ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230322

Learning in Teams: Peer Evaluation for Fair Assessment of Individual Contributions

Fedor Duzhin[a;*]

[a]Nanyang Technological University
ORCiD ID: Fedor Duzhin https://orcid.org/0000-0002-9864-1249

Abstract. We develop a game-theoretical model of a classroom scenario, where n students collaborate on a common task. We assume that there exists an objective truth known to the students but not to the course instructor. Each of the students estimates the contributions of all team members and reports her estimates to the instructor. Thus, a matrix A of peer evaluations arises and the instructor's task is to grade students individually based on peer evaluations.

The method of deriving individual grades from the matrix A is supposed to be psychometrically valid and reliable. We argue that mathematically it means that 1) the collective truth-telling is a strict Nash equilibrium and 2) individual grade of student i does not depend on the true contribution of student j for $j \neq i$.

Existing methods of peer evaluation commonly used in educational practice fail to satisfy at least one of these properties. We construct a new method of peer evaluation satisfying both desired properties for $n \geq 5$. We share a large dataset (1201 students, 220 teams, 6619 evaluations) of peer evaluations collected in undergraduate courses taught by the author, outline some practical challenges, and show how these challenges can be addressed.

1 Introduction

A vast body of literature exists on methods of assessment in tertiary education ([20]). In practice, however, written final exams prevail, even though most students will never take an exam in their life after graduation and therefore exam grades are hardly able to capture the true potential of a student to thrive in a complex work environment.

Even though most people never take formal exams after leaving school, working in teams and writing reports are typical job tasks. Teamwork and report writing are taught at universities, but grading every individual student fairly based on a team's report is a challenge. For instance, if all the team members get the same grade, then a free-rider problem may occur — see [14], [12], [5], or [1].

The most obvious solution to the free-rider problem is peer evaluation ([7]). The simplest and yet popular approach to peer evaluation is letting each student grade contribution of each of the team members in absolute terms, i.e., out of 10, a 100, as A, B, C or in a similar way — see [11], [16], [18], or [4]. According to our experience, peer evaluation in absolute terms results in almost all students giving maximal scores to each other just not to jeopardize their friends' final grades. Thavikulwat and Chang criticize the whole idea of peer evaluation in [23] and propose to replace it with a different procedure based on students choosing their preferred group size.

There exist sophisticated peer evaluation systems. The system described in [5] is based on each student allocating a certain number of points between their teammates. Kauffman et al. introduce in [13] a mixed system where students give each other ratings from a list of nine terms such as "excellent", "very good" etc. but these ratings are then converted into a numeric value by dividing everyone rating by the team's average.

A system of peer evaluation is a procedure of calculating the "true" (at least, as it is perceived by team members) contribution of each of the team members to the common task based on mutual evaluations reported by team members. A system of peer evaluation may or may not have certain desired qualities. Among most important sought–after qualities of educational assessment are *validity*, *reliability*, and *practicality* ([10], [17], [19]).

A valid assessment measures what it is supposed to measure. For example, a test on factual knowledge cannot be a valid assessment for critical thinking. Likewise, a system of peer evaluation that allows students to manipulate their own scores by purposely lying about the contributions of their teammates is not valid since it assesses a skill of clever deception rather than honest teamwork.

A reliable assessment yields the same results each time it is used in the same setting. For example, an oral exam where a student is supposed to explain a randomly picked topic from the syllabus is not reliable since the grade of a student who learned half of the syllabus will depend on random chance. In terms of assessing individual contribution to teamwork, all students who worked in a team of, say, 5 and who contributed, say, 25% of the total team effort (i.e., more than the average in their respective teams), should get the same grade in the end.

A practical assessment should be easy to implement in a real classroom given existing constraints. For example, a written final exam is not practical for massive open online courses. Likewise, an assessment method for individual contribution to teamwork that is based on the instructor interviewing all students rather than on peer evaluation is not practical for large classes.

In this paper, we develop a mathematical model of peer evaluation for individual contribution to teamwork. We prove that the collective truth–telling in our model is a strict Nash equilibrium and argue that it means that the assessment system is valid. We also show that our assessment method is reliable. Finally, we explain how the instructor's judgement can be integrated into our assessment method.

* Corresponding Author. Email: fduzhin@ntu.edu.sg

2 Related works

While we are constructing a mathematical model of peer evaluation in learning teams, there are a few superficially similar problems in game theory literature.

The first of them is peer grading or peer assessment — see, for example [25], [22], [9], and [24]. Despite essentially the same name, the main scenario is completely different. In peer grading, a large group of students is required to submit their *individual* work (essays, assignments, reports etc.) to a common pool and then each student gets to evaluate a small number of peers' works according to criteria designed by the course instructor. The challenges are to distribute grading, write clear guidelines for students, and convert several grades given by peers to one final grade that is fair in the sense that it must be a close approximation to a grade that the instructor would give.

The second problem that is related to ours is peer nomination — see, for example, [2], [3], [15]. In peer nomination, the real-life motivation is a scenario where a number of researchers are competing for grants and they themselves select proposals that are worthy of funding. This setup is somewhat similar to ours in that there is assumed to be a ground truth — ranking, i.e., an order on the set of proposals. The key differences with our setup are that in peer nomination the group is large, i.e., every agent only evaluates a fraction of other agents' proposals, and that the output is binary while the output in our scenario is numeric.

The third game–theoretical problem that somehow resembles ours is fair division — see, for example, [8] and [21]. In fair division, n agents, too, compete for some common resource. However, the similarity ends here. The main difference is that in fair division, each agent has their own opinion on the value of each resource and the objective is to distribute resources between the agents in a desirable way. For example, it may be desired that each agent's fraction of resource is at least $1/n$ (proportional division), that everyone values their own share at least as anyone else's share (envy-free division) etc. In peer evaluation of individual contribution to teamwork, the resource (grades) is equally valued by all the agents, but there exists a ground truth about who deserves more and who deserves less of that resource. The main objective is to design a procedure that encourages all agents to evaluate others consistently with the ground truth and the main mathematical tool is a strict Nash equilibrium.

A rigorous mathematical theory of peer evaluation is outlined in [6]. However, their theory is very broad and does not provide a specific reliable method of grading (reliability is not even mentioned). Here, we are going to narrow the scope of the theory and, borrowing some ideas from [6], provide a practical method of grading.

3 Mathematical Theory

3.1 Setup

We assume that $n \geq 3$ students collaborate on a common goal of completing a set of well-defined tasks and there exists the objective truth — individual contribution of each student to teamwork. If the true contribution of the ith student is t_i, then the objective truth is the vector

$$t = (t_1, t_2 \dots, t_n) \in \Delta^{n-1} \subset \mathbb{R}^n,$$

where

$$\Delta^{n-1} = \{(t_1, \dots, t_n) : t_1 \geq 0, \dots, t_n \geq 0, \sum_{i=1}^{n} t_i = n\} \subset \mathbb{R}^n$$

is the $(n-1)$-simplex. Note that we require the mean individual contribution rather than the sum to equal 1.

The instructor observes the final product of teamwork (e.g., a report or a presentation) and evaluates the team with a score T. Intuitively, if the course instructor just wanted to give all students individual scores proportional to their effort, then the "fair" score given to student i should be $t_i \times T$.

The vector t is known to students but can't be observed by the instructor directly. What students report to the instructor is a matrix

$$A \in \mathcal{M}_{n\times n}(\mathbb{R}_{\geq 0})$$

of evaluations of each student by each student, where $\mathcal{M}_{n\times n}(\mathbb{R}_{\geq 0})$ is the set of $n \times n$ matrices with non-negative real entries. Note that even though practical systems of peer evaluation with negative scores exist, one can convert such a system to a system with non-negative scores simply by taking the exponential function of each score.

Further, let the entry in row i, column j of matrix A, i.e., a_{ij}, be evaluation of student i by student j. Then, for each i, the vector $A_{i*} = (a_{i1}, a_{i2}, \dots, a_{in})$ is the vector of evaluations received by student i and, for each j, the vector $A_{*j} = (a_{1j}, a_{2j}, \cdots, a_{nj})$ is the vector of evaluations reported by student j. The system of peer evaluation should motivate students to report truthful evaluations, i.e, A_{*j} should be proportional to t for all $j = 1, 2, \dots, n$ if all students are truthful.

Some additional conditions may be imposed on A in practice. For example, in all practical systems of peer evaluation known to the author, students cannot report that everyone contributed nothing, i.e., A_{*j} can't be the zero vector. Further, if self-evaluations are not collected, then $a_{ii} = 0$ for all i. It is also often required that the sum of evaluations reported by each student should be constant, i.e., that $A_{*j} \in \Delta^{n-1}$.

3.2 Mechanism

Definition 1 *A* mechanism *(a term adopted from [6]) is an algorithm for calculating the vector of individual grades, i.e., a function*

$$\begin{aligned} f : \mathcal{M}_{n\times n}(\mathbb{R}_{\geq 0}) &\longrightarrow \Delta^{n-1}, \\ A = (a_{ij})_{1\leq i\leq n, 1\leq j\leq n} &\longmapsto f(A) = g = (g_1, g_2 \dots, g_n). \end{aligned}$$

Note that the output of the mechanism, i.e., the vector g of individual grades may or may not be equal to the vector t of true contributions.

3.3 Valid and reliable assessment

Definition 2 *A mechanism is* incentive–compatible *if collective truth-telling is a strict Nash equilibrium, i.e., lying decreases one's own score given that others tell the truth.*

A mechanism is reliable *if, assuming that all students report the truth, then g_i is an increasing function of t_i. In particular, g_i does not depend on t_j for $j \neq i$. It means that*

$$g_i(t_1, \dots, t_n) = g_i(t'_1, \dots, t'_n)$$

holds for two different vectors $(t_1, \dots, t_n)$ and $(t'_1, \dots, t'_n)$ as long as $t_i = t'_i$.

We will, at least for the time being, assume that students report their evaluations independently, i.e., they do not form coalitions. This may sound like a strong assumption, but in practice, such coalitions are usually noticed by students themselves, reported to the instructor, and acted against. In that case, the only practical consideration that will discourage them from being truthful is the possibility of increasing their own score by deliberately reporting incorrect evaluations of other students. In other words, if a mechanism is incentive-compatible, then it translates to a valid assessment method.

To understand reliability, think of students A and B from different teams, whose teams got the same score, and whose true individual contribution is the same. For the assessment to be reliable, students A and B should get the same final grade.

Now we will outline two mechanisms that are commonly used in educational practice and show that one of them is not valid and the other one not reliable.

Pie-to-all mechanism

The *pie-to-all* mechanism works as follows. Each student gets a pie of size n and then distributes her pie among all team members, including herself, in proportion to their contribution to the team effort. The final individual grade of a team member is the average piece of pie that he received from all the team members. It means that we have

$$a_{ij} \geq 0, \quad \sum_{i=1}^{n} a_{ij} = n, \quad g_i = \frac{1}{n}\sum_{j=1}^{n} a_{ij}. \tag{1}$$

The pie-to-all mechanism is reliable if all students are truthful in their evaluations of themselves and others. Indeed, if all students report the truth, then the output of the mechanism is, clearly, the objective truth. It means that a hypothetical student whose contribution, is, say, 1.1 (i.e., 10% more than the average) will get a score of 1.1 regardless of the contribution of each of her teammates.

However, the pie-to-all mechanism is not valid because it includes self-evaluation and students have incentives to overestimate their own contribution. In the extreme case, g_i is maximized when $a_{ii} = n$ and $a_{ij} = 0$ for $i \neq j$ regardless of what the rest of the students report. The collective truth–telling is not a Nash equilibrium. In a practical educational setting, it is hardly possible to just claim "I did all the work and my teammates did nothing", but the very existence of the incentive to inflate one own's contribution may shift the focus of assessment from fairly evaluating everyone to cleverly justifying one's own high score. Still, the pie-to-all mechanism has been used in practice — see, for example, [13].

Pie-to-others mechanism

The *pie-to-others* mechanism works as follows. Each student gets one pie of size n and then distributes her pie among all her teammates, not including herself, in proportion to their contribution to the team effort. The final individual grade of a team member is the average piece of pie that he received from all the team members. It means that we have

$$a_{ij} \geq 0, \quad a_{ii} = 0, \quad \sum_{i=1}^{n} a_{ij} = n, \quad g_i = \frac{1}{n}\sum_{j=1}^{n} a_{ij}. \tag{2}$$

The only difference with the pie-to-all mechanism (1) is the absence of self-evaluation, which is expressed by $a_{ii} = 0$.

The pie-to-others is a very popular mechanism, probably the most popular one. Its clear advantage is that the collective truth–telling is a (non-strict) Nash equilibrium, for one can't change one's own score by deliberately lying. However, a huge issue with the pie-to-others mechanism is that it does not yield a reliable assessment method as the following examples show.

Example 1 *Consider a hypothetical team of three students whose vector of true contributions is*

$$t = \left(\frac{3}{2}, \frac{3}{4}, \frac{3}{4}\right).$$

If everyone is truthful, the matrix of peer evaluations and the vector of individual grades are

$$A = \begin{bmatrix} 0 & 2 & 2 \\ 3/2 & 0 & 1 \\ 3/2 & 1 & 0 \end{bmatrix}, \quad g = \begin{bmatrix} 4/3 \\ 5/6 \\ 5/6 \end{bmatrix} \neq \begin{bmatrix} 3/2 \\ 3/4 \\ 3/4 \end{bmatrix}$$

We see that the pie-to-others mechanism does not output the objective truth even when all students report the truth. In our experience, students who major in mathematics figure this out and become unhappy if they hear that they will be evaluated through the pie-to-others mechanism. However, this does not make the pie-to-others mechanism unreliable. To understand why it is unreliable, let us look at one more example.

Example 2 *Consider a hypothetical team of three students whose vector of true contributions is*

$$t = \left(\frac{3}{2}, \frac{3}{2}, 0\right).$$

The matrix of peer evaluations and the vector of individual grades are

$$A = \begin{bmatrix} 0 & 3 & 3/2 \\ 3 & 0 & 3/2 \\ 0 & 0 & 0 \end{bmatrix}, \quad g = \begin{bmatrix} 3/2 \\ 3/2 \\ 0 \end{bmatrix} = t,$$

which means that now all the students are fairly rewarded.

Examples 1 and 2 show that the pie-to-others mechanism is not reliable since the contribution of the first student is the same in both examples but her final score is different.

The unreliability of the pie-to-others mechanism is not its only issue. The other issue comes from the fact that it is profitable for the strongest students to just do all the work by themselves without letting their teammates do anything. It's then (and only then) that they will be fairly rewarded for their hard work. In the author's experience, this behaviour is common in real classrooms and a serious weakness of the pie-to-others mechanism is that it incentivizes such behaviour.

4 Reliable incentive–compatible mechanism

We assume that at most one entry of the objective truth vector t is 0, i.e., $t_i + t_j > 0$ and $a_{ik} + a_{jk} > 0$ whenever $i \neq j$. The author admits that this assumption may sound too strong. However, it comes from the author's experience and is probably specific to the cultural background. In the author's experience, situations, when some students do not contribute to teamwork in the beginning, are not unheard of. However, whenever this happens, students usually communicate it to the instructor so that he could encourage or warn non-contributing team members to start working. Students need to get a clear message that if they contribute nothing to teamwork, they will fail the course but if they do a bare minimum, they will probably pass, albeit with a low grade. A specific trick that helps is a preliminary non-graded round of peer evaluation. Another helpful action is meeting with all the learning teams and discussing their progress a few weeks before the deadline.

Also, we assume that $n \geq 5$. Note that our mechanism is reliable even for $n = 3$ and $n = 4$, but we do not know even if incentive–compatible mechanisms exist for $n = 3$ or $n = 4$.

4.1 Relative contributions

The key ingredient of our mechanism is the *relative contribution*

$$r_{ij}^k = \frac{a_{ik}}{a_{ik} + a_{jk}}, \tag{3}$$

of student i to student j according to student k. The quantity r_{ij}^k is the share of i's contribution in the combined i's and j's contribution in k's opinion. Note that if all students are truthful, i.e., evaluations they submit are proportional to the objective truth, then we have $a_{ik} = c_k t_i$, where c_k is the coefficient of proportionality, and

$$r_{ij}^k = \frac{a_{ik}}{a_{ik} + a_{jk}} = \frac{c_k t_i}{c_k t_i + c_k t_j} = \frac{t_i}{t_i + t_j}.$$

Thus, for every i and every $j \neq i$, we have the vector

$$r_{ij} = \left(r_{ij}^1, \cdots, \widehat{r_{ij}^i}, \cdots, \widehat{r_{ij}^j}, \cdots, r_{ij}^n, \right) \in \mathbb{R}^{n-2} \tag{4}$$

of relative contributions, where the terms with a hat are discarded. Note that $r_{ij} + r_{ji} = (1, \ldots, 1)$.

4.2 Auxiliary matrix

Now let

$$b_{ij} = \begin{cases} 1, & i = j, \\ \frac{\text{median}(r_{ij})}{\text{median}(r_{ji})}, & i \neq j. \end{cases} \tag{5}$$

Note that $b_{ij} \in \mathbb{R}_{\geq 0} \cup \{\infty\}$. We will call $B = (b_{ij})_{1 \leq i,j \leq n}$ the *auxiliary matrix* for the raw peer evaluation matrix A.

Lemma 1 *If all students report the truth, then*

$$b_{ij} = \frac{t_i}{t_j}$$

Note that, according to our assumption, t_i and t_j cannot be both equal to 0, i.e., $t_i/t_j \in \mathbb{R}_{\geq 0} \cup \{\infty\}$.

Proof 1 *If all students report the truth, then all entries of the vector r_{ij} are equal to $\frac{t_i}{t_i+t_j}$ and hence the median of r_{ij} is $\frac{t_i}{t_i+t_j}$. In the same manner, the median of r_{ji} is $\frac{t_j}{t_i+t_j}$. We can see that $b_{ij} = t_i/t_j$ if $i \neq j$ and it is obvious that $b_{ii} = 1 = \frac{t_i}{t_i}$.*

4.3 Perceived contributions

Now we will outline the main idea of our mechanism. Given a matrix of peer evaluations A, we first construct the auxiliary matrix B. Note that, if all students report the truth, then columns of B are proportional to the objective truth t with the coefficient of proportionality chosen so that $b_{ii} = 1$ for all i. Now, in order to extract the vector s of *perceived contributions* from B, we divide each column of B by the mean of its entries, then take row medians (here, medians are needed to avoid issues with infinite entries that may occur when a student contributed zero to teamwork), and normalize the result by dividing it by its mean. Below is a worked example.

Example 3 *In this example, we have a hypothetical team of 5 students. Let the matrix of peer evaluations be*

$$A = \begin{bmatrix} 1 & 1 & 5 & 2 & 2 \\ 1 & 2 & 4 & 2 & 3 \\ 1 & 3 & 3 & 3 & 4 \\ 1 & 4 & 2 & 3 & 5 \\ 1 & 5 & 1 & 3 & 6 \end{bmatrix}$$

Note that we do not require a common normalization. The full calculation of relative peer evaluations r_{ij}^k for all i, j, k will take too much space, so we will show r_{12}, r_{21}, and r_{23} here just to demonstrate the method. Rows of this matrix are indexed by pairs (i, j) and columns by k, the median of each r_{ij} is highlighted:

i	*j*	*1*	*2*	*3*	*4*	*5*
1	*2*			*5/9*	***1/2***	*2/5*
2	*1*			*4/9*	***1/2***	*3/5*
2	*3*	*1/2*			*2/5*	***3/7***

The auxiliary matrix B whose entries are row medians of R and their inverses is

$$B = \begin{bmatrix} 1 & 1 & 1/2 & 2/5 & 2/3 \\ 1 & 1 & 3/4 & 1 & 1 \\ 2 & 4/3 & 1 & 4/5 & 1 \\ 5/2 & 1 & 5/4 & 1 & 1 \\ 3/2 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Its normalized version with 2 decimal places and highlighted row medians is

$$B_{norm} = \begin{bmatrix} \boldsymbol{0.63} & 0.94 & 0.56 & 0.48 & 0.71 \\ 0.63 & \boldsymbol{0.94} & 0.83 & 1.19 & 1.07 \\ 1.25 & 1.25 & \boldsymbol{1.11} & 0.95 & 1.07 \\ 1.56 & 0.94 & 1.39 & \boldsymbol{1.19} & 1.07 \\ 0.94 & 0.94 & 1.11 & 1.19 & \boldsymbol{1.07} \end{bmatrix}$$

The vector of perceived contributions is the normalized vector of row medians, i.e.,

$$s = (0.63, 0.95, 1.13, 1.21, 1.09).$$

Note that our method of calculating perceived contributions still works if the contribution of one student is 0 and other students report it as 0. In that case, one column of B will have infinite entries and it will be impossible to normalize that column, but row medians will be well-defined.

Lemma 2 *Consider the mechanism $f(A) = s$, where s is the vector of perceived contributions calculated as described above. Let $n \geq 5$ and suppose that $n - 1$ out of n students report evaluations that are perfectly consistent, i.e., $n - 1$ out of n columns of the matrix A are proportional to each other. Then s is independent of the remaining column of A, i.e., evaluations reported by the last student don't affect the vector s of perceived contributions.*

Proof 2 *If $n \geq 5$, then each r_{ij} has at least 3 entries. If all students but one report perfectly consistent evaluations, then all but one entries of r_{ij} are equal. Changing the remaining entry won't change the median of r_{ij} and hence can't change the auxillary matrix B, from which s is calculated.*

Note that Lemma 2 implies that the collective truth–telling is a (non-strict) Nash equilibrium if $n \geq 5$.

4.4 Relative error of reported evaluations

Consider a peer evaluation matrix A and the vector of perceived contributions s calculated as described in section 4.3. Consider normalized columns of A, i.e.,

$$v_{ij} = \frac{n a_{ij}}{\sum_{i=1}^n a_{ij}}$$

Then

$$v_j = (v_{1j}, v_{2j}, \ldots, v_{nj}) \in \Delta^{(n-1)}$$

is j th version of truth and

$$\left| \frac{v_{ij} - s_i}{s_i} \right|$$

is the relative error of evaluation of student i by student j. Let

$$E_j = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{v_{ij} - s_i}{s_i} \right| \tag{6}$$

be the average relative error of student j's version of truth. Notice that $E_j = 0$ if and only if $v_j = s$, i.e., evaluations reported by j are perfectly consistent with perceived contributions s. It may happen that $s_i = 0$ for some i (we assume that at most one entry of s is 0). In that case, our convention is that $0/0 = 0$ and $1/0 = n$.

4.5 Our mechanism

Finally, our mechanism takes into account the vector of perceived contributions s calculated as described in section 4.3 and consistency of every student's version of truth with s defined by (6). The final grade of student j is

$$g_j = (1 - \varepsilon)s_j + \varepsilon \max(1 - E_j, 0), \tag{7}$$

where $\varepsilon > 0$ is some small number. The author sets $\varepsilon = 0.05$ in his classroom, but the exact value of ε is not important. Note that the mean grade is less than 1 unless all students report perfectly consistent evaluations. In the author's experience, the fact that the mean grade is less than 1 does not lead to any practical issues. However, the vector of final grades calculated according to (7) can be normalised if needed.

Lemma 3 *Consider the mechanism defined by* (7). *The collective truth–telling is a strict Nash equilibrium if* $n \geq 5$.

Proof 3 *Suppose that all but one student report the truth, i.e.,* A_{*j} *is proportional to* t *for all* $j \neq k$. *According to Lemma 2, it follows that* $s = t$ *regardless of evaluations* A_{*k} *reported by* k. *Thus, the grade of student* k *is*

$$g_k = (1 - \varepsilon)t_k + \varepsilon \max(1 - E_k, 0)$$

and it is maximized if and only if $E_k = 0$, *which happens if and only if evaluations* k *reports are also perfectly consistent with the truth.*

Theorem 1 *The mechanism defined by* (7) *is incentive–compatible and reliable.*

Proof 4 *Incentive compatibility is Lemma 3. Reliability of our mechanism follows from Lemma 1.*

5 Practical considerations

5.1 Instructor's judgment

It is not realistic to expect that actual students will be completely truthful and accurate in their evaluations. It is therefore desirable to have some procedure of discrediting evaluations that are not trustworthy.

Our solution is that students not only give numeric evaluations to each other, but also provide justifications, i.e., write short explanations on why they gave a particular evaluation. The instructor will then read these explanations and give each a score for trustworthiness. Let the instructor's score for justification of evaluation given to student i by student j be w_{ij}.

The scores w_{ij} are then used in the updated definition of the auxiliary matrix B as follows. We begin with (4), the familiar definition of vectors of relative contribution r_{ij}, but now we introduce weights into it. The relative contribution of students i and j according to k, i.e.,

$$\frac{a_{ik}}{a_{ik} + a_{jk}}$$

is now an entry of r_{ij} with weight $w_{ik} + w_{jk}$. Equation (5) that defines b_{ij} via medians of r_{ij} will now use weighted medians instead. The rest of our mechanism, i.e., the definition of the auxiliary matrix B, the vector of perceived evaluations s, the errors E_j of evaluations reported by j, and the vector of final grades g remain unchanged.

In practice, we use non-negative integers as scores given by the instructor to students' explanations of their evaluations, which means that the weighted median is simply the median of the vector obtained from r_{ij} by repeating each of its entries as many times as the weight of the entry. However, this is not important and other instructors can use non-integer weights if they think it is more appropriate.

5.2 Missing values

In practice, students may fail to submit evaluations of some or all of their teammates, i.e., an actual matrix A may have missing values. If student j did not submit evaluations at all, then the instructor will just set $w_{ij} = 0$ and $a_{ij} = 1$ for all i, i.e., grades will be calculated as if student j reported that all team members contributed equally to teamwork, but her evaluations are just ignored in the calculation of the vector of perceived contributions. If student j submitted evaluations of some but not all of her teammates, then the instructor will set missing evaluations to be the average of existing evaluations with, again, zero scores for trustworthiness. Our mechanism remains incentive compatible if at least 5 and reliable if at least 3 students submit their evaluations.

5.3 Coalitions

In practice, coalitions may occur. For instance, if two friends are in the same team, they may be tempted to give each other unfairly high evaluations. Our assessment mechanism prevents this behaviour on several levels.

First, from the proof of Lemma 2, it is clear that its stronger version actually holds — namely, in a team of n students, if $n - c$ are perfectly consistent in their evaluations, then the remaining c can't change the vector s of perceived contributions if $c \leq \frac{n-2}{2}$. It means that teams of 7 or 8 students are resistant to coalitions of size 2, teams of 9 or 10 students are resistant to coalitions of size 3 etc.

Second, students know that a part of their final grade is given for consistency of evaluations they report with the vector of perceived truth. It means that two friends are explicitly discouraged from giving each other unfairly high evaluations because they know that by giving each other too generous evaluations they will sacrifice a part of their grades. At the same time, their mutual evaluations a_{ij} and a_{ji} won't affect the median of r_{ij}, i.e., they can't be confident that their evaluations of each other will actually increase their grades.

Third, if the instructor suspects a coalition, he can reduce the weight of evaluations reported by students involved in the coalition.

6 Empirical results

6.1 *Educational setup*

Between 2017 and 2021, the author taught a number of relatively large undergraduate courses where a main assessment component (between 35% and 50% of the total course mark) was a team project. Teams had 5 or 6 students. Some teams were formed by students themselves and some by the author. Most students majored in mathematics.

Individual contribution to team project was assessed by some version of mechanism (7). The author also graded the final product of teamwork, i.e., a report and a presentation, according to a rubric. Final individual grade to student i for the project was given by the formula

$$0.9 \times (q + (1-q)s_i) \times T + 0.05 \times e_i + 0.05 \times \max(1 - E_i, 0),$$

where s_i is the perceived contribution of student i calculated as described in section 4.3, E_i is the relative error of evaluations submitted by i defined by (6), e_i is the average score given by the instructor to student i for providing justifications for evaluations that i reported, T is the score given to the whole team according to the rubric, and q is a hyperparameter defining the trade-off between two extremes — all team members get the same score and team members who contribute very little get a near-zero score. The value of q in the courses taught by the author varied between 0 and 0.5, depending on the university's current policies.

We anonymized the data by assigning random identifiers to courses, students, and teams. The summary is given in Table 1.

Table 1. Information on the courses taught by the author the data were collected from — the number of students, the number of teams, and the number of records (evaluations) for each course are given.

	course ID	Students	Teams	Records
1	HV	210	38	1170
2	LS	175	32	965
3	QN	210	38	1170
4	RS	163	30	895
5	TD	190	36	1016
6	XY	253	46	1403
	Total	1201	220	6619

6.2 *Data*

We collected peer evaluations in each team, normalised the matrix of peer evaluations dividing each column by its mean, and computed the vector of perceived truth in each case. The full dataset is available here: http://peerdata.kdl.me/

6.3 *Statistics*

Do we actually need a non-trivial method for assessing individual contributions to teamwork? Let us look at statistics. First, 28 out 220, or 13% of teams in our undergraduate courses submitted all equal evaluations ($a_{ik} = a_{jk}$ for all i, j, k), i.e., all students in those teams were fine with getting equal grades for the project. Out of all 1201 students, 11, or 1%, contributed nothing to teamwork, i.e., their perceived contribution was 0.

Further, there are 192 teams who reported non-equal evaluations. Out of 1048 students in those teams, 134, or 13% still reported all equal evaluations (i.e., $a_{ik} = a_{jk}$ for all i, j and for a given k); 46, or 4% reported incomplete evaluations (at least one a_{ik} is undefined for a given k); and 30, or 2% did not report peer evaluations at all (all a_{ik} are undefined for a given k). Now if we exclude all students who reported equal peer evaluations and students who missed at least one evaluation, there still remain 886, or 72% of 1201 students who have put some serious thought into the peer evaluation exercise.

It means that some method for grading individual contributions to teamwork is very important for educational practice.

6.4 *Consistency of empirical evaluations*

To understand to which extent our model of teamwork is applicable to real life, let us do some data analysis. Our "ideal" assumption is that there exists a vector t of objective truth whose entries are actual contributions of all the team members to teamwork and that the vector t is known to students. If it were the case and if all students reported the truth, then all matrices of peer evaluation would have rank 1. Of course, this is not observed in practice — everyone has their own version of the truth, which leads to all sorts of disagreements between evaluators.

For every two students from the same team, we can measure how well their versions of truth agree. To do it, we compute the Spearman correlation coefficient between evaluations reported by them. The Spearman correlation coefficient measures the strength of monotonic relationship between two variables. It equals 1 if and only if two students rank all team members from highest to lowest in exactly the same order and -1 if and only if two students rank all team members in the opposite order.

Our dataset contains 1704 pairs of evaluators for whom the Spearman correlation coefficient can be computed. This excludes students who submitted all equal evaluations, students who did not complete peer evaluation, and self-evaluations. The histogram of empiric correlations is shown in Figure 1.

We can tell that a good number of evaluations are quite consistent. To interpret Figure 1, let us think of a simplified scenario, in which some students try to fairly and honestly evaluate their team members' contributions to teamwork and the rest of the students just submit some rubbish evaluations. Let's say that we expect the correlation between honest evaluators to be 0.8 or above and that the fraction of honest evaluators is q. Then the fraction of correlations that are 0.8 or above will be q^2. The empirical fraction of correlations that are 0.8 or above is 0.54, i.e., 73% of students are honest evaluators.

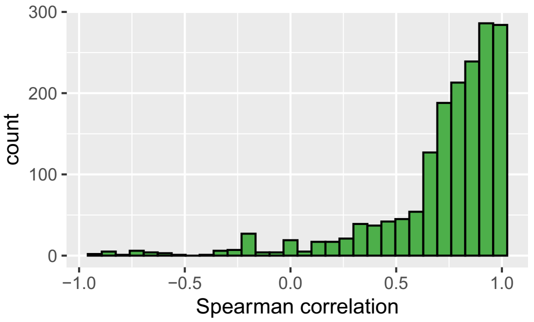

Figure 1. Empirical distribution of the correlation coefficient between pairs of different evaluators.

The fraction of 73% of honest evaluators found here is consistent with the 72% of students who submitted non-equal evaluations found in the previous section. In either case, it seems that at least 70% of students take the peer evaluation exercise seriously.

6.5 *Justifications*

We also required students to justify their evaluations by writing a short testimonial that reflects on the contribution of the evaluatee, her strengths, and suggestions for improvement. Here is an example of such a justification (it comes with a score of 0.8, i.e., below average):

> [NAME] has played his part by contributing to the introduction and literature review portion of the project. He has been proactive in discussions and tries his best to help us out if we have any doubts. He has good interpersonal skills and is a comfortable team member to work with. However, one area of improvement could be diligence as he could have put in more effort in his work.

The instructor read all such justifications and assigned them scores for trustworthiness according to the following rule: 5 out of 10 if there is an explanation of the contribution of the student who is evaluated, 3 for strengths, and 2 for suggestions for improvement. It means that the score for the testimonial cited above was 10 out of 10.

To scale our method of assessing individual contribution to teamwork to very large classes, including massive open online courses, one can train an AI that will assign trustworthiness scores to justifications of peer evaluations automatically. This is, however, a topic for a different study.

7 Discussion

We have introduced a mathematical model of teamwork with the main objective of developing a valid, reliable, and practical assessment method of individual contribution to teamwork. The main assumption is that there exists an objective truth — a vector t whose entries measure individual contributions to teamwork precisely. Our empirical results show that this idealistic assumption is not too far from reality.

We defined mathematically desirable properties of an assessment method, namely, incentive compatibility (called validity in psychometrics) and reliability. Then we have argued that existing assessment methods that are widely used in actual classrooms do not satisfy these properties.

We then introduced a new assessment method that is incentive–compatible and reliable for teams of at least 5 students such that at most one of the students contributes nothing to teamwork. We proved that our assessment method is incentive–compatible by showing that the collective truth–telling is a strict Nash equilibrium. Since the objective truth is unobservable, it means that any matrix of peer evaluations of rank 1 is a strict Nash equilibrium. However, we do not know if other Nash equilibria exist and we do not know if incentive–compatible assessment methods exist if the number of students in the team is 3 or 4.

References

[1] Praveen Aggarwal and Connie L O'Brien, 'Social loafing on group projects: Structural antecedents and effect on student satisfaction', *Journal of Marketing Education*, **30**(3), 255–264, (2008).

[2] Noga Alon, Felix Fischer, Ariel Procaccia, and Moshe Tennenholtz, 'Sum of us: Strategyproof selection from the selectors', in *Proceedings of the 13th Conference on Theoretical Aspects of Rationality and Knowledge*, pp. 101–110, (2011).

[3] Haris Aziz, Omer Lev, Nicholas Mattei, Jeffrey Rosenschein, and Toby Walsh, 'Strategyproof peer selection: Mechanisms, analyses, and experiments', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, (2016).

[4] James R Beatty, Robert W Haas, and Donald Sciglimpaglia, 'Using peer evaluations to assess individual performances in group class projects', *Journal of Marketing Education*, **18**(2), 17–27, (1996).

[5] Charles M Brooks and Janice L Ammons, 'Free riding in group projects and the effects of timing, frequency, and specificity of criteria in peer assessments', *Journal of Education for Business*, **78**(5), 268–272, (2003).

[6] Arthur Carvalho and Kate Larson, 'Sharing rewards among strangers based on peer evaluations', *Decision Analysis*, **9**(3), 253–273, (2012).

[7] Robert Conway, David Kember, Atara Sivan, and May Wu, 'Peer assessment of an individual's contribution to a group project', *Assessment & Evaluation in Higher Education*, **18**(1), 45–56, (1993).

[8] Geoffroy De Clippel, Herve Moulin, and Nicolaus Tideman, 'Impartial division of a dollar', *Journal of Economic Theory*, **139**(1), 176–191, (2008).

[9] Fedor Duzhin and Amrita Sridhar Narayanan, 'Peer grading reduces instructor's workload without jeopardizing student learning in an undergraduate programming class', *New Directions in the Teaching of Physical Sciences*, (2020).

[10] Alison Laver Fawcett, *Principles of assessment and outcome measurement for occupational therapists and physiotherapists: theory, skills and application*, John Wiley & Sons, 2013.

[11] Norma C Holter, 'Team assignments can be effective cooperative learning techniques', *Journal of education for Business*, **70**(2), 73–76, (1994).

[12] William B Joyce, 'On the free-rider problem in cooperative learning', *Journal of Education for Business*, **74**(5), 271–274, (1999).

[13] Deborah B Kaufman, Richard M Felder, and Hugh Fuller, 'Accounting for individual effort in cooperative learning teams', *Journal of Engineering Education*, **89**(2), 133–140, (2000).

[14] Jane H Leuthold, 'A free rider experiment for the large class', *The Journal of Economic Education*, **24**(4), 353–363, (1993).

[15] Omer Lev, Nicholas Mattei, Paolo Turrini, and Stanislav Zhydkov, 'Peernomination: A novel peer selection algorithm to handle strategic and noisy assessments', *Artificial Intelligence*, 103843, (2022).

[16] S Dolly Malik and Daniel R Strang, 'An exploration of the emergence of process prototypes in a management course utilizing a total enterprise simulation', *Developments in Business Simulation and Experiential Learning*, **25**, (1998).

[17] John R McClure, Brian Sonak, and Hoi K Suen, 'Concept map assessment of classroom learning: Reliability, validity, and logistical practicality', *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, **36**(4), 475–492, (1999).

[18] Kenneth O Morse, 'International management virtual teamwork: A simulation', *Developments in Business Simulation and Experiential Learning*, **29**, (2014).

[19] Barbara M Moskal and Jon A Leydens, 'Scoring rubric development: Validity and reliability', *Practical assessment, research, and evaluation*, **7**(1), 10, (2000).

[20] Catherine A Palomba and Trudy W Banta, *Assessment Essentials: Planning, Implementing, and Improving Assessment in Higher Education. Higher and Adult Education Series.*, ERIC, 1999.

[21] David C Parkes, Ariel D Procaccia, and Nisarg Shah, 'Beyond dominant resource fairness: Extensions, limitations, and indivisibilities', *ACM Transactions on Economics and Computation (TEAC)*, **3**(1), 1–22, (2015).

[22] Victor Shnayder and David Parkes, 'Practical peer prediction for peer assessment', in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, pp. 199–208, (2016).

[23] Precha Thavikulwat and Jimmy Chang, 'Pick your group size: A better procedure to resolve the free-rider problem in a business simulation', *Developments in Business Simulation and Experiential Learning*, **37**, (2014).

[24] James R Wright, Chris Thornton, and Kevin Leyton-Brown, 'Mechanical ta: Partially automated high-stakes peer grading', in *Proceedings of the 46th ACM Technical Symposium on Computer Science Education*, pp. 96–101, (2015).

[25] Hedayat Zarkoob, Greg d'Eon, Lena Podina, and Kevin Leyton-Brown, 'Better peer grading through bayesian inference', *arXiv preprint arXiv:2209.01242*, (2022).