received set of peer-assessments can be calculated in several different ways. The most common option is to calculate the average of received grades for each criteria. For example, column-wise mean for each criteria on the data in table ?? yields a tuple

\[(2.66, 1.66, 2, 0.66, 1.0)\]

as a final grade for this set of peer-assessments. Other options for calculation of the final grade are discussed in [7].

Assessment tasks at PM course-

<table>
<thead>
<tr>
<th>Assessment task</th>
<th>Percentage of the total mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artifcat—summary building Group work 3.4 with connected topics Recommendation with links</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Example data for scoring rubric peer-assessment

The maximal amount of available points for criteria \( C_i \) are very important and have to be determined with care by teachers and/or by a more sophisticated method. Divjak in [5] demonstrated an approach based on the multi-criteria group-decision making with groups formed by representatives of several stakeholders (teachers, students, former students, employers, etc.).
3.2 Case Study

LMSs offer packages for integration of self-assessment, peer-assessment and summative assessment. These packages support some automatic analysis of learner data. In our case study on a Project Management (PM) course we used the Moodle package Workshop (see [20]) for assessment support and data collection. In the course peer-assessment has been used for two assessment activities: peer-assessment for essay writing (low stake assessment) and peer-assessment of projects (high stake assessment). The first peer-assessment is used as a preparation for the assessment of projects that accounts for 30% of the final grade. Scoring rubrics were used for peer-assessment of both activities.

We analyzed students’ comments on peer assessment gathered in academic years 2013/2014 and 2014/2015. In the academic year 2014/2015 of the PM course we have analyzed peer-assessment data at the criteria for two peer-assessment activities: essays and projects. The goal here was to analyze and compare the reliability of peer-assessment at summative and criteria level.

4 Validity and Reliability of Peer-Assessment

In this section we try to answer the first research question regarding address the question of validity and reliability of peer assessment peer-assessment.

1. “Assessment is valid if it has to measure what was intended . . . Assessment is reliable if an equivalent grading would be given if marked again shortly afterwards or by another person. If assessment is not reliable, it cannot be valid; but an assessment can be reliable and yet be invalid, by accurately measuring the wrong thing.” (see Entwistle in [11], p. 157).

Checking and assuring validity Validity and assurance of assessment is a hard problem quite hard. Preparation of the teaching, learning and assessment with the use of constructive alignment is the first step in this process. Validity of assessment is evaluated relative with to the intended learning outcomes of study programme and consequently the course. Correspondence of assessment with the LOs—learning outcomes (LOs) can be prepared in many ways. Besides LOs—learning outcomes, the type of assessment depends on the must be chosen to take into account students’ prior knowledge, the size of a class, teacher’s workload, available resources etc. One possible assessment structure for the PM course is given in Table ?? and Table ??.

Additional options for verification of validity of the assessment can be performed through the use of student questionnaire querying about on the achievement of LOs and through tracking continued tracking of students in their career. Students’ perspective on the results of peer assessment will be presented in the next section.
Our aim in the case study is the analysis of the peer assessment of essays. Activities, type of students work, available help and duration are presented in Table ??.

Reliability for peer assessment for PM was checked by comparing the gradings from academic year 2014/2015 ($n = 62$ students) with two previous academic years ($n = 34 + 35$ students) when only teachers graded essays according to the criteria (scoring rubrics). As can be seen from Table ??, the results are comparable. These results correspond with the research in [24]. later career.

Comparison of assessments by academic year

**Academic year Average** 2014/15 ($n = 62$) 6.40/9 2013/14 ($n = 34$) (5.62/10) = 6.24/9 2012/13 ($n = 35$) (6.11/10)= 6.79/9 Reliability is studied through analysis of variations that occurs across raters and across different peer-assessment of the same individual rater. The former is commonly referred to as *intra-rater reliability*, while the latter is known as *intra-rater reliability* [16].

Comparison of assessment results during three years provides a starting point in analysis of assessment reliability (cf. [11]).

For a measure of reliability we have considered *Summative reliability, as analysis of the span of totals of peer assessments of the same work.* Peer gradings whose span is within 2 points (i.e. less than or equal) are considered reliable (consistent); peer gradings that exceed 2 point span indicate inconsistent gradings and such gradings are flagged as unreliable (requiring supervision).

Table ?? presents some data on reliability at the overall grade level and suggests that students’ evaluations are sufficiently reliable received totals, for the dataset of the *Project Management* course was analyzed in [6] and suggested that about 85% of assignments received sufficiently reliable peer-assessments. However, we argue that summative reliability is inadequate as it is lacks the necessary to provide valuable feedback to students. Therefore, by utilizing only summative reliability, an opportunity for deeper learning may be missed.

First check of reliability

Number of grades within the 2-point span Number of grades that equal or exceed the 2-point span 51–11 85%–15%

It is even more interesting to compare grading–

### 4.1 Reliability Measure

Detailed look of peer-assessments analyzed on the criteria level. For such analysis we have to introduce appropriate metrics. is interesting and reveals more information about reliability of peer-assessment.

Let $S = (c_1, c_2, \ldots, c_n)$ and $S' = (c'_1, c'_2, \ldots, c'_n)$ be tuples which represent gradings $S$ and $S'$ of the same essay according to the criteria $C_1, C_2, \ldots, C_n$. Gradings $S$ and $S'$ can be imagined as points in $n$-dimensional space.
The common and naïve approach is the use of Euclidean metric as a distance measure. Instead, we propose the use of the normalized 1-metric (known as taxicab of Manhattan distance, cf. [2]) in \( n \)-dimensional space, where \( n \) is the number of criteria in the rubrics. Let \( S = (c_1, c_2, \ldots, c_n) \) and \( S' = (c'_1, c'_2, \ldots, c'_n) \) be tuples describing two student gradings \( S \) and \( S' \) of the same essay according to the criteria \( C_1, C_2, \ldots, C_n \). \( S \) and \( S' \) can be imagined as points in \( n \)-dimensional space.

Distance between points \( S \) and \( S' \) can be calculated as normalized Manhattan distance: \( d(S, S') = \frac{1}{n} \mathcal{N} \left( \frac{|c_1 - c'_1|}{r_1} |c_1 - c'_1| + \cdots + \frac{|c_n - c'_n|}{r_n} |c_n - c'_n| \right) \), where \( \mathcal{N} \) is a total (maximum) number of available points for this activity. Normalized taxicab distance

Let \( S \) be a set of peer assessments (peer-assessments for the same work). As a measure for divergence of the assessment set \( S \) we propose taking assignment. We base the modeling of the reliability on the diameter of the peer-assessment set, i.e. a maximal pairwise distance between points gradings in \( S \):

\[
\max_{S, S' \in S} d(S, S').
\]

Manhattan distance (based on taxi-cab norm) is used because of the discrete nature of the assessment data. Normalization is introduced to allow future comparison with calculations based on different metrics. A divergent set of assessments \( S \) has a larger diameter, whereas a small diameter is expected for a coherent set of assessment. We propose the reliability measure \( \text{rel}(S) \) of the grading set \( S \) as

\[
\text{rel}(S) = 1 - \text{diam}(S).
\]

Reliability of the grading set takes values in the \([0, 1]\) range with 1 being the maximal reliability of the perfectly uniform grading set.

Weight \( r_i \) of criteria \( C_i \) can be determined by teachers and/or by group decision making with the use of multi-criteria decision making. A group in decision making is usually heterogeneous and consists of representatives of teachers, students and other stakeholders (former students, employers etc., cf. [5]).

We consider two options for gradings sets that are not acceptable, i.e. insufficiently reliable, and propose adequate solutions.

1. To demand supervision, i.e. to ask for teacher’s grading of the assignment and in that case teacher’s grade becomes the final grade. This approach is feasible in classroom of school settings with manageable number of students.
2. As low reliability of a grading set can be a result of an outlier grading (a singular odd grading, quite different that the rest), it is possible to exclude the outlier grading and consider the reliability of the remaining set.

This idea has application in situations with a sufficient number of available peer-assessments (MOOC setting) with the potential to decrease the teacher’s workload.

4.2 On the choice of metric and criteria of assessment

The use of the taxicab distance metric also known as Manhattan or rectilinear metric is advocated for several reasons. Distance between gradings can be calculated without the need for paper or a calculator. Also, a grade change on a single criterium exactly equals the total change. Figuratively, we may say that a grade "moves" in a rectangular grid, just like the Manhattan taxicab. Finally, the total of points, as the value most interesting to the student, is exactly the Manhattan norm of the grading vector. As taxicab metric is a metric in a mathematical sense, i.e. it satisfies the axioms of a metric, it has no disadvantages when compared with Euclidean metric.

Normalization, realized as scaling of the metric to the [0, 1] range, is introduced to allow for comparison of results and to facilitate recommendations in forthcoming implementations. For example, to allow a recommendation to take a grading set with a relative diameter less than 0.3 as acceptable.

4.3 Granular vs. summative reliability

Peer-assessments are commonly analyzed only at the summative level, i.e. by measuring differences of sums (totals) of received assessments from different raters. Tentative comparison of summative and granular reliability of received peer-assessments for our datasets suggests that a significant proportion of about 10 to 15% of assessment sets seem reliable at the summative level, but are not reliable when looked at the criteria level. In these assessment sets raters show significantly greater variety at the criteria level than at the summative (total) level in rating these assignments. This is a missed opportunity for deeper learning when raters would receive inadequate feedback for these assessments.

As mentioned before, our case study was performed on Project Management course in academic year 2014/2015 and on peer-assessment of student essays. Total of 62 students participated in the peer-assessment. Calculated Pearson’s correlation of diameter and summative difference of assessment sets is $r = 0.59$ for the PM course’s "essay activity" dataset and $r = 0.66$ for the PM course's "projects activity" dataset. This indicates that granular reliability is not a trivial extension of summative analysis.
Students’ Perception about Peer-Assessment

Finally, by utilizing the learning analytics collected in LMS we answer the following research questions:

1. **What is student perception about peer-assessment, assessment standards and criteria and mutual learning activity?**

   Answers on the second and the third research questions are based on students’ perception. There are two principal ways how students

2. **Is deeper learning encouraged by peer-assessment?**

   Students’ views on peer-assessment were collected through closed questions in questionnaire and by open questions in a form of e-journal in LMS. Students’ questionnaire was filled out by 45 students out of 62 for the academic year 2014/2015. The question relevant for peer-assessment was asked in the form of agreement with the claim: “Peer assessment of essay and projects motivated me on new way of thinking and learning.”

   Agreement with: “Peer assessment of essay and projects motivated me and introduced me to a new way of thinking and learning.” \(n = 45\) answers

   The results are presented in Figure ?? . It follows that 73.33% of students agree or even strongly agree with the claim that peer-assessment and mutual learning is motivating and that it opened new ways of learning for them. More details in [6].

   Furthermore students commented the peer-assessment exercise in a free form journal in a period of two consecutive academic years.

   Students’ perspective on whether deeper learning was encouraged through peer assessment was taken in the form of the e-journal where students answer the following four questions:

   1. **What you have learned through peer learning?**

   2. **Do you see link of peer learning to course learning outcomes?**

   3. **Was peer learning interesting?**

   4. **How to enhance the peer learning exercise?**

   Additional students’ comments were welcomed.

   Most common comments on learning and importance were:
Interesting and important (both—the course and the peer assessment) Students’ comments are systematized and presented in Table 2 to support advantages and disadvantages recognized in the literature. We confirm the existing groupings and additionally recognized one advantage (Empowering) and one disadvantage (Accountability) based on the received feedback. Namely, students commented that peer-assessment is very useful and enhances the development of their own skill set through assessment capability. Some students, at the same time, feel uncomfortable with the responsibilities that arise in peer-assessment. They consider themselves incompetent for assessment and consider it a teacher’s job.

I benefited from reflection on my own work—I had to see where I was not so good and I had to spot my own errors

I learned from others how to better structure an essay and how to do it in a more interesting way

We have learned more from assessment than in student’s presentation of Power Point slides

To learn how to assess is not easy, especially when you perform criteria-based assessment

- I appreciate think that most assessments are based on superficial impressions because they lack the link between theory-time and practice
  /or ambition for deeper analysis.
- I found out that assessing essays in a short period of time is hard—now I have much more respect for teachers’ work This was my first assessment of peer’s work.
- Could be enhanced/improved with a quiz activity to check what was learned and memorized.
- Interesting, should be introduced before final year.
- Had to think out-of-the-box to grasp the significance of this type of activity. As we haven’t done this before, it is not easy.

Assessment of peer’s assignment is very useful and closely related to the learning outcomes of this course.

Peer-assessment of essays and projects was hard and stressful (especially for projects). Nevertheless, I find it was helpful in reflection of my own skillset and assessment capability.
- I don't want to take responsibility for the grades of other students.
- We (the students) don't have the necessary competencies for assessment.
- I think that teachers must read all of the assignments and have a final say.
- I wish we tried this on a neutral example.
- I assessed the work of my friends, not my employees. Assessment is teacher's job, not students!

Certain useful suggestions from students were implemented in the second peer assessment task in the course. Students suggested that:

More recommendation on structure should be given:

Criteria should be explained in detail and to introduce more criteria and subcriteria;

6 Implementation

students dislike binary criteria (language, references); A support for transparent and meaningful peer-assessment learning analytics is lacking in assessment analytics in general. Reliability check of assessment is the core issue assessment and especially peer-assessment. We analyze the current implementation in the Moodle LMS that is Open Source, used at FOI.

More time in classroom should be dedicated to discuss how to write assess it and what results should look like;

6.1 Moodle Workshop plug-in

Assessment should be anonymized. Peer-assessment activity in the Moodle LMS is provided by the Moodle Workshop module. In peer-assessment activity, students receive two grades: a grade for their work and a grade for the quality of their assessments of other student’s assignments.

Each participant in workshop is graded first for his submission, and later for his peer-assessment(s). These grades are visible as separate grade items in student’s gradebook.

Calculation of the assignment grade for student's submission in Workshop is fairly simple. It is calculated as weighted mean of all received assessment grades without any reliability check. If the teacher wishes to influence the final submission grade, she/he can provide her assessment and set its weight (to a perhaps high value). The teacher can also entirely override received peer-assessments and set the final grade of the submission. The implemented logic behind the grade for assessment is more complex; it tries to estimate the quality of each assessment.
This functionality is available through a subplugin "Comparison with the best assessment". In order to estimate the quality of assessment, this subplugin selects the best assessment and rewards other assessments according to their distance from the best assessment.

Due to constraints of existing calculations and omission of reliability check we are developing a new Workshop subplugin that addresses the problem of grade calculation in peer-assessment, offers alternative methods for final grade calculation and most importantly offers the analysis of the reliability of peer-assessment at the criteria level. This subplugin implements the ideas and methods presented in this paper as well as in \cite{7, 8}.

7 Conclusion

Assessment guides learning and therefore it has to be carefully prepared, conducted, analyzed and enhanced. Especially important characteristics of assessments are their continuously improved. Especially important issues of assessments that need to be addressed are validity and reliability. For validity analysis it is important to introduce constructive alignment with intended learning outcomes but also to take students’ perspective on their achievements as well as to track their careers after graduation. For peer assessment reliability has to be carefully checked because several disadvantages can challenged reliability of results. We propose the modified Manhattan metrics (based on taxi cab norm) to be used in order to check on reliability and further develop in the scope of learning analytics. Future research in the interdisciplinary field of learning analytics of assessment that include modeling by different metrics arising from non-Euclidean geometry or multi criteria decision making is needed. In the case study of the PM course presented in the paper \cite{6} it was shown that peer assessment peer-assessment can be constructed to be valid and reliable.

Further, student perception is that peer assessment together with peer learning is motivating and opens new learning paths and that it trigger deeper learning approach. In that respect further research should be done specially in the peer assessment of more complex tasks such as problem solving or project tasks. Here we analyze reliability issues for peer-assessment where a difference can be observed in case of using granular approach based on assessment criteria instead of the total score.

We are acutely aware of the limitations of our current research—data is limited (since it is gathered from a single course). Therefore, it is early for generalization. So far, our results findings agree with previous related research (cf. \cite{24}) —and confirm the inadequacy of summative analysis of reliability. We propose the use of scaled Manhattan metric based on taxicab norm to model both inter-rater and intra-rater reliability.

Further research should be directed toward investigation of appropriate proposed
metrics for evaluation of peer assessment (especially for peer assessment for complex tasks such as problem solving, projects, etc.) and new pedagogical applications of learning analytics, etc.

Students perceived the value of peer-assessment as they reported that peer-assessment combined with peer-learning is motivating; opens new learning paths and triggers deeper learning. Based on the literature and received students’ feedback we support previously recognized advantages and disadvantages of peer-assessments. We recognize additional advantage and disadvantage category. Finally, there are the following systematization of advantages: Logistical, Pedagogical, Metacognitive, Affective and Empowering. Disadvantages are categorized as Logistical, Reliability, Equalizing, Metacognitive and Accountability.

References


