Using the method of Normalisation for mapping group marks to individual marks: Some observations

Mehdi Chowdhury
Bournemouth University, UK

October 2019

The method of normalisation combined with peer ratings is utilised to provide the solution to the biased rating problem of mapping group work marks to individual marks. We critically evaluate the method of normalisation following the findings of an article of the author which argued against the use of self and/or peer rating mechanism. We demonstrate that the findings of that article also hold for the method of normalisation as the influence of human behavioural factors are not accommodated in the designs. Additionally, we argue that the method (and its variants) is rather complicated, where all possible contingencies are not pre-specified. It makes the arrangement between tutors and students in conducting peer assessments incomplete and unverifiable by a third party.

Key Words: Game theory, Peer Assessment, Normalisation, Group Work

Correspondence to: Mehdi Chowdhury, Business School, Bournemouth University, D165 Dorset House, Talbot Campus, Fern Barrow, Poole, BH12 5BB, UK.
E-mail: mchowdhury@bournemouth.ac.uk
Tel: +44 (0)1202 961 397

Introduction:

A mechanism known as the Self and Peer Assessment or Ratings (SPA) is utilised in higher education to map group work marks to individual marks. This mechanism is also known as Self and Peer Ratings. The SPA originated from the idea that in a group work based assessment, it is important to assess both the process and product. As the process is not always apparent from the final product (e.g. essays, reports), SPA attempts to obtain information by asking group members to rate each other's contribution to the product of the group (Eberly Center, 2018; Zhang and Ohland, 2009). Previous research (Magin, 2001; Zhang and Ohland, 2009) has identified that this mechanism suffers from a number of biases resulting in students' not revealing the true information of actual contribution of group members to the tutor. Therefore, a number of articles have attempted to design SPA enabling truthful information transmission from students to tutors. In a previously authored article (Chowdhury, 2019), I questioned the effectiveness of these attempts and demonstrated that the interpretation of the ratings in the SPA is not possible because of the inherent weakness of not accommodating human behavioural factors in those designs. This finding implied that SPA should not be utilised to map group marks to individual marks as the tutors essentially do not know what the ratings actually mean.

A challenge to the findings of my previously authored article (Chowdhury, 2019) comes from the method of normalisation combined with peer assessment only. A student cannot influence his or her own mark when peer assessment is combined with normalisation. Hence it can be argued that normalisation combined with peer assessment takes care of the biases and invalidates the findings of my article. A number of variants of normalisation are available. We in this paper particularly analyse the method developed by Li (2001) and then further developed by Bushell (2006) and Spatar et al. (2015). Spatar et al. (2015) claimed to have developed a comprehensive method using normalisation taking a number of factors into consideration including the problem of biased reporting.

The objective of our paper is to establish that the findings of my previous article still hold as the weakness of the mechanism that comes from inherent behavioural factors is not taken care of in these methods. The ratings therefore remain uninterpretable and can not be granted as truthful. Neither can the ratings be treated as false. In order to accomplish the objective in this paper, we first briefly summarise the findings of Chowdhury (2019) followed by an illustration of the method of normalisation as in Li (2001), Bushell (2006) and Spatar et al. (2015). The discussion and criticism then follows.

It should be noted that the paper is not against the practice of group work in higher education. Group work helps to develop skills considered important in the professional world and a vital requirement of graduate recruitment criteria (Eberly Center, 2019; University of Birmingham, 2019). Our article is only addressing the issue of assignment of marks in group work based assessments and in this context evaluating the practice of normalisation through peer assessment which claims to assign individual grades in a fair manner.

Findings of Chowdhury (2019):

I, in the article Chowdhury (2019) identified that the practice of self and peer assessment can be represented as a strategic form game. A strategic form game is defined by three elements (i) the set of players (ii) the set of the strategies of the players, and (iii) the outcomes (utilities) of players from different combinations of strategies played (Maschler et al. 2013, page 77). In the practice self and peer assessment (SPA) these three distinct elements can be observed i.e. (i) the group members as players (ii) The ratings of SPA as the strategies, and (iii) the individual grades as the outcomes of the combinations of strategies.

I, therefore, proposed that game theoretic arguments can be utilised to predict the possible outcomes of a SPA. According to the game theory literature, a player has a dominant strategy if a strategy exists which is always a best response for any combination of other players' strategies (Maschler et al. 2013, page 85-

86). A rational player is expected to play the dominant strategy of a game. The article demonstrated that if utility/satisfaction increases with marks and a player is motivated by self-interest only (i.e. indifferent about the mark/satisfaction of peers) then the dominant strategy constitutes reporting the ratings as such the best possible mark irrespective of the ratings of the other group members is ensured. In a generic SPA a student therefore is predicted to report the highest self and the lowest peer contributions. The paper also demonstrated that if only peer rating is used, a student still has the dominant strategy of rating peers low to increase his or her marks.

However, through studying the data of an actual SPA, I, in the article demonstrated that whilst the design allowed for a dominant strategy for each player (i.e. students), they did not play their dominant strategies. Hence it is not possible to infer from the ratings whether students have reported truthfully or not. It implies that the ratings are not interpretable following the criteria set up by the tutor. The paper suggested that issues like trust, reciprocity, altruism and guilt aversion can influence the reporting of students in similar manners observed by the behavioural economists (see Fehr and Schmidt 2006). The assumption of indifference, which is vital for working of the SPA, also is unlikely to hold.

In the article, I therefore concluded that it is impossible to tell whether the ratings are true, false or to what extent human behavioural factors influence the reporting. Consequently the ratings cannot be relied upon to map group work marks to individual marks.

Normalisation and the solution of biased reporting

The problem of biased reporting has been identified by the previous researchers even though the structural similarities between SPA and/or PA with a game was not identified. No attempt was previously made to understand the dominant strategies of a SPA permitted by the design structure. However, as mentioned before, the previous research recognised that a student can make an attempt to increase his/her grade by falsely reporting SPA. A solution to this has been claimed to come through applying normalisation.

At first, we will see that the utilising peer assessment only, which appears to be a natural solution does not solve the problem of incentivising misreporting. It is demonstrated through Table 1.

(Insert Table 1 Here)

Table 1 uses the example proposed by Li (2001), which was later used and further developed by Bushell (2006) and Spatar et al. (2015). Li (2001)'s basic design consisted of a group with 4 members. The group mark of a piece of work submitted by that group was 70. To assign individual mark from the group mark, the group members were asked to rate each other's except for themselves. Li (2001) used 4 categories by which ratings of the level of participation (or contribution) were conducted. In each category, a rater (student) submits a rating on a 5 point scale ranging from poor to outstanding contribution. The final total rating of a student is calculated by adding the ratings under the 4 categories (The final rating was used in Bushell 2006 and Spatar et al. 2015). The rating scales were,

Poor	-	1
Below average	-	2
Average	-	3
Above average	-	4
Outstanding	-	5

Following these scales, as there are 4 categories, the maximum possible peer rating from a student to another student is 20 and the minimum possible is 4. Therefore, the sum of peer ratings of a student from 3 peers ranges from 12 to 60.

Table 1 presents the mapping of marks in two panels. The upper panel is the benchmark case (as in Li 2001, Bushell 2006 and Spatar et al. 2015) and the lower panel shows the effects of misreporting. In the upper half, the names of those providing the rating are in the first column and the names of the receivers of rating are given in the first row. For example, Peter rates the contribution of John, Mary and Janet respectively 14, 17, and 16. The last column shows the total rating awarded by a rater, e.g. the total rating awarded by Peter is 47. The sixth row shows the total rating received by a rater. For example Peter received rating 16, 16 and 17 provided respectively by John, Mary and Janet. The total rating received by Peter is 49. The last cell of the sixth row shows the average of total rating received (or equivalently awarded) by the four raters which is equal to 47.75. The seventh row shows the Individual Weighting Factor (IWF) calculated by dividing the rating received by the average of the total rating awarded. For example, Peter's IWF is 49/47.75=1.026. Therefore, his Individual Mark is GM× IWF = 70×1.026 = 71.83.

It is possible, however, for a student to utilise the conversion method for his/her own benefit. It has been depicted in the lower panel in Table 1, where Peter awards John the lowest possible rating of 4 instead of 14, which results in him effectively awarding himself a higher individual mark 75.80. Note that the IM of John has gone down to 52.60, and IMs of Mary and Janet have gone up to 80.44 and 71.16 respectively.

The solution proposed to this problem by Li (2001), and which has been largely utilised and extended by Spatar et al. (2015) is implemented by normalisation of ratings. In Table 1, we observe that students can increase their own marks by rating peers low. This is not possible in the normalisation method as is demonstrated in Table 2.

(Insert Table 2 here)

The first half of Table 2 is the benchmark case which is identical to the benchmark case in Table 1. However, in the second half, individual marking under simple normalisation (NIM) and adopted method of Spatar et al. (2015) have been presented.

In order to normalise, the total grade awarded by a rater has been added, and then the rating received by an individual from the rater has been divided by that added rating. For example, the total rating awarded by Peter is 47. The rating awarded to John by Peter is 14. Therefore John's normalised rating from Peter is 14/47=0.298. Similarly John's normalised rating from Mary is 0.356 and 0.286 from Janet. By adding 0.298, 0.356 and 0.286 John's Normalised Individual Weighting Factor (NIWF) is calculated as 0.939. By multiplying GM and NIWF John's Normalised Individual Mark (NIM) is calculated as 65.74. The modified method suggested by Spatar et al. (2015) results in John receiving a mark of 67.87.

The method of normalisation claimed to provide a solution to biased reporting or the problem of manipulation observed in Table 1. In Table 3, the effects of biased reporting on individual marks have been demonstrated.

(Insert Table 3 here)

In order to make the comparison simple, the top half of Table 3 is constructed in the same format as the bottom half of Table 1. Table 3 also needs to be compared with Table 2 to understand the impact of normalisation. The lower half of Table 3 shows the grade under the simple method of normalisation together with a grade using the method proposed by Spatar et al. (2015). Table 3 shows that Peter awarded John the lowest possible rating of 4. However under normalisation, this does not increase Peter’s grade which remains the same as the 71.57 awarded in Table 2. Using the Spatar et al. (2015) method, it increases the mark marginally from 71.17 to 71.43. The normalisation method therefore gives Peter negligible incentive to misreport. It should also be noted that we observe some changes of the marks of the peers.

As Peter’s mark does not change, one may feel tempted to state that normalisation provides a solution to the biased or non-truthful reporting problem. However, following the findings of my previous article (Chowdhury, 2019), we propose that the method does not solve the problem but adds further complications and this is discussed in the next section.

Criticism of Normalisation:

In this section we provide criticisms of the use of normalisation which has been developed by Li (2001) and furthered by Bushell (2006) and Spatar et al. (2015). Our main argument in a nutshell is that problem of SPA is not a problem of misreporting instead is a problem of interpretation. However the solutions proposed by previous papers address the misreporting problem instead of interpretation problem. We also identify that the methods proposed are complicated and impractical given the current higher education environment and expose tutors to challenges by students.

i. Interpretation problem, friendship bias and sabotage:

I, in my previous article identified that the designs of the Self and Peer Assessment or only Peer Assessment do not take the utility of students into consideration. The utility function captures the satisfaction received by the students from marks. When the characteristics of a utility function are unknown, the interpretation is impossible. The article showed that if the students are indifferent, i.e. only motivated by own utility, then false reporting will take place in a way that self will be reported above peers. However, it did not take place in the actual ratings of the paper. Hence, the paper concluded that it is impossible to identify whether students lied or told the truth.

The method of normalisation claims to have solved this problem as students can not affect own marks when only peer rating is used. However, it can only work if the students are indifferent about the mark of the peers. However it does not imply the absence of misreporting as there is no incentive to report what is true.

On the other hand misreporting is likely if students are not indifferent about the marks of peers. One possible source of misreporting is friendship or reciprocity bias which has been previously addressed in the literature (Magin 2001; Falchikov 2005). It is illustrated in Table 4.

(Insert Table 4 here)

In Table 4, we assumed a benchmark case where the group members agreed to rate 12 to each other. Now assume that Peter likes John and decides to award 20. It does not decrease Peter's mark however it increases that of John's and decreases Mary and Janet's. This is illustrated in the lower panel of Table 4.

Similarly it is also possible to sabotage ratings. Assume for some reason Peter does not like John and wants to sabotage by awarding a lower rate. It does not change Peter's own mark but reduces John's. It also increases Mary and Janet's mark and this is illustrated in Table 5.

(Insert Table 5 here)

It should be now clear that normalisation is subject to friendship bias and sabotage of group members. Essentially the method does not account for human behavioural factors. There is no incentive for truthful reporting when students are indifferent about the mark of peers. It makes the interpretation of ratings impossible as the tutor cannot distinguish between true or false reports. Hence whatever the modification is, such as proposed by the normalisation, the inherent problem of interpretation remains unresolved.

ii. Incomplete contract:

In addition to the above stated interpretation problem, the normalisation also suffers from the incomplete contract problem. Incomplete contract is defined as a contract that cannot specify all contingencies and cannot be verified by a third party (Tirole 1999). The arrangement of self and/or peer assessment implies a formal contract between group members and the tutor. In the arrangement the students are instructed to submit the ratings. However the way normalisation works can be described by following instruction:

'The total rating of your other group members is equal to 1. Allocate ratings to the other members according to their contributions as such total rating is equal to 1.'

Therefore the instruction to submit ratings and instruction specifying actual utilisation are different. Any method that normalises the original rating modifies the meaning of the instruction given to students. It consequently constitutes a divergence from the original arrangement/contract between tutors and students in conducting peer assessment.

Additionally, although the normalisation uses only simple mathematical operations; it requires substantial training for both the tutors and students. The training should explain all contingencies including the effects of the friendship bias and sabotage. A proper implementation of the method has a substantially high cost (takes more time) implication for both tutors and students. The complexity of the mapping process introduces the possibility that the outcomes could be the subject of an appeal to a higher academic committee. Because of the ambiguity/complexity of the mapping method the decision of the appeal is likely to go against tutors. Hence a tutor faces a greater risk from the incomplete contract scenario.

Many tutors in this regard would interview students ex-post of reporting. It further complicates the matter. Modifying marks following ex-post interview implies that the students have non-truthfully reported the ratings which again is subject to further deliberation of a higher academic committee and this may constitute a form of academic offence. It also has substantial additional cost (in time) implications for both tutors and students.

Conclusion:

We praise the academic endeavours that have attempted to find methods of mapping group marks to individual marks using self and/or peer assessment. However, in light of the criticisms in this paper we would question whether such a method is actually attainable.

Many now view group work as an integral part of learning in higher education. We are also supportive of this view and believe that group learning generates substantial experience to students and is valuable in their post education careers. However, assigning marks to an individual student in a group work based assessment is problematical due to information asymmetry between tutors and students. We believe that the subject matter would benefit from an interdisciplinary research approach, combining education literature with behavioural economics, in an attempt to reconcile the issues highlighted in this paper. We would also solicit further endeavour from academicians and educators to suggest, if possible, alternatives of self and/or peer assessment in summative assessment of learning.

References

Bushell, G. 2006. "Moderation of peer assessment in group projects." *Assessment & Evaluation in Higher Education* 31(1): 91-108. DOI: 10.1080/02602930500262395.

Chowdhury, M. 2019. "Reliability of self and peer assessment in group work in higher education." *Review of Behavioural Economics* 6(2):147-158. http://dx.doi.org/10.1561/105.00000104.

Eberly Center. 2018. "Assess Learning & Teaching." *Eberly Center for Teaching Excellence and Educational Innovation*, Carnegie Mellon University, Available on http://www.cmu.edu/teaching/assessment/index.html. Date accessed 16 March, 2018.

Eberly Center. 2019. "What are the benefits of group work?" *Eberly Center for Teaching Excellence and Educational Innovation*, Carnegie Mellon University, Available on https://www.cmu.edu/teaching/designteach/design/instructionalstrategies/groupprojects/benefits.html. Date accessed 26 September, 2019.

Falchikov, N. 2005. "Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education." Routledge Falmer.

Fehr, E. and Schmidt, K. 2006. "The economics of fairness, reciprocity and altruism: experimental evidence and new theories". In S. C. Kolm & J. M. Ythier (Eds.), *Handbook of Economics of giving, altruism and reciprocity* (pp. 615-691). Kidlington, Oxford: North-Holland.

Li, L. 2001. "Some refinements on peer assessment of group projects." *Assessment & Evaluation in Higher Education* 26(1), 5-18. DOI: 10.1080/0260293002002255.

Margin, D. 2001. “Reciprocity as a source of bias in multiple peer assessment of group work.” *Studies in Higher Education* 26(1): 53-63, DOI: 10.1080/03075070020030715.

Maschler, M., E. Solan, and S. Zamir. 2013. “Game Theory.” Cambridge University Press.

Spatar, C., N. Penna, H. Mills, V. Kutija, and M. Cooke. 2015. “A robust approach for mapping group marks to individual marks using peer assessment.” *Assessment & Evaluation in Higher Education* 40(3): 371-389. DOI:10.1080/02602938.2014.917270

Tirole, J. 1999. “Incomplete contracts: Where do we stand?” *Econometrica* 67(4): 741-781.

University of Birmingham. 2019. “Why work in groups?” ‘https://www.birmingham.ac.uk/schools/metallurgy-materials/about/cases/group-work/why.aspx, Date accessed 27th September, 2019.

Zhang, B., and Ohland , M. 2009. “How to assign individualized scores on a group project: An empirical evaluation.” *Applied Measurement in Education* 22(3): 290-308. DOI:10.1080/08957340902984075.

Table 1:

The benchmark case in Li (2001), Bushell (2006) and Spatar et al. (2015)

The Benchmark case					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		14	17	16	47
John	16		17	17	50
Mary	16	16		13	45
Janet	17	14	18		49
Total rate received and average rating	49	44	52	46**	47.75
IWF	1.026	0.921	1.089	0.963	
IM=GM* × IWF	71.83	64.50	76.23	67.43	
Manipulation of rating					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		4	17	16	37
John	16		17	17	50
Mary	16	16		13	45
Janet	17	14	18		49
Total rate received and average rating	49	34	52	46	45.25
IWF	1.083	0.751	1.149	1.017	
IM=GM × IWF	75.80	52.60	80.44	71.16	

*GM (Group mark)=70

** We have found that in Li (2001) the rating received by Janet adds to 47, however Bushell (2006) and Spatar et al. (2015) used 46, we therefore in the tables use 46.

Table 2:

Use of Normalisation in mapping group marks to individual marks

The Benchmark case					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		14	17	16	47
John	16		17	17	50
Mary	16	16		13	45
Janet	17	14	18		49
Total rate received and average rating	49	44	52	46	47.75
IWF	1.026	0.921	1.089	0.963	
IM=GM × IWF	71.83	64.50	76.23	67.43	
Normalisation of rating and Individual marks					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		0.298	0.362	0.340	1
John	0.32		0.34	0.34	1
Mary	0.356	0.356		0.289	1
Janet	0.347	0.286	0.367		1
NIWF	1.022	0.939	1.069	0.969	
NIM= GM × NIWF	71.57	65.74	74.83	67.85	
Spatar at el. (2015)	71.17	67.87	73.72	68.70	

Table 3:

Manipulation of grades and effects on different assessment methods

The Manipulated Benchmark case (Table 1)					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		4	17	16	37
John	16		17	17	50
Mary	16	16		13	45
Janet	17	14	18		49
Total rate received and average rating	49	34	52	46	45.25
IWF	1.083	0.751	1.149	1.017	
IM=GM × IWF	75.80	52.60	80.44	71.16	
Normalisation of rating and Individual marks					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		0.108	0.459	0.432	1
John	0.320		0.34	0.34	1
Mary	0.356	0.356		0.289	1
Janet	0.347	0.286	0.367		1
NIWF	1.022	0.749	1.167	1.061	
NIM= GM × NIWF	71.57	52.46	81.68	74.29	
Spatar at el. (2015)	71.43	61.23	77.89	72.96	

Table 4:

Effects of Friendship

The Benchmark case with friendship bias					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		20	12	12	44
John	12		12	12	36
Mary	12	12		12	36
Janet	12	12	12		36
Total rate received and average rating	36	44	36	36	38
IWF	0.947	1.158	0.947	0.947	
IM=GM × IWF	66.32	81.05	66.32	66.32	
Normalisation of rating and Individual marks					
Rater/Assesse	Peter	John	Mary	Janet	Total Awarded
Peter		0.455	0.273	0.273	1
John	0.333		0.333	0.333	1
Mary	0.333	0.333		0.333	1
Janet	0.333	0.333	0.333		1
NIWF	1.000	1.121	0.939	0.939	
NIM= GM × NIWF	70.00	78.48	65.76	65.76	
Spatar at el. (2015)	70.00	74.03	66.82	66.82	

Table 5:

Effects of Sabotage

<table>
<tr><td colspan="6">The Benchmark case with sabotage</td></tr>
<tr><td>Rater/Assesse</td><td>Peter</td><td>John</td><td>Mary</td><td>Janet</td><td>Total Awarded</td></tr>
<tr><td>Peter</td><td></td><td>4</td><td>12</td><td>12</td><td>28</td></tr>
<tr><td>John</td><td>12</td><td></td><td>12</td><td>12</td><td>36</td></tr>
<tr><td>Mary</td><td>12</td><td>12</td><td></td><td>12</td><td>36</td></tr>
<tr><td>Janet</td><td>12</td><td>12</td><td>12</td><td></td><td>36</td></tr>
<tr><td>Total rate received and average rating</td><td>36</td><td>28</td><td>36</td><td>36</td><td>34</td></tr>
<tr><td>IWF</td><td>1.059</td><td>0.824</td><td>1.059</td><td>1.059</td><td></td></tr>
<tr><td>IM=GM × IWF</td><td>74.12</td><td>57.65</td><td>74.12</td><td>74.12</td><td></td></tr>
<tr><td colspan="6">Normalisation of rating and Individual marks</td></tr>
<tr><td>Rater/Assesse</td><td>Peter</td><td>John</td><td>Mary</td><td>Janet</td><td>Total Awarded</td></tr>
<tr><td>Peter</td><td></td><td>0.143</td><td>0.429</td><td>0.429</td><td>1</td></tr>
<tr><td>John</td><td>0.333</td><td></td><td>0.333</td><td>0.333</td><td>1</td></tr>
<tr><td>Mary</td><td>0.333</td><td>0.333</td><td></td><td>0.333</td><td>1</td></tr>
<tr><td>Janet</td><td>0.333</td><td>0.333</td><td>0.333</td><td></td><td>1</td></tr>
<tr><td>NIWF</td><td>1.000</td><td>0.810</td><td>1.095</td><td>1.095</td><td></td></tr>
<tr><td>NIM= GM × NIWF</td><td>70.00</td><td>56.67</td><td>76.67</td><td>76.67</td><td></td></tr>
<tr><td>Spatar at el. (2015)</td><td>70.00</td><td>63.33</td><td>74.70</td><td>74.70</td><td></td></tr>
</table>