Toward Valid and Reliable Assessment of Individual Contributions to Teamwork

Fedor Duzhin, Nanyang Technological University, Singapore
Megan Zheng Chi Lee, Nanyang Technological University, Singapore

The Paris Conference on Education 2024
Official Conference Proceedings

Abstract
In typical classroom settings, students collaborate on tasks and submit group work for grading, often relying on peer evaluations to determine individual grades. We are concerned with the method of converting pairwise peer evaluations into individual final grades. The most common way to do it is as follows: every group member has a certain number of points to distribute among the rest of the group, and the final score of every student is the average number of points she receives from other group members. We call this Pie-to-others. Assessments should be psychometrically valid and reliable. We argue that the Pie-to-others method of evaluating individual contributions to group work is valid but not reliable. Moreover, by constructing a mathematical model of peer evaluation, we can measure exactly how much Pie-to-others (or, more generally, any assessment method of individual contribution to group work) deviates from being reliable. We will explain the worst-case scenario, i.e., derive the theoretical largest possible difference between the outcome of the Pie-to-others and the fair grade a student deserves. By analyzing a large dataset (1201 students, 220 project groups, 6619 evaluations) collected in large undergraduate classes in an Asian university, we estimate that, in practice, about 1% of all students are misgraded by the Pie-to-others. Finally, we will present an easy fix to the pie-to-others method that makes it reliable.

Keywords: Mathematical Model, Group Work, Peer Evaluation, Game Theory

1. Introduction

Group work and peer evaluation are widely employed in professional settings, schools, and educational settings across different fields. In classroom settings, it is recognised for its benefits in developing soft skills such as fostering active participation among students, teaching students responsibility (Weaver & Cotrell, 1986), and hard skills such as fulfilling learning objectives of a course (Tu & Lu, 2004; Weaver & Cotrell, 1986). Therefore, instructors need to grade students individually and distinguish their grades for the project. Furthermore, as collaborative settings have the possibility of unequal contribution within groups (Kennedy, 2005), peer evaluation serves as a way to rate and compare an individual's contribution to the project. The instructor determines the final contribution scores of each student by submitting the peer evaluation scores to a grading mechanism.

In practice, students work together on a common task such as a project in a team of at least three students. The instructor observes and grades the end result of the project. However, as the instructor is unaware of the individual contributions of team members unlike the team members, the instructor has to grade their individual contributions in a practical, valid, and reliable manner.

2. Setup

Idealised Assumption. There exists an objective truth that is known to the students but not to the instructor, with the objective truth being n numbers whose average is 100.

Definition 2.1 (Peer evaluation matrix). A matrix of peer evaluation A is created based on the contribution scores reported by each student. Each column j (A_{*j}) represents the scores reported by student j to all other students i and each row i (A_{*j}) represents the scores received by student i from all other students j. Furthermore, the average of each column is 100.

The diagonal entries of the matrix may not necessarily be defined. In other words, students may or may not do self-evaluations.

Definition 2.2 (Pie-to-others). A group of students work as a team on a project. At the end of the project, everyone evaluates the contributions of their teammates (except his own) by distributing an average of 100 points among the rest of the group. The final score of every student is the product of the average number of points he receives from his teammates and the group score.

Example 2.1 Suppose the peer evaluation matrix for a group of four students, A, B, C, D is:

Student	A	B	C	D	Average
A	-	120	150	120	130
B	110	-	100	120	110
C	90	75	-	60	75
D	100	105	50	-	85

Table 1: Example of a Peer Evaluation Matrix for a Team of Size Four

The individual contribution of a student is given by the percentage of their contribution as compared to the average in the team. For example, the average here is 100, and student A contributed 130% as compared to the average of 100, and is awarded a score of 130 out of 100 for their individual contribution.

Suppose that the group score for project is 72 out of 100. Then the final individual grade for A for the project is given by (72/100)*130 = 93.6. The individual grades for this project are reflected in Table 2 below.

Student	A	B	C	D
Final grade	93.6	79.2	54	61.2

Table 2: Final Individual Grades

Definition 2.3 (Pie-to-all). Pie-to-all works in the same way as Pie-to-others except that self-evaluations are permitted. Therefore the diagonals of the peer evaluation matrix are non-zero.

Definition 2.4 (Mechanism). A mechanism is a method of converting pairwise peer evaluations into individual final grades. Ideally, a mechanism is reliable and valid to encourage truth-telling. Pie-to-others and Pie-to-all are examples of a mechanism.

Since peer evaluation typically counts towards a student's final grade and students understand how mechanisms work, students are interested in maximising their peer evaluation scores by gaming the system. Therefore, it is ideal for these mechanisms to be psychometrically valid and reliable.

Definition 2.5 (Validity). A mechanism or peer evaluation is valid when it incentivises collective truth-telling from students.

Example 2.2 Suppose the true contributions of A, B, C, D are as follows:

Student	A	B	C	D
True contribution scores	110	120	60	110

Table 3: Example of True Contribution Scores for a Team of Size Four

A valid peer evaluation should be:

Student	A	B	C	D	Average
A	-	118	97	114	109.67
B	124	-	106	124	118.00
C	62	64	-	62	62.67
D	114	118	97	-	109.67

Table 4: Valid Peer Evaluation

Definition 2.6 (Reliability). A mechanism is reliable when a student is awarded exactly what they deserve (see Figures 1 and 2 for examples) or as close to their true individual contribution or objective truth n as possible.

However, an unreliable mechanism is one that assigns students that deserve the same grade different grades (for example, for any given true contribution score, n_i, the yellow regions in Figures 3 and 4 indicate that the students deserving n_i will receive a score lying in a range).

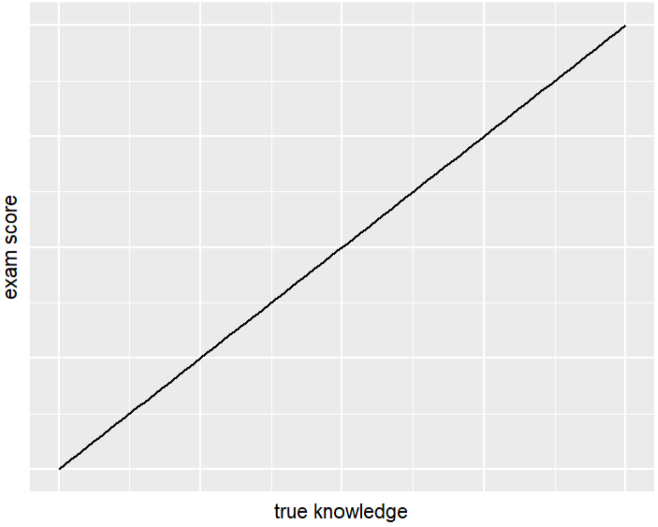

Figure 1: Example of a Reliable Mechanism

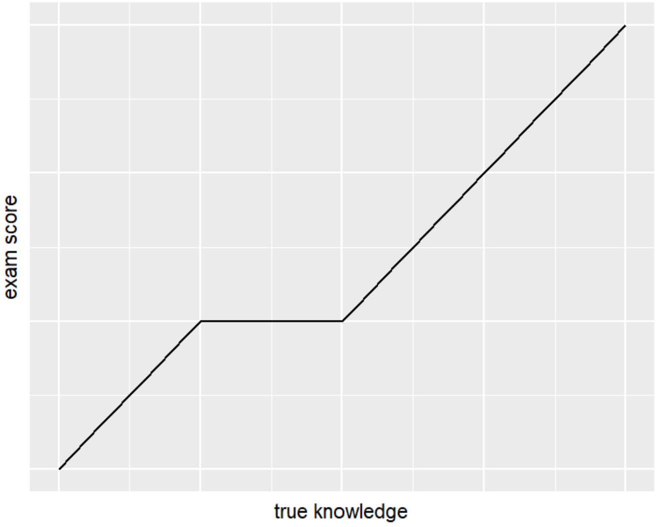

Figure 2: Another Example of a Reliable Mechanism

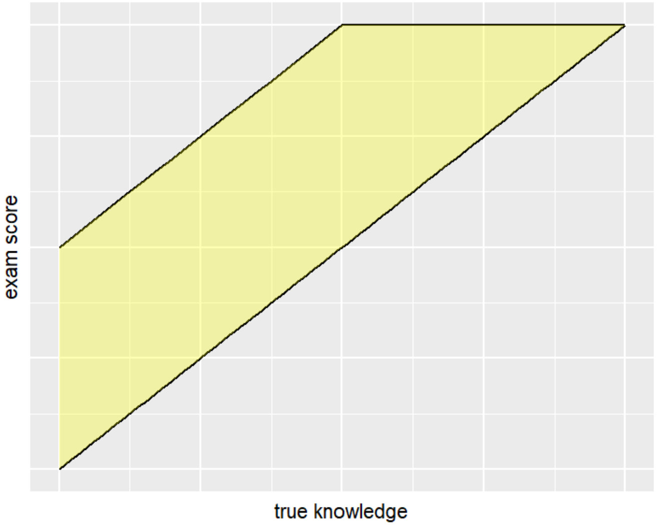

Figure 3: Example of an Unreliable Mechanism

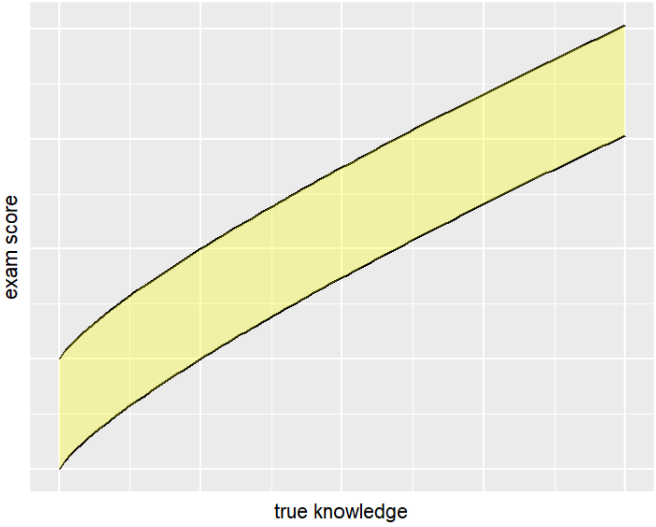

Figure 4: Another Example of an Unreliable Mechanism

3. Literature Review

In the literature, some mechanisms include awarding the same scores to team members and normalisation methods (Chowdhury, 2020; Couturier, 2018; Kaufman et al., 2000; Kennedy, 2005; Li, 2001; Malcolmson & Shaw, 2005).

One type of mechanism used in practice is awarding the same scores to everyone in the group. Although it is a fairly straightforward method, it has been criticised for its unfairness and impracticality. In practice, students contribute unequally in a team and should be graded according to their effort.

Another type of mechanism is normalisation methods. Normalisation methods are common grading methods used in practice that belong to a family of similar contribution assessments (Couturier, 2018; Kaufman et al., 2000; Li, 2001) that utilise Pie-to-others or Pie-to-all (Malcolmson & Shaw, 2005). Normalisations involve cardinal assessments and have been favoured due to its simplicity, transparency, and the preservation of the rankings (ordinal) of students' contributions (Li, 2001).

In Li (2001), the effectiveness of Pie-to-others against biased or inaccurate grading was investigated and found that without an additional bias factor, biased or inaccurate grading would skew the grades of students under Pie-to-others. This finding was consistent to an observation made in Chowdhury (2020), where normalisations were only effective if students were indifferent about the scores of their peers. In reality, students in a group tend to be partial to friends or collude to undermine their peers. Apart from biased or inaccurate grading, Pie-to-others was also used to identify free riders and ineffective team members in Couturier (2018).

However, in Kaufman et al. (2000), Pie-to-others and Pie-to-all were compared against one another to investigate the differences between self and peer ratings through statistical tests and correlations. Likewise, in Malcolmson & Shaw (2005), a similar type of investigation which compared the differences between Pie-to-others and Pie-to-all was conducted. However, they were done so in an arithmetically straightforward way which involved averages and standard deviations. A qualitative investigation was also conducted on Pie-to-others by evaluating students' feedback on their experiences with peer evaluation. Similar types of quantitative and qualitative analyses on Pie-to-others were also conducted in Kennedy (2005). A narrow spread of scores resulting from Pie-to-others was obtained in both Kennedy (2005) and Malcolmson & Shaw (2005).

4. Research Gap and Aim

Together, the literature highlighted that there were peer evaluation mechanisms studied from quantitative and qualitative perspectives. Yet, the mechanisms presented in the literature were investigated through simple arithmetic or focused on free-ridership. While free-ridership and qualitative feedback are important problems to study as free-ridership undermines the objectives of group work and qualitative feedback allow students to improve, a deeper investigation can be carried out by studying mechanisms that fairly grade students so that every student is rewarded appropriately and according to their effort. Hence, quantitatively designing a fair grading mechanism provides a more nuanced outcome than, for instance, identifying free riders. Therefore, this paper seeks to study grading mechanisms designed with game theoretic and mathematical ideas by:

1. Studying the theoretical unreliability and validity of Pie-to-others.
2. Quantify the practical unreliability of Pie-to-others using a dataset of real peer evaluations.
3. And finally, curing the unreliability of Pie-to-others.

5. Pie-to-Others

Theorem 5.1 Pie-to-others is unreliable.

Example 5.1 Suppose that the objective truth or ground truth g is (150, 75, 75) for a group of three students A, B, and C. Assume that Table 5 is a possible peer evaluation submitted.

Student	A	B	C	Average
A	-	133	133	133.0
B	100	-	67	83.5
C	100	67	-	83.5

Table 5: A Possible Peer Evaluation

Suppose that the ground truth for the same group is instead (150, 150, 0). Assume that Table 6 is a possible peer evaluation submitted.

Student	A	B	C	Average
A	-	200	100	150
B	200	-	100	150
C	0	0	-	0

Table 6: A Possible Peer Evaluation

In both versions, although A’s contribution is the same, her individual scores are different. Hence, we sought to find the theoretical range of scores a student can receive under Pie-to-others given that all team members report the truth.

Theorem 5.2 The maximum and minimum scores received by a student whose contribution is average (i.e., 100%) is:

n	3	4	5	6	7
Maximum score	133.33	150	160	166.67	171.43
Minimum score	100	100	100	100	100

Table 7: Maximum and Minimum Scores Received by an Average Student Under Pie-to-Others

Theorem 5.3 (Contribution levels and observations of scores received under Pie-to-others). Below average contributors are rewarded with higher scores than they deserve while students that contributed to more than half of the work are rewarded with lower scores than they deserve under Pie-to-others.

Example 5.2 Suppose that the below average contributor has a true contribution of 30% while the above average contributor has a true contribution of 150%. Table 8 reflects the minimum scores received by the below average and above average students.

n	3	4	5	6	7
Below average student	36	32.53	31.37	30.86	30.59
Above average student	133.33	142.11	145.45	147.06	147.95

Table 8: Minimum Scores Received by a Below Average Contributor and Above Average Contributor Under Pie-to-Others

Deducing from Theorem 5.3, students who contributed below average receives more than what they deserve, and are better off. However, for a student whose true contribution is more than half the work (above average), the maximum score the student could receive is less than what they deserve and are worse off.

These observations made are highlighted by the purple and red points in Figures 5 and 6, where the purple points represent below average contributors, and the red points represent above average contributors.

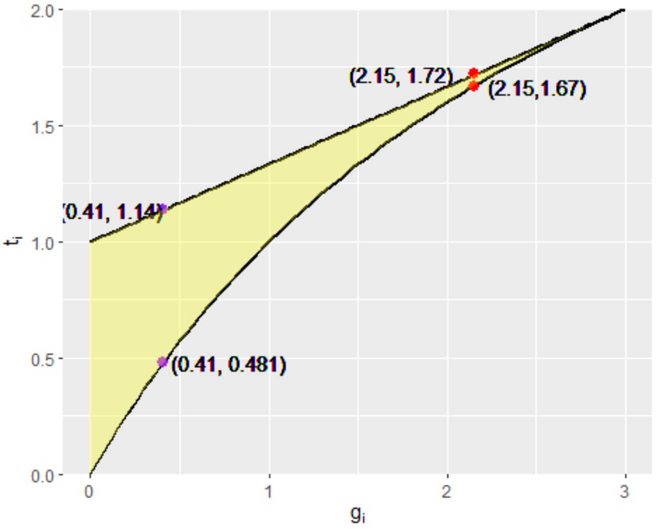

Figure 5: Range of Theoretical Contribution Scores for $n = 3$

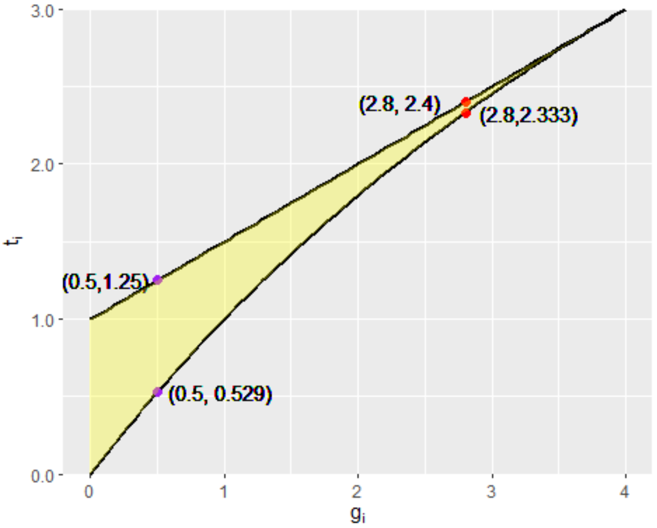

Figure 6: Range of Theoretical Contribution Scores for $n = 4$

In both Figures 5 and 6, the yellow regions reflect the ranges of scores student i may possibly receive depending on how much the rest of their teammates contribute to the project for a given true contribution g_i. Furthermore, the figures reveal the discrepancies in scores between the true contribution g_i and individual score t_i (for example g_i= 0.5 and 0.529 $\leq t_i \leq$ 1.25 in Figure 6). Therefore, there are errors in the scores under the mechanism. These observations and Theorem 5.3 allow us to conclude, theoretically, that Pie-to-others is an unreliable mechanism that does not reward students appropriately according to their effort.

Theorem 5.4 (Score discrepancies in the worst-case scenario). Under Pie-to-others, the maximum possible error, $t_i - g_i$, is 100 $-\frac{2}{n} g_i$.

From Theorem 5.4, we can deduce that for a student who contributed below average or close to zero can improve their scores by at most 100%, while a student who contributed above average will always receive strictly less than what they deserve.

The absolute relative error, $E_i = \left|\frac{t_i - g_i}{g_i}\right|$, was also calculated for average, below average, and above average contributors.

Theorem 5.5 The largest absolute relative error, E_i, of student i is largest when their contribution is minimal (i.e., below average). The varying contribution levels and the respective E_i are summarised below in Table 9, where $n \geq 3$.

Performance	Largest E_i
Below average ($g_i < 1$)	$\frac{1}{g_i} - \frac{2}{n}$
Average ($g_i = 1$)	$1 - \frac{2}{n}$
Above average ($g_i > \frac{n}{2}$)	$1 - \frac{(n-1)^2}{n^2 - 2n + g_i}$

Table 9: Largest E_i Corresponding to Different Performances

The next analysis conducted was calculating the practical unreliability of Pie-to-others using the Peer eval dataset. This dataset was obtained from several mathematics courses offered over a period in Nanyang Technological University (NTU) that had a total of 1201 students, 220 project groups and 6619 evaluations.

The true scores g_i in Peer eval were sorted in descending order. Then a letter grade was awarded to each g_i and resulting score t_i in the following manner (which was modelled after an old system used in NTU):

- A+: top 5%
- A: next 10%
- A–: next 15%
- B+: next 40%
- B: next 15%
- B–: next 10%
- C+: last 5% (all letter grades lower than B– were aggregated into C+)

Finally, we counted the number of instances where a student was misgraded, i.e., the letter grades of g_i differed from the letter grades of t_i. A sample of the results can be found in Table 10 below.

Student ID	g_i	t_i
fqHNjQ	A	A
retOzp	A	A–
vOiJFL	B	B
RdHXGD	B	B
NIAWWh	B–	B
sjyGpd	B–	B

Table 10: Sample of Results of Group nm8 From Peer Eval Dataset

Theorem 5.6 (Unreliability of Pie-to-others). From Peer eval, under collective truth-telling, Pie-to-others is 1.17% (2 d.p.) unreliable. In other words, 1.17% of the students were misgraded.

We also noted that Pie-to-others was not too unfair as extremes were uncommon, as shown below by the small percentage of cases with higher score differences in Table 11.

Score difference	Number of students	% cases
0	1124	93.6
1	71	5.9
2	4	0.3
3	2	0.2

Table 11: Percentages of Cases With Score Differences

Theorem 5.7 (Validity of Pie-to-others). Pie-to-others is a valid assessment.

Proof

Misreporting by a student does not affect their score when everyone else reports the truth. Hence, the best strategy is to report truthfully.

For example, suppose a ground truth g = (50, 100, 150) and the peer evaluation is:

Student	A	B	C	Average
A	-	30	60	45
B	70	-	140	105
C	130	170	-	150

However, if student A decides to misreport, the new peer evaluation matrix is:

Student	A	B	C	Average
A	-	30	60	45
B	110	-	140	125
C	95	170	-	132.5

Student A's individual score remains the same while his teammates' changes. Therefore, there is no incentive for A to misreport and the best strategy for each student is to report honestly.

6. The Cure for Pie-to-Others

In Section 5, we proved that Pie-to-others is valid but unreliable. We will modify Pie-to-others to allow self-evaluations and use normalised medians to evaluate individual grades. This improved assessment is called Median Pie-to-all.

Example 6.1 Table 12 is an example of a peer evaluation matrix for a group of students A, B, C, D under Pie-to-all.

Student	A	B	C	D	Median	Normalised median
A	140	130	130	110	130.0	128.4
B	100	120	120	110	115.0	113.6
C	75	70	90	80	77.5	76.5
D	85	80	60	100	82.5	81.5

Table 12: Example of a Peer Evaluation Matrix Under Pie-to-All

Theorem 6.1 (Validity and reliability of Median Pie-to-all). Median Pie-to-all is (i) valid and (ii) reliable.

Proof

(i) Valid: Misreporting by a student does not their score when everyone else reports the truth. Therefore, the best strategy for students is to report truthfully.

For example, suppose a ground truth $g = (x_1, x_2, \dots, x_n)$ for a group of n students A, B, … , N. Under collective truth-telling, the peer evaluation is:

Student	A	…	I	…	N	Median
A	x_1	…	x_1	…	x_1	x_1
…	…	…	…	…	…	…
I	x_i	…	x_i	…	x_i	x_i
…	…	…	…	…	…	…
N	x_n	…	x_n	…	x_n	x_n

However, suppose student A decides to misreport, where $x'_1 \neq x_1$ and the average of $(x'_1, \dots, x'_n)$ is 100. The new peer evaluation matrix is:

Student	A	…	I	…	N	Median
A	x'_1	…	x_1	…	x_1	x_1
…	…	…	…	…	…	…
I	x'_i	…	x_i	…	x_i	x_i
…	…	…	…	…	…	…
N	x'_n	…	x_n	…	x_n	x_n

Student A's individual score remains the same while his teammates' changes. Therefore, the best strategy for each student is to report truthfully given that everyone else is honest.

(ii) Reliable: Using the Peer eval dataset, Median Pie-to-all was found to be reliable. Furthermore, reconsidering Example 5.1:

Suppose that the ground truth g is (150, 75, 75). Assume that Table 13 is a possible peer evaluation submitted under collective truth-telling.

Student	A	B	C	Median
A	150	150	150	150
B	75	75	75	75
C	75	75	75	75

Table 13: A Peer Evaluation Submission

Suppose that the ground truth for the same group is instead (150, 150, 0) and that Table 14 is a peer evaluation submitted under collective truth-telling.

Student	A	B	C	Median
A	150	150	150	150
B	150	150	150	150
C	0	0	0	0

Table 14: A Peer Evaluation Submission With Different Ground Truth

In both versions, while A's contribution is the same, her individual scores also remain the same.

7. Conclusion

As pedagogical methods evolve and group projects become increasingly integral to the educational curriculum, the need to fairly assess students is also increasingly salient. As unequal contributions often happen in collaborative settings, peer evaluations offer insights into an individual's contribution to the task to course instructors for individual grading. Unsurprisingly, as students are often interested in maximising their scores and may game the system, they can be dishonest during their peer evaluations. Therefore, from a psychometric perspective, it is pertinent to employ valid and reliable mechanisms.

Although the prior mechanisms presented in the literature were successful in identifying free riders or provided qualitative feedback to aid a student's learning, this paper set out to use mathematical approaches to conduct a nuanced investigation into Pie-to-others. We had evaluated the theoretical and practical unreliability of Pie-to-others and found it was about 1% unreliable. However, Pie-to-others was proven to be a valid assessment. Pie-to-others was enhanced by permitting self-evaluations and replacing normalised averages with normalised medians, also known as Median Pie-to-all.

As this study had assumed the existence of an objective truth, it may be challenging to definitively quantify it in practice. Notwithstanding this limitation, the assessments in this study are easy to implement for educators.

References

Chowdhury, M. (2020). Using the method of normalisation for mapping group marks to individual marks: Some observations. Assessment & Evaluation in Higher Education, 45(5), 643–650. https://doi.org/10.1080/02602938.2019.1686606

Couturier, M. F. (2018). Identification of Ineffective Team Members Using Normalized Peer Ratings. Proceedings of the Canadian Engineering Education Association (CEEA). https://doi.org/10.24908/pceea.v0i0.13046

Kaufman, D. B., Felder, R. M., & Fuller, H. (2000). Accounting for Individual Effort in Cooperative Learning Teams. Journal of Engineering Education, 89(2), 133–140. https://doi.org/10.1002/j.2168-9830.2000.tb00507.x

Kennedy, G. (2005). Peer-assessment in Group Projects: Is It Worth It? 42, 59–65.

Lee, A. C. (2023). Mechansim design for collaborative work [Thesis, University of Illinois at Urbana-Champaign]. https://hdl.handle.net/2142/120291

Li, L. K. Y. (2001). Some Refinements on Peer Assessment of Group Projects. Assessment & Evaluation in Higher Education, 26(1), 5–18. https://doi.org/10.1080/0260293002002255

Malcolmson, C., & Shaw, J. (2005). The use of self- and peer-contribution assessments within a final yearpharmaceutics assignment | Pharmacy Education. https://pharmacyeducation.fip.org/pharmacyeducation/article/view/173

Tu, Y., & Lu, M. (2004). Mechanism Design for Peer and Self Assessment of a Group Project. The Proceedings of the Information Systems Education Conference, 21.https://citeseerx.ist.psu.edu/document?repid=rep1&amp;type=pdf&amp;doi=128bbde143c5b3565203384bcb4f1bceb861c039

Weaver, R. L., & Cotrell, H. W. (1986). Peer evaluation: A case study. Innovative Higher Education, 11(1), 25–39. https://doi.org/10.1007/BF01100106

Contact email: fduzhin@ntu.edu.sg