Metrics for Estimating Accuracy, Reliability, and Bias in Peer Assessment

M. Fertalj, LJ. Brkić and I. Mekterović
Faculty of Electrical Engineering and Computing/Department of Applied Computing, Zagreb, Croatia
melita.fertalj@fer.hr ljiljana.brkic@fer.hr igor.mekterovic@fer.hr

***Abstract* - The challenge of teaching staff being overworked and not having enough time to provide individualized feedback and guidance to students is a significant problem. Open-ended assignments foster creative thinking but add to the workload. The abundance of solutions makes individualized assessment necessary. Peer assessment, where students evaluate their peers' assignments based on their motivation, knowledge, and resourcefulness, has proven effective in higher education courses.**

This paper provides an overview of existing metrics used in peer assessment, defines what constitutes correct and consistent evaluation, and suggests ways to choose appropriate metrics for learning objectives. The paper also highlights limitations of measures presented in the published papers, shows the advantages and disadvantages of different ways of calculating the selected measures through examples.

Keywords - peer assessment; metrics; accuracy; reliability; bias; grade calculation; criteria of assessment; open-ended works

I. Introduction

Nowadays, with platforms like Coursera, edX and Udacity online learning is considered mainstream. Online teaching has created new opportunities for improving knowledge and online courses often have large groups. Whether it's online or traditional courses, students like any other type of consumer are increasingly raising their expectation of the learning service. Therefore, it is important to define an assessment method that is suitable for large groups of students which is also useful when the number of students is considerably greater than the teaching staff.

Solving open-ended assignments encourages students to tackle the problem analytically and apply their own understanding of the given task. For example, writing an essay or an engineering design doesn’t have an unambiguous solution like closed-ended assignments. Tasks where the answer is a simple yes or no or an exact number can be graded automatically. In open-ended assignments, the solutions can be in audio or video format, thus rendering automatic grading difficult or impossible.

A possible solution to this problem is the concept of peer assessment (PA) where students review their peers’ work while providing feedback on the quality of the work they submitted. The given feedback affects the learning process and the improvement of solving assigned tasks. Teaching staff of the department of Computer Science at North Carolina State University have compared 18 machine learning models to analyze student comments [1] to their peer’s work and concluded that comments that identify problems and offer suggestions that encourage students to improve their solutions.

Students might exhibit “rogue” and “non-rogue” [2] behavior when grading. “Rogue” students grade their peers’ work with the highest, lowest, average grade or simply rate randomly. Fair assessment of students' work means reviewing their assignments according to the expected quality standard. It is important to assign a fair grade to maintain the motivation of students already interested in the course material and to increase the motivation of others. Grades assigned by graders influence the final grades of their peers, therefore it is necessary to ensure that they are graded fairly. Students are awarded for the quality of their assignment solutions as well as the quality of their grading peers’ solutions. Calculating both the grade for the student’s solution and the grade for student’s evaluation in peer assessment should take bias, reliability, and credibility into account. Peer assessment can be implemented with calibration, where students are provided with assessment criteria called rubrics [3][4]. Using rubrics, students evaluate calibration tests of different qualities. Calibration tests are assignments created and graded by the teaching staff and their grades are considered correct. Students should evaluate as similar as possible to teachers.

II. Final grade calculation

We will analyze separately different methods for calculating the final grade on a student assignment using peer assessment. For some methods (e.g. Mean and Median) a calibration test is not required as it cannot improve the result. Examples of assigned grades and the final grades are presented and compared to understand which method is better in a particular case.

A. Mean

The grades assigned by the student graders are aggregated and the sum is divided by the number of grades. We assume that the student graders did not exhibit “rogue” behavior and do not require additional calculations to improve the accuracy of the obtained grades. For student *i* and grader *j*, the final calculated grade is marked as g_i. The grades are calculated as $g_i^{(x)} = \frac{\sum_j A_{i,j}^{(x)}}{|A_i^{(x)}|}$. Set of all grades assigned to student *i* is marked as A_i, while $A_{i,j}$ is the grade assigned to student *i* by grader *j*. Let us grade each assignment task individually, rather than the whole assignment. The index *x* in the formula above marks the task in question. E.g., $g_i^{(1)}$ is the grade for student *i* for the

1st task, whereas $A_{i,j}^{(1)}$ is the grade assigned by grader *j* to student *i* for the 1st task.

B. Median

Median is an easy way to calculate the final grade instead of calculating the average. Median determines the central value of the distribution. This method is used in peer assessment applications like Coursera and Coursebank. For set of numbers X with n elements ordered by size from smallest to largest, the median is calculated depending on if n is an even or odd number. When n is even is defined as *median(X)* = $\frac{x_{(n/2)} + x_{(n/2)+1}}{2}$. Whereas n is odd, the formula is as follows: *median(X)* = $x_{(n+1)/2}$. We are interested in calculating the median for an assignment with multiple tasks. For student *i* and task *x*, grade g_i is calculated using the formula below, where A_i is the set of all grades that have been assigned to student *i* for task *x:* $g_i^{(x)} = median(A_i^{(x)})$.

C. Calibrated peer assessment

The "Calibrated Peer Review" (CPR) system was developed at the University of California [5] and was also used at the Texas University [6]. The CPR method for calculating the final grade is to evaluate the mean grade with the previously calculated weight. The students who evaluated closest to the teachers' grading are awarded the greater weight. Since their grading is considered most accurate among their peers, their grading has the most impact on the final grade. The weights are determined with assignments called calibration tests. These tests are created and graded by teachers and their grades are considered as a reference to compare how a student has graded in relation to the teacher. During peer assessment, students also grade one or more calibration tests along with their classmates' assignments. The expected grades for the calibration tests are teachers' grades who previously graded them in preparation for the peer assessment process.

The formula for calculating the mean is modified so that the weight w_j for the grader *j* is used and the final calculated grade is defined as $g_i^{(x)} = \frac{\sum_j w_j^{(x)} A_{i,j}^{(x)}}{\sum_j w_j^{(x)}}$. The weight for each student is determined from the difference between the grade they have assigned and the teacher's grade. For *N* calibration tests (assignments) each assignment consists of multiple tasks. The correct grade that should be assigned $t_n^{(x)}$ is defined for the assignment task *x* on calibration test *n*. The difference between the grade assigned by student grader j and the correct grade assigned by the teacher for task *x* is determined as $\Delta g_j^{(x)} = \frac{\sum_n \left(A_{n,j}^{(x)} - t_n^{(x)}\right)^2}{N}$. If the student hasn't graded a task on the calibration test, that test is ignored in the formula above. Only the differences for the other calibration tests are calculated, where N represents the number of calibration tests that are not disregarded. The greater the difference $\Delta g_j^{(x)}$, the smaller the credibility of the student grader. Therefore, the weight is defined as

$$w_j^{(x)} = \begin{cases} w_{max,}\ if\ \Delta g_j^{(x)} = 0 \\ \frac{1}{\Delta g_j^{(x)}}, else \end{cases}$$

. If the student has graded the calibrated test exactly like the teacher, the difference $\Delta g_j^{(x)}$ is 0. Therefore, the weight of that student will have the maximum weight value calculated for other students.

D. Comparing mean, median, and calibrated PA calculations

Let us examine the methods for calculating the final grade mentioned in the previous paragraphs. Example of assigned grades are presented in TABLE I. The correct grades for students 1, 2 and 3 will be determined by using the grades given in the peer assessment process. The teacher has evaluated the calibration test and the assigned grades are considered as correct. A Lickert scale of 1-5 is often used for grading [7].

TABLE I. ASSIGNED GRADES

Grader *j*	**Student 1**			**Student 2**			**Student 3**			**Calibration**		
Grader 1	1	1	1	1	1	1	1	1	1	2	4	1
Grader 2	5	1	2	3	4	2	5	3	1	5	3	4
Grader 3	5	2	1	4	3	4	3	5	1	4	1	3
Correct grade $t_n^{(x)}$	/	/		/	/		/	/		**5**	**2**	**3**

Before calculating the final student grades with mean, median and calibration, the CPR method has an additional step where the weights for each student are calculated. Using the formulas mentioned in the *Calibrated peer assessment* paragraph, we determine that there is *N* = 1 calibration test with *x* = [1, 2, 3] tasks. It is also necessary to set an upper limit for weights, because if the grader's grades correspond perfectly with the teacher's grades, the weight will be very high and would prevail over all other weights. We set the upper limit for the maximum weight value to 0.9. The differences between student and teacher grades as well as the calculated weights are as follows:

TABLE II. GRADE DIFFERENCES AND WEIGHTS

Grader *j*	$\Delta g_j^{(1)}$	$\Delta g_j^{(2)}$	$\Delta g_j^{(3)}$	$w_j^{(1)}$	$w_j^{(2)}$	$w_j^{(3)}$
Grader 1	$(2-5)^2 = 9$	$(4-2)^2=4$	$(1-3)^2 =4$	$\frac{1}{9} = 0.11$	$\frac{1}{4} = 0.25$	$\frac{1}{4} = 0.25$
Grader 2	0	1	1	0.9	0.9	0.9
Grader 3	1	1	0	0.9	0.9	0.9

Mean grade is calculated as a grade for the assignment, while grade for each task can also be calculated separately. The same is valid for calculating the median grade. As shown in TABLE III., the mean grades differ from the median grades. Using one method instead of the other makes more sense depending on the expected grading result. E.g., student graders can assign grades in range [1, 5] and the calculated grade should be whole or half of the grade. In this case, calculating the median results in integer or partial values.

TABLE III. COMPARISON OF MEAN, MEDIAN, AND CALIBRATED PA CALCULATIONS

		Student 1	Student 2	Student 3
Mean	$g_i^{(1)}$	3.67	2.67	3.00
	$g_i^{(2)}$	1.33	2.67	3.00
	$g_i^{(3)}$	1.33	2.33	1.00
	$\overline{g_i}$	2.11	2.56	2.33
Median	$g_i^{(1)}$	5.00	3.00	3.00
	$g_i^{(2)}$	1.00	3.00	3.00
	$g_i^{(3)}$	1.00	2.00	1.00
	$\overline{g_i}$	2.33	2.67	2.33
CPR	$g_i^{(1)}$	$\frac{0.11*1+0.9*5+0.9*5}{0.11+0.9+0.9} =$ 4.77	3.35	3.82
	$g_i^{(2)}$	$\frac{0.25*1+0.9*1+0.9*2}{0.25+0.9+0.9}$=1.43	3.19	3.63
	$g_i^{(3)}$	$\frac{0.25*1+0.9*2+0.9*1}{0.25+0.9+0.9}$=1.44	2.75	1.00

Calculating both mean and median grades is equally represented in the literature for determining assignment and task grades. The results also show that the student 1 deserved a higher grade for the 1st task than the grader 1 evaluated. Grader 1 assigned the lowest grade and has graded the calibration test poorly, so his assessment has the least weight in calculating the final grade.

III. Metrics

Metrics include measures used to monitor student activity and evaluate the success of teaching programs. Students learn through homework, quizzes, exams, or essays where the final grade is a metric comprised of the mentioned measures.

TABLE IV. contains different kind of measures found in research for bias, reliability, credibility, and accuracy. The terms are somewhat exchangeable i.e., reliability and credibility are both defined by consistency, Pearson's correlation coefficient is used for calculating reliability or accuracy depending on the literature. When calculating the final grade, it is possible to give less or more importance to the graders' evaluation, in which case the mentioned credibility index is used. Intra-rater reliability will not be analyzed in this paper because we are interested only in students reviewing the same assignment once.

TABLE IV. MEASURES FOR CALCULATING BIAS, RELIABILITY AND ACCURACY

Statement of meaning	Terminology detected in literature	Corresponding measures
Student evaluation is subjective because it is influenced by their perspective [8].	Bias	assessment median [8], average distance to diagonal [8], probabilistic models for estimating biases [9][10]
Inter-rater reliability is defined as consistent grading of different students on the same assignment. Intra-rater reliability is defined as consistent grading of the same student on the same assignment [11].	Reliability	agreement (exact, adjacent) and consistency (Pearson's and Spearman correlation coefficient) [11], probabilistic models for estimating reliabilities[9], Intraclass Correlation Coefficient [12], Krippendorff's alpha [13]
Grader credibility is defined by the dimensions [14]: Accuracy determines how close the assigned grade is to the grade we know to be correct (teacher's grade), consistency represents stability of assigned grades within the same test, and transferability provides information on how grader accuracy changes across multiple works.	Credibility	Credibility Index (CI)[14] [15], Reviewer Competency Index (RCI) [16], Competency index [2]
The teacher determines what the correct solution is and the associated grades for different qualities of the student solution depending on the deviation from the correct one.	Accuracy	Distance between gradings [17], Pearson's correlation coefficient [11], probabilistic models for improving accuracy [18] [19]

A. *Accuracy*

1) *Distance between gradings*

Distance between gradings published in the *Learning Analytics for Peer assessment: (Dis)advantages, Reliability and Implementation* article [17] considers each set of assigned grades for each student separately. From the set, the grades are compared pairwise. For example, for the student i, the grades of the student graders x and y are observed. The grades are organized into a vector and the distance between the vectors is compared. Grade vector for each task for student i that graded grader x is defined as $A_{i,x} = (A_{i,x}^{(1)}, A_{i,x}^{(2)}, \dots, A_{i,x}^{(n)})$, and for grader y the vector is defined as $A_{i,y} = (A_{i,y}^{(1)}, A_{i,y}^{(2)}, \dots, A_{i,y}^{(n)})$ where n is the total number of tasks in the assignment.

The authors of the mentioned article have opted for the modified Manhattan distance to determine the distance between these two vectors. Manhattan distance is calculated as the sum of the absolute differences of the Cartesian coordinates of the two vectors. The Manhattan distance between two points can be visually represented on a rectangular grid that is parallel to the coordinate axes where space between the gridlines is equal to 1. Distance between vectors $A_{i,x}$ and $A_{i,y}$ is calculated using the following expression, where A_{max} is the maximum possible grade on the assignment and X is the number of tasks:

$$d(A_{i,x}, A_{i,y}) = \frac{1}{A_{max}} \left(\frac{|A_{i,x}^{(1)} - A_{i,y}^{(1)}| + \cdots + |A_{i,x}^{(n)} - A_{i,y}^{(n)}|}{|X|} \right) \in [0, 1]$$

Vectors $A_{i,x}$ and $A_{i,y}$ are considered as set A. The diameter of A is defined as the maximum distance between two vectors in the set as $diam(A) = \max_{A_{i,x},A_{i,y} \in A} d(A_{i,x}, A_{i,y})$. If the student graded the other students' work exactly or similar as the teacher, the difference between assigned grades is small. Therefore, the diameter is smaller, and the accuracy is bigger. To demonstrate, let us consider that we have conducted peer assessment for an assignment thar consists of one task. Six students have participated in assignment and graded three of their peers. For student i, the grades are as follows:

TABLE V. ASSIGNED GRADES EXAMPLE

Assignments	Stud 1	Stud 2	Stud 3	Stud 4	Stud 5	Stud 6
$A_{i,1}^{(1)}$	4	5	4	4	2	3
$A_{i,2}^{(1)}$	5	4	4	5	2	3
$A_{i,3}^{(1)}$	5	4	4	5	2	4

The calculated distances are presented in TABLE VI.:

TABLE VI. DISTANCES

Assignment distances	Stud 1	Stud 2	Stud 3	Stud 4	Stud 5	Stud 6
$d(A_{i,1}, A_{i,2})$	0,2	0,2	0,0	0,2	0,0	0,0
$d(A_{i,2}, A_{i,3})$	0,0	0,0	0,0	0,0	0,0	0,2
$d(A_{i,3}, A_{i,1})$	0,2	0,2	0,0	0,2	0,0	0,2

The maximum distance $diam(A)$ is 0,2. Therefore, peer assessment accuracy is calculated and presented as a percentage: $acc(A)$ = 1 - $diam(A)$ = 1 – 0,2 = 0, 8 * 100 = 80%. As shown, the accuracy is high concluding that the students greatly agreed on the quality of the solutions they graded. The main disadvantage of the ***distance between gradings*** method is that the grades used in the described calculations are individual grades assigned by the student graders. Final grades are calculated independently and are not verified with this method. The overall agreement in grading between students with the advantage of not requiring teachers grades for calculation.

2) Pearson correlation coefficient

Pearson correlation coefficient [11] is used to measure of the linear correlation between each student grader's grade and the teacher's grade on the same assignment. The use of the Pearson's correlation coefficient is widely represented in the literature for various purposes is defined as

$$pearson(G_i, T_i) = \frac{\sum_{x\in X}(g_i^{(x)} - \overline{G_i})(T_i^{(x)} - \overline{T_i})}{\sqrt{\sum_{x\in X}\left(g_i^{(x)} - \overline{G_i}\right)^2 * \sum_{x\in X}\left(T_i^{(x)} - \overline{T_i}\right)^2}}$$

Where G_i is a set of grades calculated for student i, g_i is the calculated grade for the student i for task x, while T_i is the set of correct grades (defined by the teacher) for student i. The similarity obtaines values from the interval $[-1,1]$ where 1 represents complete similarity while -1 represents minimal similarity. The calculation is done for every student. The average value of all students is taken as the final result and cannot be expressed as a percentage. Positive Pearson's coefficient indicates that if the value of the first set of data increases, the values in the second set will also increase. Negative pearson coefficient indicates that if the value of the first set decreases, the value of the second set increases.

Let us examine an example for the Pearson's correlation coefficient so that we can inspect its properties. Vector $\overrightarrow{g_1} = (5, 4, 5, 3, 4, 2, 4, 5, 1, 5)$ is the vector of calculated grades. Vector $\overrightarrow{t_1} = (3, 1, 2, 2, 3, 1, 4, 2, 4)$ is the vector of teachers grades that are considered correct. The average grades are $\bar{g} = 3.8$ and $\bar{t} = 2.5$. To simplify the subsequent calculation, the respective averages are deducted from the vector values so now the vectors contain values as following:

$\overrightarrow{g_1} = (1.2, 0.2, 1.2, -0.8, 0.2, -1.8, 0.2, 1.2, -2.8, 1.2)$,

$\overrightarrow{t_1} = (0.5, -1.5, -0.5, -0.5, 0.5, -1.5, 0.5, 1.5, -0.5, 1.5)$. The Pearson's coefficient is calculated according to the formula mentioned above: $pearson(\overrightarrow{g_1}, \overrightarrow{t_1})$ = 0.5887. The value indicates that the correlation is relatively low which makes sense since the vectors contain distinctly different values.

For the 2nd example the $\overrightarrow{g_1}$ vector remains the same and the teacher vector is $\overrightarrow{t_2} = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1\ )$, with calculated average $\bar{t} = 1$. When average teacher grade $\bar{t}$ is deducted from the $\overrightarrow{t_2}$ vector, the values are now all 0, i.e. $\overrightarrow{t_2} = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$. The Pearson's coefficient is $pearson(\overrightarrow{g_1}, \overrightarrow{t_2})$ = 0. As seen in the example, Pearson's correlation coefficient gives a result of 0 in case when one of the vectors contains all 0 values. More precisely, the result is 0 if any of the vector components are set to the same number (e.g. the whole vector is 3). Of course, the student can solve the assignment 100% correct and will be assigned the same grade for all of the tasks but we cannot know if that is the case for certain.

The 3rd example demonstrates a Pearson's coefficient where the value cannot be interpreted as we expect. Again, the $\overrightarrow{g_1}$ vector remains the same and the teacher vector is $\overrightarrow{t_3} = (1, 1, 1, 1, 1, 1, 1, 1, 2, 1)$, with calculated average $\bar{t} = 1.1$. When average teacher grade $\bar{t}$ is deducted from the $\overrightarrow{t_3}$ vector, the new values are $\overrightarrow{t_3} = (-0.1, -0.1, -0.1, -0.1, -0.1, -0.1, -0.1, -0.1, 0.9, -0.1)$. The Pearson's coefficient is $pearson(\overrightarrow{g_1}, \overrightarrow{t_3})$ = -0.7075. In this case, the correlation coefficient is drastically different from the example with $\overrightarrow{t_2}$ vector, although the $\overrightarrow{t_3}$ vector changed only in one value. The ideal conditions would be to always expect minimal changes in the calculated results when the input g or t vectors contain slightest change. Accuracy is required to be 0 in the case with maximum error when grading, i.e. when the calculated grades of the student graders are farthest from the expected teacher's grade. For example, the teacher gave the highest grade (5) for all tasks in the assignment, but calculated grades which gave the students are the lowest grade (1). We also expect accuracy of 1 when the g and t vectors are equal.

From the previous examples it is clear that Pearson's correlation coefficient does not satisfy these properties.

B. *Reliability*

Spearman’s rank correlation coefficient is a measure for determining peer-grading reliability [11] between each student grader’s grade and the **median grade** on the same assignment. Pearson’s correlation coefficient is used to measure the linear correlation between student and teacher grades (paragraph *Accuracy*), while Spearman’s correlation measures monotonic relationships. If one value increases (or decreases), the other value also increases (or decreases). It is defined as $\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$, where n is the number of grades (X_i, Y_i) and $d_i = R(X_i) - R(Y_i)$ is the difference between the two ranks for each grade comparison. If there are no ranks that are tied i.e., there are no equal grades, then the grades are ordered from greatest to smallest with rank 1 assigned to the highest grade, rank 2 to the next highest and so on. When ranks are not tied, the above formula for ρ can be applied. When ranks are tied to two grades of the same category (e.g., two X_i grades), both grades are assigned a mean rank calculated from the two ranks in question.

Spearman’s coefficient returns a value from -1 to 1, where +1 indicates a perfect positive correlation between ranks, -1 value indicates a perfect negative correlation between ranks and 0 presents no correlation between ranks. Spearman’s coefficient presents the inter-rater reliability i.e., how consistently did different students grade the same assignment. Assigned grades are in TABLE VII., where the usual mark for grade g_i is now replaced as X_i and the median is marked as Y_i. The example in TABLE VII. presents an assignment consisting of two tasks, where two student graders evaluated three of their classmates.

TABLE VII. ASSIGNED GRADES, RANKS AND THEIR DIFFERENCES

	Student 1	Student 2	Student 3
$X_i^{(1)}$	1	3	5
$R(X_i^{(1)})$	3	2	1
$X_i^{(2)}$	5	1	2
$R(X_i^{(2)})$	1	3	2
$\overline{Y_\iota}$	3	2	3.5
$R(\overline{Y_\iota})$	2	3	1
$(d_i^{(1})^{2)}$	1	1	0
$(d_i^{(2)})^2$	1	0	1

$$\rho^{(1)} = \rho^{(2)} = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6*2}{3*(9-1)} = 0.5$$

In this case, we calculated how consistently did two student graders evaluate a task. For both tasks the reliability ρ is 0.5 i.e., grading is dispersed around the median. Therefore, the graders are strongly consistent in their evaluation. If the students’ grading is compared to the correct grades defined by teacher instead of the median, the Spearman’s correlation shows how consistently students grade in relation to the teacher. The **positive ρ** (closer to +1) indicates that students give higher grades for better solved assignments i.e., assignments that the teachers also grade higher. The **negative ρ** (closer to -1) indicates that students grade consistently but the exact opposite of how the teacher will grade the same assignment. The **closer ρ is to 0**, the relationship between the student grade and median is weaker, and the less consistent the grading is among the students. In this case, students assign high grades for an assignment, but is equally possible they will give lower grades for the same assignment. Therefore, their grading is not reliable.

C. *Bias*

To guarantee reliability in peer assessment, it is necessary to detect bias and provide objective grading of student assignments. In the previous paragraph, we examined the students’ lack of consistency in grading their peers. Student bias is displayed as the tendency to evaluate every assignment with the same grade or to consistently give the assignments higher or consistently lower grade then deserved. As mentioned in the *Introduction*, these students exhibit “rogue” behavior and their impact on the calculation of the final grade should be minimized.

There are different types of biases presented in the paper on metrics written by the teaching staff at University of Alicante [8]: restriction of range bias is the tendency to evaluate every assignment with the same grade; central tendency bias is the tendency to evaluate every assignment with the median or mean grade; leniency bias is the tendency to overrate, and harshness bias is the tendency to underrate assignments. We will examine the average distance to diagonal and its standard deviation measures to determine leniency and harshness biases.

For student *i*, the grader *j* assigns the grade a_{ij}. The grade assigned by the teacher (correct grade) is marked as c_i and n_j presents the number of grades assigned by the grader *j*. The average distance to diagonal is defined as $\overline{DD_j} = \frac{\sum_{\forall i} a_{ij} - c_i}{n_j}$ and presents the mean of differences between the grade assigned by the student grader and the correct grade assigned by the teacher. Every distance is positive if the student graded higher than the teacher, otherwise it is negative when the student graded lower. The standard deviation for the student grader j is defined as $S_{DDj} = \sqrt{\frac{\sum_{\forall i}(a_{ij} - DD_j)^2}{n_j - 1}}$. Considering both the DD_j and S_{DDj} values, it is possible to confirm the presence of leniency and harshness bias. If DD_j obtains a high positive value, the grader overrates his peers’ assignments which indicates leniency bias. Whereas low negative values of DD_j indicate that the grader underrates the assignments which represents harshness bias. The low values of standard deviation S_{DDj} shows that the graders bias is more pronounced.

For the assigned grades in TABLE VIII., the grader 1 has a high positive value for DD_j, therefore leniency is present i.e., grader 1 overrated the three tasks of the student 1 assignment. This makes sense for the given data since we know that grader 1 rated all tasks with the highest grade. Grader 2 has a negative DD_j value i.e., underrates the student 1 which concludes harshness bias. For grader 3 the DD_j is 0, thus indicating perfect evaluation equal to the teacher’s grades. All three graders have a high standard deviation value, so their grading habits are highly noticeable.

TABLE VIII. ASSIGNED GRADES, AVERAGE DISTANCE TO DIAGONAL AND STANDARD DEVIATION

Grader j	Student 1			DD_j	S_{DDj}
Grader 1	5	5	5	2	3.67
Grader 2	2	3	1	-1	3.8
Grader 3	3	4	2	0	3.8
Correct grade $t_n^{(x)}$	3	4	2	/	/

IV. Conclusion

Using peer assessment for calculating grades of assignments decreases time and effort pressure on the teaching staff. It is important to consider student behavior and ensure that the students are graded according to their grading habits. Students should be awarded for their assignment solutions as well as their grading of other students' work. The research shown in this paper has pointed out the interconnection of accuracy, reliability and bias when calculating the final grade in peer assessment process. "Non-rogue" students are reliable graders, and their grade should carry more weight when calculating the final grade. "Rogue" students tend to underrate or overrate. Therefore, they are considered unreliable and biased which can negatively affect the final grade. Their grade should carry less weight for the final grade calculation depending on the difference degree from the correct expected grading.

Analysis of different measures for accuracy indicates that Pearson's correlation coefficient isn't suitable in all grading cases. Choosing the right measure is critical for estimating accuracy, reliability, and bias because it affects the final grade calculation. Looking into calibrated PA, it is worth considering that if only one calibration test is used, which is for most cases, we are not able to evaluate its weight and the calculated evaluations will not be used. In other words, the weight is 0, which is one of the disadvantages of calibrated evaluation.

Peer assessment process is conducted using the Edgar system [20] for the *Databases* course at Faculty of Electrical Engineering and Computing. The PA process generates student data that can provide insight into student evaluation. Since every option that student selects in the Edgar system is saved, there is several years of log event data available for analysis. We plan to expand current analysis of metrics for estimating the best peer assessment grade and use the mentioned metrics on real student data collected from the Edgar system to improve the final grade calculation.

References

[1] Y. Xiao, Q. Jia, and J. Cui, "What kind of peer-assessment comments help improve learning outcomes? Evidence from a programming course Machine learning in peer assessment View project Understanding Butterfly Effects and Predictability in Lorenz Models and Global Models View project," 2021. [Online]. Available: www.aaai.org

[2] J. Hamer, K. T. K. Ma, and H. H. F. Kwong, "A Method of Automatic Grade Calibration in Peer Assessment."

[3] Y. Song, E. F. Gehringer, Z. Hu, J. Morris, J. Kidd, and S. Ringleb, "Toward Better Training in Peer Assessment: Does Calibration Help? Paper presented at the CSPRED 2016: Computer-Supported Peer Review in Education," 2016. [Online]. Available: https://digitalcommons.odu.edu/teachinglearning_fac_pubs

[4] "An Analysis of Calibrated Peer Review (CPR) in a Science Lecture Classroom," 2008.

[5] H. K. Suen, "Peer Assessment for Massive Open Online Courses (MOOCs) l sis (SNA) in OnlineCourses."

[6] Y. Hartberg, B. Gunersel, N. J. Simspon, and V. Balester, "Development of Student Writing in Biochemistry Using Calibrated Peer Review," 2008. [Online]. Available: http://cpr.molsci.ucla.edu/

[7] J. R. Rico-Juan, A. J. Gallego, J. J. Valero-Mas, and J. Calvo-Zaragoza, "Statistical semi-supervised system for grading multiple peer-reviewed open-ended works," *Comput Educ*, vol. 126, pp. 264–282, Nov. 2018, doi: 10.1016/j.compedu.2018.07.017.

[8] R. Molina-Carmona, R. Satorre-Cuerda, P. Compan˜-Compan˜, C.-Rosique, F. Farao´, and F. Llorens-Largo, "Metrics for Estimating Validity, Reliability and Bias in Peer Assessment*."

[9] C. Piech, J. Huang, Z. Chen, C. Do, A. Ng, and D. Koller, "Tuned Models of Peer Assessment in MOOCs," Jul. 2013, [Online]. Available: http://arxiv.org/abs/1307.2579

[10] A. A. Namanloo, J. Thorpe, and A. Salehi-Abari, "Improving Peer Assessment with Graph Neural Networks." [Online]. Available: http://www.tml.cs.

[11] Y. Song, Z. Hu, and E. F. Gehringer, "Closing the circle: Use of students' responses for peer-assessment rubric improvement," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9412, pp. 27–36. doi: 10.1007/978-3-319-25515-6_3.

[12] Y. Tong, C. D. Schunn, and H. Wang, "Why increasing the number of raters only helps sometimes: Reliability and validity of peer assessment across tasks of different complexity," *Studies in Educational Evaluation*, vol. 76, Mar. 2023, doi: 10.1016/j.stueduc.2022.101233.

[13] F. Garcia-Loro, S. Martin, J. A. Ruipérez-Valiente, E. Sancristobal, and M. Castro, "Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform," *Comput Educ*, vol. 154, Sep. 2020, doi: 10.1016/j.compedu.2020.103894.

[14] Y. Xiong and H. K. Suen, "A proposed Credibility Index (CI) in peer assessment The TRACE Center for Excellence In Early Childhood Assessment View project Cognitive Diagnostic Models (CDM) View project," 2014. [Online]. Available: https://www.researchgate.net/publication/281783815

[15] "What is 'Calibrated Peer Review' (CPR TM) and Why Are We Doing It?" [Online]. Available: https://cpr.tamu.edu/cpr/cpr/login.asp

[16] J. Russell, S. van Horne, A. S. Ward, E. A. Bettis, and J. Gikonyo, "Variability in students' evaluating processes in peer assessment with calibrated peer review," *J Comput Assist Learn*, vol. 33, no. 2, pp. 178–190, Apr. 2017, doi: 10.1111/jcal.12176.

[17] B. Divjak and M. M. Mareti´c, "Learning Analytics for Peer-assessment: (Dis)advantages, Reliability and Implementation," 2017.

[18] J. Xu, J. Liu, P. Lv, and P. Yang, "Improving Peer Assessment Accuracy by Incorporating Grading Behaviors," Dec. 2021, pp. 1162–1169. doi: 10.1109/ictai52525.2021.00184.

[19] A. Darvishi, H. Khosravi, S. Sadiq, and D. Gašević, "Incorporating AI and learning analytics to build trustworthy peer assessment systems," *British Journal of Educational Technology*, vol. 53, no. 4, pp. 844–875, Jul. 2022, doi: 10.1111/bjet.13233.

[20] I. Mekterovic, L. Brkic, B. Milasinovic, and M. Baranovic, "Building a comprehensive automated programming assessment system," *IEEE Access*, vol. 8, pp. 81154–81172, 2020, doi: 10.1109/ACCESS.2020.2990980.