Peer reviewed version

University of Bristol - Explore Bristol Research
General rights

Quantitative analysis of approaches to group marking

Hugh Harvey[a,c], James Keen[a,c,*], Chester Robinson[a,c], James Roff[a,c] and Thilo Gross[b]

[a]University of Bristol, Department of Engineering Mathematics, Woodland Road, Bristol, UK;

[b]University of California at Davis, Department of Computer Science, Kemper Hall, Shields Blvd. Davis, CA 95616, USA

[c]These authors contributed equally

ABSTRACT
Group work, where students work on projects to overcome challenges together, has numerous advantages, including learning of important transferable skills, better learning experience and increased motivation. However, in many academic systems the advantages of group projects clash with the need to assign individualised marks to students. A number of different schemes have been proposed to individualise group project marks, these include marking of individual reflexive accounts of the group work and peer assessment. Here we explore a number of these schemes in computational experiments with an artificial student population. Our analysis highlights the advantages and disadvantages of each scheme and particularly reveals the power of a new scheme proposed here that we call pseudoinverse marking.

KEYWORDS
Quantitative; group marking; group projects; individual marks

Introduction

Group projects, where a small group of students tackles a challenge, can enhance the learning experience (Abelson and Babcock 1986; Hmelo-Silver 2004). This form of teaching encourages students help each other and learn to overcome problems together. This is beneficial as it not only allows students to develop technical skills but also transferable skills in project management and leadership (Hmelo-Silver 2004; Davies 2009).

To an assessor, group projects can present a considerable challenge (Fellenz 2006; Davies 2009). Many academic systems require the assessor to assign an individualised mark to each student participating in the course. These marks must be fair and unbiased. However, marks should also take the transferable skills that the student acquired into account (Jackel et al. 2017). While the technical progress made by a group is generally easy to assess, e.g. on the basis of a project report, it is more difficult to judge the technical and collaborative skills of individual members of a group. In a typical setting, each student works on several shorter projects with different groups. The challenge is then to compute a student's individualised mark from the set of group marks received.

A simple approach is to assign to each student the average of the marks of the groups in which the student participated. As we show below, this has the undesirable effect of reducing the variance of the marks, producing a cluster around the mean. However, several alternatives exist. The assessors can individualise marks based on their own supervision of the project or they can ask the students to write reflexive accounts detailing the individuals contribution to the project, which can then be taken into account in the marking (Dyment and O'Connell 2011).

A recent review of group marking (Dijkstra et al. 2016) noted that the most common approach to individualised marking is peer assessment. Students are asked to mark the other members of the groups they participated in, by assigning actual marks, e.g. (Abelson and Babcock 1986; Fellenz 2006), or grouping them in categories, e.g. (Lejk and Wyvill 2001). It has been noted that peer marking is an effective tool if used appropriately, but there can be some flaws when dealing with students who are either particularly generous or harsh with their marking style (Goldfinch 1994).

Other options include self-assessment, where you ask the student within the group to perform an additional piece of work and assign a portion of their mark based on that. It is noted that, while self-assessment is a simple approach, it often deviates from the markers assessment significantly, partly due to factors beyond the scope of the project (Sharp 2006; Suñol et al. 2016).

An interesting scheme proposed in (Ko 2014) reveals another alternative, here called Iterated Individual Weighting Factor (IWF-it). This method involves peer assessment as above, but then adjusting the weighting of each group member's opinion depending on how close to the true group mark they predicted for their group. This approach does attempt to discount potential bias in the peer assessments by inaccurate individuals, but is computationally expensive.

There has been an attempt to simplify the process of marking group projects by the Australian Learning and Teaching Council. This resulted in a software package called SPARK[PLUS] (Self and Peer Assessment Resource Kit) (Freeman and McKenzie 2002). This software has been tested in a university-level engineering setting (Wu, Chanda, and Willison 2014). The test showed that, while the idea was sound, a high proportion of the test subjects were dissatisfied with the program's attempt to distribute the group marks.

Here, we consider the effectiveness of different approaches through quantitative analysis of fairness. Our analysis highlights two promising approaches. First, a particular peer assessment method in which each group member is asked to assess the value of each member of the group, in order to then fairly divide the marks between them. While there is a risk with this approach that marks can be manipulated by coordinated responses, it otherwise yields fair results from a transparent and easily understood procedure. Second, pseudoinverse marking can be implemented to eliminate the need for peer assessment. This approach uses the marks of each individual student for a series of projects. Then the best estimate of the mark the student should receive is computed. This approach avoids the risk of strategic coalitions and the administrative effort of peer assessment at the cost of transparency.

Methods

In the following, we test various marking schemes by applying them to a virtual student population. We numerically generate populations of N students. Each individual i is assigned a number which represents their `ideal mark' q_i, i.e. the mark that they should be awarded. We draw the ideal marks from a Gaussian

distribution with mean mark of 60 and standard deviation of 12. This mimics the UK academic system.

After the virtual population has been created, we assume that each student undertakes p projects in groups of size m over the course of the unit. The overall mark for each student is then found from the marks of the groups the student participated in, and potentially some other information, such as peer assessment.

Without much analysis, it is intuitive that having larger groups (greater m) makes it harder to determine accurate individualised marks, whereas having more group marks for each student (greater p) makes it easier. We verified the former is valid for most schemes considered. As the number of projects increases, the individual mark predictably leads to an averaging, with the exception of the final scheme. For the analysis presented here, we vary group size and number of projects per student together, considering specifically the case $m = p$. Hence, the results can be presented more concisely.

In our virtual population simulation, we assume that the group mark w_j for group j is the mean of the ideal marks of the group's participants. Mathematically, the mean is expressed as

$$\omega_j = \frac{\sum_i M_{ij} q_i}{n_j} \tag{1}$$

where $n_j = \sum_i M_{ij}$ is the number of participants of project j, and $\boldsymbol{M}$ is the participation matrix which is defined by

$$M_{ji} = \begin{cases} 1 \text{ if student } i \text{ participated in project } j \\ 0 \text{ if student } i \text{ did not participate in project } j \end{cases} \tag{2}$$

For illustration, consider the matrix

$$\boldsymbol{M} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix} \tag{3}$$

This matrix describes the partitioning of 4 students into 4 groups, such that student 1 and student 2 participate in group 1 (first row), student 3 and student 4 participate in group 2 (second row), student 1 and student 3 participate in group 3 (third row) and student 2 and student 4 participate in group 4 (fourth row).

Once the group marks have been determined, we apply a set of different marking schemes (explained below) to assign individualised marks x_i to the students. We analyse the accuracy of the marking schemes first by considering their performance in a scenario where the size of the population $N = 52$ and group size $m = 4$. For this scenario, we can draw scatter plots showing the individualised mark x_i of a student against their ideal mark q_i. Furthermore, we study the accuracy as a function of the variables n and m. For this comparison, we quantify the accuracy in terms of the maximal absolute error

$$E_{max} = max_i |q_i - x_i| \tag{4}$$

the mean absolute error

$$E_{mean} = \sum_i \frac{|q_i - x_i|}{N} \tag{5}$$

and the root mean square error

$$E_{rms} = \sqrt{\sum_i \frac{(q_i - x_i)^2}{N^2}} \tag{6}$$

Among these, the mean absolute error provided a measure of the overall accuracy, whereas the maximum error provides an estimate of `unfairness' by focussing on the most unfair case. The mean square error is an intermediate between these two extremes, averaging over the population but assigning a higher weight to individual marks that deviate strongly.

Results

We now compare six different schemes for generating individualising marks. Starting from the simplest and building up to the most mathematically complex. The first is simply to assign the group mark to all the group members (Sec. 3.1), which provides a baseline or null-model against which other methods can be judged. We then consider using additional information such as reflexive accounts (Sec. 3.2), before proposing an improvement on this scheme that we call mark-adjusted reflexive accounts (Sec. 3.3). Subsequently, we consider two methods for peer marking, normalised peer assessment (Sec. 3.4) and peer ranking (Sec. 3.5). Peer ranking is a scheme we introduce based on an approach that has been proposed for sports rankings (Park and Newman 2005; Motegi and Masuda 2012). Finally, we propose a mathematical method that aggregates results from different projects to infer individualised marks using the Moore-Penrose pseudoinverse (Sec. 3.6).

Self-organised peer pressure (SOPP)

In our simplest scheme, the final mark for a student is the average of the marks of the groups that the students participated in, expressed as

$$x_i = \sum_j \frac{M_{ij} w_j}{n_i} \tag{7}$$

where $n_i = \sum_j M_{ij}$ is the number of students in each group. We call this scheme the self-organised peer pressure (SOPP) method because every student has a direct interest in the success of their groups. Other marking schemes might create secondary objectives such as to improve standing within the group (possibly at a cost to other group members or group success) to optimise outcomes from peer assessment.

We note that even the very simple SOPP method leads to individualised marks unless multiple students participate in exactly the same groups.

The disadvantage of this method is that it leads to a `regression to the mean'-type of effect, where each student is assigned a mark close to the class average. This can be seen in (fig. 1) by plotting the expected assigned mark versus the ideal mark for a group of students. The best students receive on average less than their ideal mark whereas weaker students receive more than their ideal mark. This result is intuitive as, under this scheme, good students suffer from being grouped with students that are on average weaker than themselves. Conversely, weak students benefit from being groups with better students.

Considering the effect of students participating in more, larger projects (fig. 2) shows that average errors saturate. However, this saturation occurs at a relatively high level where the root mean square error reaches almost 10 percentage points, which corresponds to a whole degree classification in the British academic system. Perhaps more worrying is that the maximum error is high and keeps increasing, indicating that SOPP is increasingly unfair to some individual students.

In summary, the advantage of the SOPP scheme is that it is very easy and intuitive. This makes the scheme easy to apply and marks are found by a highly transparent procedure. Furthermore, it creates incentives that are well aligned with the spirit of project work. For each student, the only way to improve their marks is

to improve the mark of the groups they participate in. This mimics a typical workplace situation where ultimately the success of a project matters, rather than the contributions of individual team members.

The good alignment of the assigned mark and project goals is undermined by the large error in the marks, which is approximately the difference between a grade classification. One may be tempted to rescale the marks to address the regression-to-the-mean, resulting in a narrow distribution of marks. However, such a rescaling would exacerbate the individual errors to levels that could be judged unacceptable.

Reflexive accounts (RA)

A common approach to individualise marks is to pair group projects with individual assignments such as reflexive accounts on the project. Here we assume that for every group project a student participates in, they also carry out an individual assignment. We assume that the assessor does not arrive at the student's ideal mark when marking the individualised component. This represents the student performing better or worse for a reflexive report than in their group report, and possible subjectivities in the assessors marking. The assessor instead arrives at the students ideal mark with an error, drawn from an uniform distribution with a range of ± 16.

In the simplest case, considered in this section, the mark that the student receives in a given project is then found as a linear combination of the group mark and the mark for the individual assignment. Based on personal experience and some preliminary trials, we focus on the case where the group mark (w_j) enters with weight $\alpha = 0.7$ and the individualised component (y_{ij}) enters with weight $1 - \alpha = 0.3$. This means that, taking multiple projects into account, the final mark of a student is computed as

$$x_i = \frac{1}{n_i}\sum_j \alpha\, w_j M_{ij} + (1-\alpha)y_{ij} = \frac{0.7 \cdot \sum_j M_{ij}w_j + 0.3 \cdot \sum_j y_{ij}}{n_i} \tag{8}$$

where $y_{ij}(q_i)$ represents the mark assigned to student i on project j for the reflexive piece of work, depending on their ideal mark q_i. As the final mark is to a significant proportion based on the ideal mark, we expect better overall performance of this scheme. The `regression to the mean' that we identified as the major problem of the SOPP method is ameliorated but not eliminated (fig. 1). Also considering the participation in larger groups shows a quantitatively better, but qualitatively similar picture to the SOPP method.

In summary, the RA approach offers few surprises. It uses a linear combination of the SOPP method with the ideal result and hence produces results that interpolate between the SOPP outcome and the ideal outcome. The advantage of the RA scheme is that we can somewhat ameliorate the poor performance of the SOPP scheme, while largely maintaining transparency and mathematical simplicity. However, the improvement comes at the cost of abandoning the spirit of project work to some extent as the project outcome is supplemented by a secondary objective which is an individual written assignment. We thus lose some of the beneficial alignment of the project with real-life workplace scenarios and introduce significant non-project workload for both the student and the assessor.

Changing the weighting of group and individualised components in the marking gives some control over the balance between the advantages of project work and marking fairness. What makes this scheme somewhat unsatisfactory is that the trade-off remains linear. Introducing a small individualised component only marginally improves mark accuracy. To significantly improve accuracy requires sacrificing most of the advantages of project-based assessment.

<u>Mark-adjusted reflexive accounts (MRA)</u>

We now explore an alternative use of reflexive accounts, where the reflexive account is instead used to judge the individual's contribution to the project. We make the same assumptions as previously. We then estimate the contribution of student i to project j by

$$c_{ij} = \frac{y_i}{\sum_k y_k M_{kj}} \, . \tag{9}$$

The total number of mark points t_j that we can distribute for project j is the project mark w_j times the number of students involved in the project n_j, i.e.

$$t_j = w_j n_j \, . \tag{10}$$

Instead of distributing these points equally we now partition them according to the estimated contribution to the project, such that student i receives for project j their contribution multiplied by the marks available. Combining this yields

$$r_{ij} = c_{ij} t_j = \frac{w_j y_i \sum_k M_{kj}}{\sum_k y_k M_{kj}} \, . \tag{11}$$

The final mark for the student after participating in multiple projects is then the mean of their mark for each project

$$x_i = \frac{\sum_j M_{ij} r_{ij}}{n_i} \, . \tag{12}$$

[Figure 1 about here.]

[Figure 2 about here.]

Considering the outcomes of the computational experiment we see that good students still receive marks that are systematically less than their ideal mark (fig. 1). Compared to the simpler RA scheme, the performance of MRA is almost identical. This is also confirmed when considering different group sizes and project iterations (fig. 2). A small difference exists for very small group sizes, but this is mostly cosmetic. Consider that, under the MRA scheme, the reflexive account does not affect the marking at all. Thus any error made in the marking of the reflexive account does not affect the student's mark. So this difference between the schemes appears due to specific assumptions of the computation experiment and is of little practical relevance.

Perhaps more significant is another advantage of the MRA scheme that is not considered in the computational experiment. This is the case where one student in an otherwise well-performing group does not engage with the project at all.

Consider the (somewhat extreme but not unheard of) example where a student in a group of 4 does not do any work for the project or the reflexive account. Regardless, the project still achieves 60. Under the assumptions made in this report the average ideal mark for the other 3 students on the project would have to be 80 to compensate for the work not done by the defecting student. Assuming that the 3 engaged students achieve on average the same level of marks in their reflexive accounts, their mark under the RA scheme would be

$$0.7 \cdot 60 + 0.3 \cdot 80 = 42 + 24 = 68 \, . \tag{13}$$

while the defecting student receives

$$0.7 \cdot 60 + 0.3 \cdot 0 = 42 \, . \tag{14}$$

In the British system, the defecting student still passes while the very strong, engaged students only receive a disheartening second class mark. By contrast, in

the MRA scheme the engaged students would receive on average their ideal mark of 80, a solid first class result. The defecting student receives the zero mark they deserve.

Even if the defecting student decides to invest effort in the reflexive account, but not the project, the student would need to achieve a mark of 51 in the reflexive account to receive the same mark of 42 that would be assigned under the RA scheme. We note that the MRA scheme can mean that students can fail although both their reflexive account and their overall group mark are sufficient to pass. However, this is unlikely to happen unless a student only participates in groups where all others receive vastly higher marks on the reflexive accounts.

In summary, the MRA scheme is slightly more complex than the RA scheme, but the calculations are still simple enough to be carried out very quickly with pen and paper, a calculator or a simple spreadsheet. On average, the improvement of performance over the RA scheme does not seem to justify the extra complexity of the MRA. However, MRA is superior in the extreme case where students do not engage with the project at all. Because the scheme is still quite transparent, this means that MRA can create strong incentives for the students to engage with the projects. We caution it is prudent to be conservative in marking reflexive accounts using this scheme and to avoid extreme marks unless they are clearly indicated.

Normalised peer assessment (NPA)

In peer assessment, the students assess the performance of each member of their group. The immediate advantage of this method is that information on the relative contributions to the project is sourced directly from the students involved. The obvious drawback is that the students are given a way to influence their mark other than delivering a good project outcome. Thus; there is a risk that students use peer assessment strategically to maximise their own marks rather than to provide genuine information about project contributions.

A variety of different peer assessment schemes have been proposed to address these advantages and disadvantages (Spatar et al. 2014). To gain insights into the student's ability to manipulate their marks, one can distinguish between strategies that students can implement for themselves and those that require the formation of coalitions with other students. An individual strategy would be a student giving themselves a high mark and/or all other group members a low mark to maximise their outcome. Alternatively, a coalition of three members of a group of four could conspire to increase their marks at the expense of the fourth student.

The topic of coalition formation in groups is very complex and subject of active research. However, a simple argument extending the coalition example above can be made to show that peer assessment can fail if multiple students conspire to mark others strategically. In the not very far-fetched case above, the three conspirators could mark each other highly while assigning a zero mark to the fourth student. In the absence of any other sources of information, this situation is indistinguishable from a scenario in which one of the students did not engage with the project at all. Thus, no peer marking scheme can deliver a satisfactory outcome in both of these scenarios at the same time.

Here, we particularly consider a normalised peer marking scheme (Spatar et al. 2014). In this scheme, every member of a group marks all other members, but not themselves. Not allowing students to mark themselves prevents inflating their own mark, the most direct form of mark manipulation.

In the next step the marks assigned by each student are normalised. For example, let S_{ki} be the mark that student k assigned to student i. The mark is normalised by dividing all marks by the total marks student k assigned to their group members. The normalised marks,

$$s_{ki} = \frac{S_{ki}}{\sum_i S_{ki}}. \tag{15}$$

are in the range [0,1] and reflect the student k's opinion of the proportional contribution of the other group members. Taking only these normalised marks into account prevents students from inflating their own contribution by marking all others contributions lowly.

To determine the mark r_{ij} that each student i receives for a project j, we first compute the total mark points t_j as in the MRA scheme above (Eq. 10). Then these points are distributed proportionally to the normalised marks a student has received. This mark is calculated by

$$r_{ij} = \frac{t_j \sum_k s_{kj}}{n_j}. \tag{16}$$

which includes normalising the mark using the number of students in the group, n_j.

In numerical simulations, each student is assumed to rank another student based on their ideal mark with some random error. This random error represents students not accurately judging the ability of other group members. As with similar errors for other schemes, a uniform distribution with a range of ± 16 was used. Under these assumptions, the method achieves good results, particularly as there is no systematic bias against stronger students (fig. 1). Unlike all other schemes in this report, the error of this method decreases when group size is increased. This is intuitive as the assigned mark is computed from a larger number of observations, and thus profits from a `Wisdom of the Crowd' effect (Crosscombe and Lawry 2017). Additionally, larger groups make effective coalitions harder to form.

While the performance of this method in our test is very good, the clear drawback is a vulnerability to manipulation by coalition formation. One could argue that this can be somewhat mitigated by adjusting the mark if indications of a coalition formation within a group exists (e.g. from observation of the group work). However, allowing for such adjustments to some extent defeats the purpose of having a transparent procedure. We therefore recommend this method particularly for projects with large group sizes; where the method performs particularly well and with reduced impact of small coalitions.

Peer ranking (PR)

We now turn to two unusual methods. For sports rankings, network-based approaches have recently received attention (Park and Newman 2005; Motegi and Masuda 2012). A similar approach can be taken for peer marking. Instead of marking each group member with a precise mark, students rank the other members of their group. The ranking consists of a pairwise comparison of the perceived contribution of every pair of other members. We can interpret each such ranking as a link in the network leading from the weaker to the stronger student. Figure 3 shows a complete example, detailing each pairwise comparison.

[Figure 3 about here.]

Network structures can be encoded in matrices. Here we use a modified adjacency matrix A, defined by

$$A_{ik} = n_a a + n_b b \,. \tag{17}$$

where n_a is the number of times that student i was ranked lower than student k, n_b is the number of times that student i was ranked higher than student k. Finally, a and b are constant parameters such that $b < a$ so greater importance is placed on being ranked higher. Figure 3 also shows the adjacency matrix, A, for the previous example.

The advantage of the matrix notation is marks can be generated using a spectral approach, similar to the famous PageRank algorithm (Page et al. 1999). We compute the leading eigenvector v of A and normalise it such that the elements sum to the group size so $\sum v_n = n_j$.

After the normalisation, the entries of the eigenvector are in the range $[0, n_j]$ with a mean of one. This mean of one constrains the average of the individual assigned marks being equal to the group mark. The n-th entry v_n is a proxy for the relative contribution for student n in the group.

We then use the eigenvector entries to personalise the marks such that the final mark that student i receives for the project j is

$$r_{ij} = w_j(\alpha + (1 - \alpha)v_n) \,, \tag{18}$$

where $n = n(i, j)$ is index assigned to student i within group j and α is another scalar parameter used to control the weight attributed to the peer ranking.

Based on preliminary tests, we used the parameters $a = 0.25$, $b = 1$ and $\alpha = 0.65$. We assumed that students rank each other student in order of a perceived mark, which is the ideal mark and a random error. Like for the normalised peer assessment method (Sec. 3.4), this random error is drawn from a uniform distribution with range ± 16. These perceived marks are used to rank each other student, and then only the set of ranked lists from each student is used.

The results of our computational experiment show that these choices of parameters receive a relatively low absolute error, but has a slight systematic bias favouring weaker students. This bias could be reduced by reducing α, albeit at the cost of increasing non-systematic error. Individual error appears primarily in groups of equal or almost equal strength where the ranking method amplifies the small differences. One may suspect that this is a lesser problem in reality where students are not able to rank each other perfectly based on tiny differences. However, a detailed investigation of this point would likely require a study of ranking behaviour of real students, which is beyond the scope of the present paper.

In summary, the peer ranking scheme performs slightly worse than the NPA method (particularly in case of large group sizes). It is also considerably less transparent. We therefore judge that the NPA method will be superior in most cases. However, this method is in its infancy and so presents a promising area for further improvements.

Pseudoinverse Marking (PiM)

All methods considered so far were applied to one project group at a time. However, in a setting where students participate in multiple projects, additional information can be gained by considering how they perform in groups of different compositions.

We continue working on the assumption that the mark of a project report reflects the average ideal mark of the participants in the project, i.e.

$$w_j = \frac{\sum_i M_{ij} q_i}{n_j}. \tag{19}$$

In vector notation, this can be written as

$$\boldsymbol{w} = \boldsymbol{Q}\,\boldsymbol{q}. \tag{20}$$

where the matrix $\boldsymbol{Q}$ is defined by $Q_{ij} = M_{ij}/n_j$. In marking, we have determined the project marks $\boldsymbol{w}$, and we know the $\boldsymbol{Q}$ matrix as it follows from the partitioning of students into groups. Our aim is to compute the ideal marks $\boldsymbol{q}$ that the students should receive.

Formally, we can compute $\boldsymbol{q}$ by multiplying by $\boldsymbol{Q}^{-1}$ (the inverse of $\boldsymbol{Q}$) which yields

$$\boldsymbol{q} = \boldsymbol{Q}^{-1}\boldsymbol{w}. \tag{21}$$

If students were partitioned into groups such that $\boldsymbol{Q}$ is invertible, this relationship should yield the desired marks exactly.

In typical cases, another slight complication arises because common ways of dividing students into groups lead to singular matrices $\boldsymbol{Q}$ such that an inverse does not formally exist. In this case, the method can still be applied if we replace the inverse with the Moore-Penrose pseudoinverse (Barata and Hussein 2012). If each student participates in only a single project, then calculating the pseudoinverse of $\boldsymbol{Q}$ corresponds to implementing the SOPP method (Sec. 3.1). Clearly, this is undesirable so the number of projects each student participates in should be at least the number of students in the group. In an ideal case, every student completes multiple projects and interacts with a completely different set of people every time. In all cases, Eq. (21) can be used to deconvolute the contributions and recover the ideal mark using the pseudoinverse.

In our numerical experiment, we consider such a case where every student contributes to four projects and interacts with $4 \cdot 3 = 12$ distinct other students in the process. The assignment resulted in a singular $\boldsymbol{Q}$. The results (fig. 1) show that the estimated ideal mark computed with the Moore-Penrose pseudoinverse are in excellent agreement with the ideal marks. This method produces the smallest errors of all considered here.

One could criticise that the accuracy of the method relies on our assumption that the project marks achieved are the mean of the ideal marks of the project participants. However, this is actually not so much an assumption as a definition of what is meant by the ideal mark.

Apart from the high accuracy, this method has many other advantages. For example, it does not require any additional information (peer marking, ranking, accounts). Thus, the additional workload for both students and assessors is eliminated. Moreover, the method is robust and could be easily implemented, as the Pseudoinverse is supported by libraries in several programming languages.

[Figure 4 about here.]

Another advantage is that, unlike for any other method, the measured errors all decrease as the population size increases (fig. 4). Intuitively, as the algorithm can draw on more information, the accuracy of the assigned marks should increase. Additionally, for larger populations, more combinations of students for groups are possible, allowing a better assignment of groups with fewer repeated student interactions.

A drawback is that the final marks are only available after all projects have been completed. Also, students and assessors may find the method intransparent.

Perhaps the main disadvantage is that there is a minimum number of projects in which a student needs to participate. This de-facto limits the applicability of the method to projects that are carried out in small groups.

Conclusion

In this paper, we explore several schemes for individualising group project marks. The most suitable scheme depends on the specific circumstances, including the number of students and the number of projects each student participates in. An overview of the errors for a typical setting is shown in Table 1.

[Table 1 about here.]

The simplest scheme (SOPP), where students receive the average of the group marks for the projects that they participated in, systematically favours weaker students. Advantageously, no additional workload is generated and this scheme represents a real life workplace situation. Marking of reflexive accounts (RA) did not fully remove the systematic bias from the marks and has the disadvantage of generating significant additional workload. Among the two RA marking schemes, the mark-adjusted reflexive accounts (MRA) proposed here provides stronger incentive to engage with the project.

Normalised Peer Assessment (NPA) resulted in very accurate marks, particularly for projects with many members, but entails the risk of mark manipulation by coalition formation. A peer ranking scheme (PR) introduced here performed similarly but is significantly less transparent. Finally, the new pseudo-inverse marking (PiM) posed here achieves very accurate results without the risk of mark manipulation or additional workload. Contrarily, this scheme requires that the number of projects to which a student contributes is at least as large as the group size in the projects.

In absence of factors favouring a certain approach, our analysis highlights normalised peer assessment as the best scheme for projects with large group sizes. However, pseudoinverse marking is highlighted as the best scheme for marking a series of projects carried out by small groups.

This paper also illustrated how agent based computational models can be used to explore the fairness and accuracy of marking schemes. Here we have used only a very simple model, and plenty of opportunities for improvements and refinements still exist. For example, one could allow the students to allocate their time investment into the project strategically or build in social dynamics. These extensions are beyond the scope of the current paper. We hope that in the future more of this analysis will be carried out to yield deeper insights into the mathematical properties of group marking schemes.

References

Abelson, M. A., and J. A. Babcock. 1986. "Peer evaluation within group projects: A suggested mechanism and process." *Organizational Behavior Teaching Review* 10 (4): 98-100.

Barata, J., and M. Hussein. 2012. "The Moore-Penrose Pseudoinverse: A Tutorial Review of the Theory." *Brazilian Journal of Physics* 42 (1-2): 146-165.

Crosscombe, M., and J. Lawry. 2017. "Exploiting vagueness for multi-agent consensus." In *Multi-agent and Complex Systems*, 67-78. Springer.

Davies, W. M. 2009. "Groupwork as a form of assessment: common problems and recommended solutions." *Higher Education* 58 (4): 563-584.

Dijkstra, J., M. Latijnhouwers, A. Norbart, and R. A. Tio. 2016. "Assessing the I in group work assessment: State of the art and recommendations for practice." *Medical Teacher* 38 (7): 675-682. https://www.tandfonline.com/doi/full/10.3109/0142159X.2016.1170796.

Dyment, J. E., and T. S. O'Connell. 2011. "Assessing the quality of reflection in student journals: A review of the research." *Teaching in Higher Education* 16 (1): 81-97.

Fellenz, M. R. 2006. "Toward fairness in assessing student groupwork: A protocol for peer evaluation of individual contributions." *Journal of Management Education* 30 (4): 570-591.

Freeman, M., and J. McKenzie. 2002. "SPARK, a confidential web-based template for self and peer assessment of student teamwork: benefits of evaluating across different subjects." *British Journal of Educational Technology* 33 (5): 551-569.

Goldfinch, J. 1994. "Further developments in peer assessment of group projects." *Assessment & Evaluation in Higher Education* 19 (1): 29-35.

Hmelo-Silver, C. E. 2004. "Problem-based learning: What and how do students learn?" *Educational Psychology Review* 16 (3): 235-266.

Jackel, B., J. Pearce, A. Radloff, and D. Edwards. 2017. "Assessment and Feedback in Higher Education: A Review of Literature for the Higher Education Academy." *Higher Education Academy*.

Ko, S. 2014. "Peer assessment in group projects accounting for assessor reliability by an iterative method." *Teaching in Higher Education* 19 (3): 301-314.

Lejk, M., and M. Wyvill. 2001. "Peer assessment of contributions to a group project: a comparison of holistic and category-based approaches." *Assessment & Evaluation in Higher Education* 26 (1): 61-72.

Motegi, S., and N. Masuda. 2012. "A network-based dynamical ranking system for competitive sports." *Scientific Reports* 2: 904.

Page, L., S. Brin, R. Motwani, and T. Winograd. 1999. *The PageRank citation ranking: Bringing order to the web*. Technical Report. Stanford InfoLab.

Park, J., and M. E. J. Newman. 2005. "A network-based ranking system for US college football." *Journal of statistical mechanics* 10: P10014.

Sharp, S. 2006. "Deriving individual student marks from a tutors assessment of group work." *Assessment & Evaluation in Higher Education* 31 (3): 329-343.

Spatar, M., N Penna, H. Mills, V Kutija, and M. Cooke. 2014. "A robust approach for mapping group marks to individual marks using peer assessment." *Assessment and Evaluation in Higher Education* 40 (3): 371-389.

Suñol, J. J., G. Arbat, J. Pujol, L. Feliu, R. M. Fraguell, and A. Planas-Lladó. 2016. "Peer and self-assessment applied to oral presentations from a multidisciplinary perspective." *Assessment & Evaluation in Higher Education* 41 (4): 622-637.

Wu, C., E. Chanda, and J. Willison. 2014. "Implementation and outcomes of online self and peer assessment on group based honours research projects." *Assessment & Evaluation in Higher Education* 39 (1): 21-37.

Table 1. Overview of the errors of various methods for a typical scenario. In a group of 52 students each student undertakes 4 projects in groups of 4 people.

Method Name	Average Absolute Error	Maximum Error
Self Organised Peer Pressure	5.5	13.8
Reflexive Accounts	3.7	10.6
Mark-Adjusted Reflexive Accounts	3.5	11.3
Scaled Peer Assessment	1.7	4.5
Ranked Peer Assessment	3.3	12.3
Inverse Problem Approach	1.3	1.9
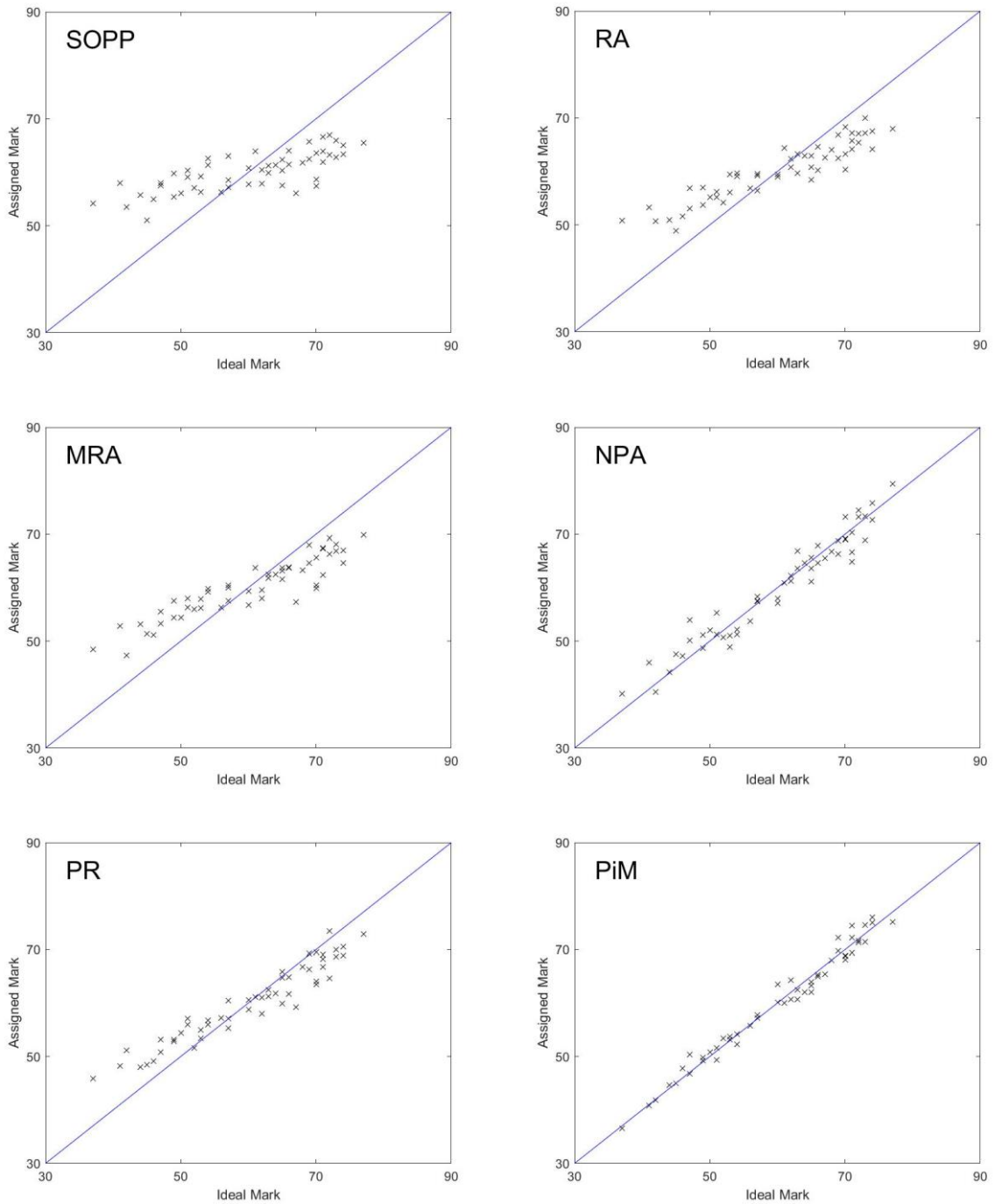
Figure 1. Scatter plot of assigned mark against ideal mark for various marking schemes. Shown are results for Self organised peer assessment (SOPP - top left), Reflexive accounts (RA - top right), Mark-adjusted reflexive accounts (MRA - middle left), Normalised peer assessment (NPA - middle right), Peer ranking (PR - bottom left) and Pseudoinverse marking (PiM - bottom right). Diagonal lines indicate the ideal distribution of assigned marks. Pseudoinverse marking is the best distribution of marks. ($N = 52$, $m = 4$, and $p = 4$).
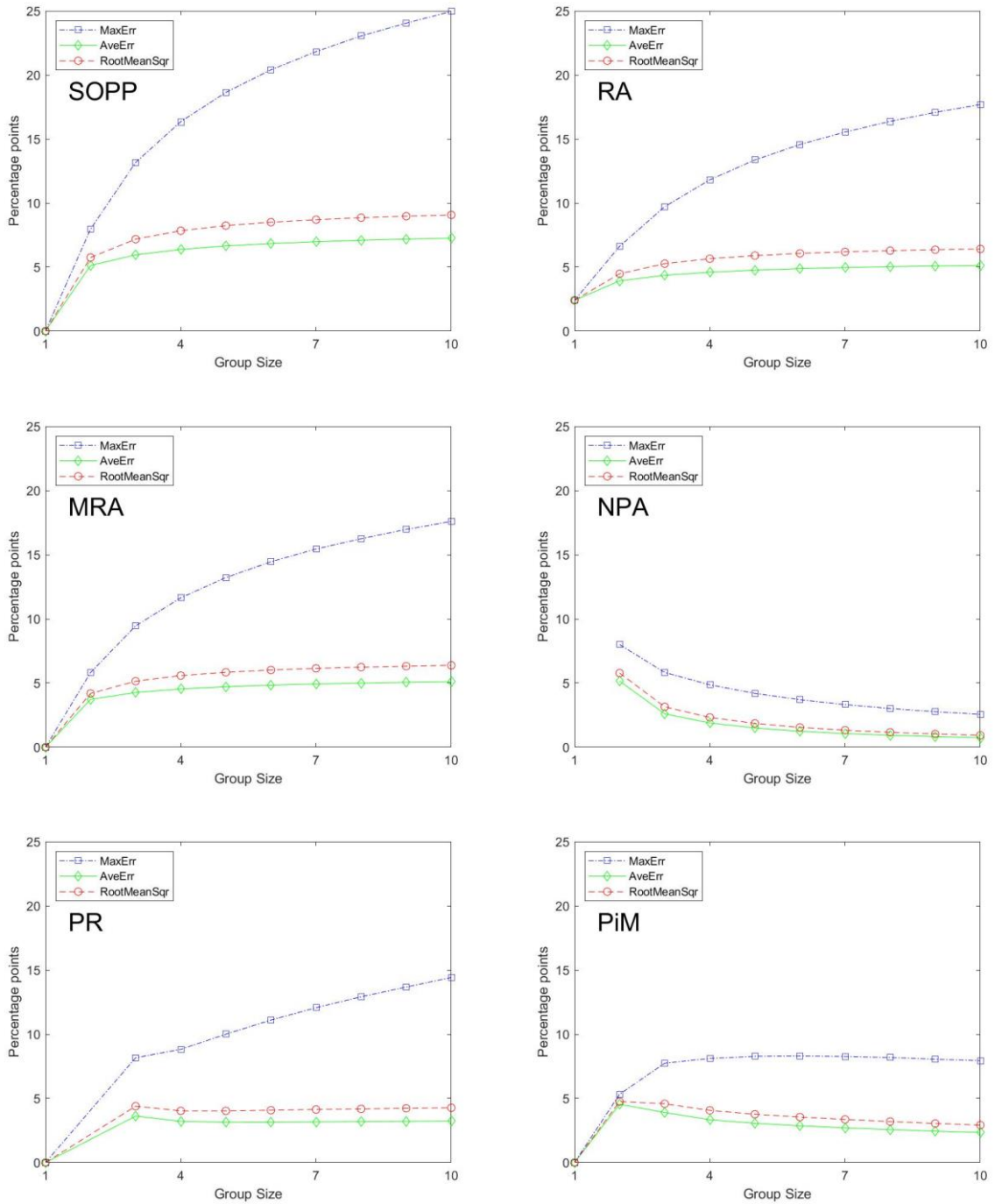
Figure 2. Plot of the different measures of error against different group sizes m for various marking schemes. Shown are results for Self organised peer assessment (SOPP - top left), Reflexive accounts (RA - top right), Mark-adjusted reflexive accounts (MRA - middle left), Normalised peer assessment (NPA - middle right), Peer ranking (PR - bottom left) and Pseudoinverse marking (PiM - bottom right). Normalised peer assessment and Pseudoinverse marking show interesting trends as using larger group sizes reduces all three errors.
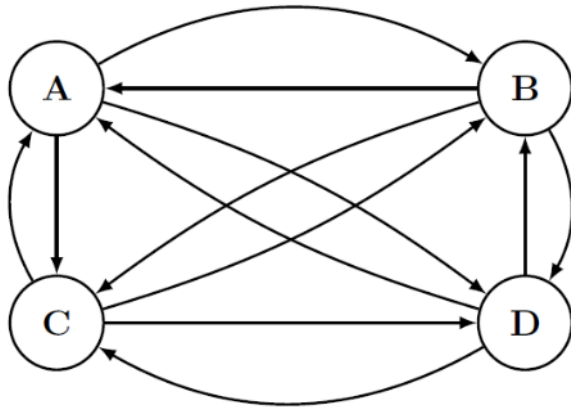
$$A = \begin{pmatrix} 0 & 2a & a+b & 2b \\ 2b & 0 & 2b & 2b \\ a+b & 2a & 0 & 2b \\ 2a & 2a & 2a & 0 \end{pmatrix}$$

Student A: $B > C > D$
Student B: $C > A > D$
Student C: $B > A > D$
Student D: $B > A > C$

Figure 3. Example of peer ranking. In a set of four students each student forms an opinion about the relative contributions of the other three students (Bottom right). This information can be represented as a directed graph (left), which can be in turn represented as a matrix (top right). The leading eigenvector of this matrix provides an aggregated measure for the relative contributions.
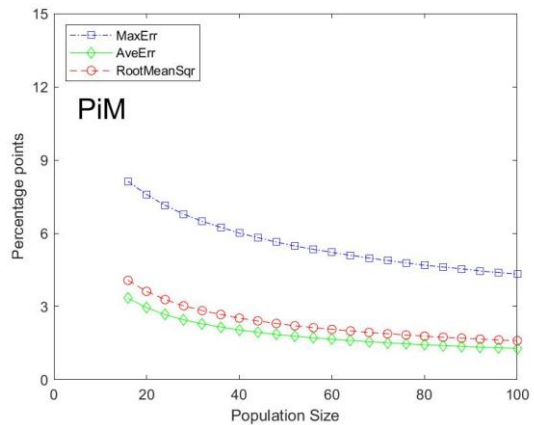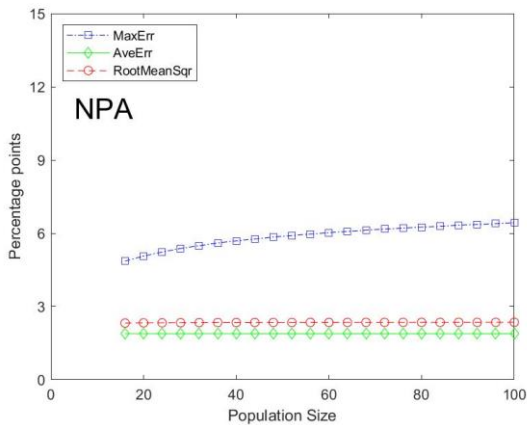

Figure 4. Plot of the different measures of error against different population sizes N. Shown are results for Normalised peer assessment (NPA - left) and Pseudoinverse marking (PiM - right). Results for other methods are similar to NPA and hence have been omitted. Unlike the other methods, Pseudoinverse marking decreases all three measures of error as the population size increases. ($m = 4$, and $p = 4$)

16