An Extended Approach to Adjust Inconsistent Minority Peer- and Self-assessment Scores of Teamwork using Assessor's Reliability

Kali Prasad Nepal
Deakin University, Australia
Email: knepal@deakin.edu.au

CONTEXT

Teamwork is generally assessed either solely by academic staff or by both academic staff and students themselves confidentially as well as collaboratively. Peer- and self-assessments have been used primarily to assess *teamwork process* and teacher assessment to assess *teamwork product*. Peer- and self-assessments are useful to elicit team members' contribution towards teamwork and to convert team mark into individual marks, provided the scores are reliable (the extent to which the scores are consistent). However, not all peer- and self-assessment scores are reliable. Anecdotal and literature evidence suggest that there are several cases of inconsistencies in these scores. Individual contribution scores given by teammates to an assessee (including himself/herself) can sometimes vary significantly due to both intentional and unintentional reasons. Simply using total individual rating scores without considering an assessor's reliability to estimate individual contribution factors can sometime results unfair grades and becomes hindrance to learning through teamwork.

PURPOSE

This study proposes an extended approach to adjust inconsistent and/or distorted minority peer and self-assessment scores of teamwork using standard normal probability concept.

APPROACH

In order to adjust inconsistent and/or distorted minority peer-and self-assessment scores of teamwork, an extended approach has been proposed. The approach uses the reliability of assessor's scores of an assessee using standard normal probability curve. The evaluation of the extended approach is conducted by comparing with the existing approaches using two case examples of peer- and self-assessment of teamwork where minority team members' scores are inconsistent.

RESULTS

The evaluation of the extended approach shows that the proposed method is superior to the available approaches in order to adjust inconsistent peer- and self-assessment scores for special cases where scores of minority team members are inconsistent. The extended approach helps both to automatically detect such scoring anomalies and to adjust the scores so that the fairer contributions to the teamwork would be obtained and utilised.

CONCLUSIONS

The extended approach is useful in that it helps both to automatically detect scoring anomalies and to devise the methods to adjust them. However, the approach does not address the issue of scoring inconsistencies by majority of team members as it uses average score as a basis for identifying inconsistencies. Moreover, the approach needs to be implemented in the real teamwork environment in order to identify the impacts of these scoring adjustments in teamwork process and teamwork product.

KEYWORDS

Self- and peer-assessments, assessor's reliability, teamwork.

Introduction

Teamwork in engineering profession is highly valued. As a result, engineering schools have included teamwork as an important graduate attribute in their engineering curriculum. The benefits of learning and practising teamwork in engineering programs/courses and subjects/units are well documented (e.g., Hansen, 2006; Michaelsen et al., 2002). At the same time, issues and problems of learning through teamwork at engineering schools are also well recognised (e.g., Hansen, 2006; Li, 2001; Lejk, Wyvll, and Farrow, 1996). Assessment of individual student in a team has been identified as one of the major issues in learning through teamwork. If not assessed fairly and adequately, it may result in free-riding and/or eat-it-all behaviours, which may hinder the teamwork process as well as the quality of teamwork product. It may also demotivate students to contribute to teamwork resulting poor achievement of teamwork-related learning outcomes.

Teamwork is generally assessed either solely by academic staff (teacher assessment) or by both academic staff (teacher assessment) and students themselves (peer- and self-assessments) confidentially as well as collaboratively (co-assessment) (Nepal, 2016). The importance of self-assessments are also well documented in existing literature (e.g., Willey and Gardner, 2009; Boud 2013). Existing literature suggests a number of methods to assess teamwork process and teamwork product and to award team mark and individual marks. Lejk, Wyvll, and Farrow (1996) have summarised nine (9) methods. Seven (7) approaches were explored by Race (2000). Among these methods, assessing teamwork product for a team mark and adjusting team mark by using individual contributions for individual marks has been a popular choice (Conway *et al.,* 1993; Goldfinch, 1994; Nepal 2012).

Moreover, there is a common practice to elicit individual contribution towards teamwork by confidential peer- and self-assessments (individual contribution rating scores) by students themselves and to assess teamwork product (team mark) by academic staff (Nepal 2012; Alias, Masek, and Salleh, 2015). Individual contribution rating scores are then converted into individual contribution factors (*ICF*), also known as individual weight factors (*IWF*). *ICF* are then multiplied by team mark to award individual marks, provided the peer- and self-assessment scores from which *ICF* are derived are reliable (the extent to which the scores are consistent). However, not all peer- and self-assessment scores are reliable. Individual contribution rating scores given by teammates to an assessee (including himself/herself) can sometimes vary significantly due to both intentional and unintentional reasons. Simply using total individual rating scores without considering an assessor's reliability to estimate *ICF* can sometime results unfair grades and becomes hindrance to the learning through teamwork.

In this study, an extended approach which takes an assessor's reliability to an assessee into account using standard normal probability concept is proposed. To discuss the characteristics of this extended approach, mathematical equations and computations are presented and discussed with the help of two typical teamwork example cases of inconsistent minority peer- and self-assessment scores. Although, the proposed approach can be extended to categorical criterion-based self-and-peer assessment scores, the discussion here is based on holistic norm-based individual contribution rating scores where each assessor is asked to assign holistic individual contribution rating score to each assessee (including himself/herself) in such a way that the total score by an assessor to all assessees is constant (say, 100). This condition applies to all assessors.

Extended Approach

Let us assume that s_{ij} is a peer- and self-assessment score given by an assessor i to an assessee j for his or her contribution to a teamwork ($i = j$ is a self-assessment score). Total individual contribution rating of an assessee, j (denoted by ICR_j) is obtained by summing up the rating scores given by all assessors ($i = 1, 2, \ldots, N$) to an assessee j as in Equation (1).

$$\boldsymbol{ICR_j = \sum_{i=1}^{i=N} s_{ij}} \qquad \forall \boldsymbol{j \in N} \tag{1}$$

Where N is the number of members in a team.

The average contribution rating ($\overline{ACR}$) of all team members is calculated by summing up ICR_j of all assessees and dividing it by the number of members in a team N as in Equation (2).

$$\boldsymbol{\overline{ACR} = \frac{1}{N}\sum_{j=1}^{j=N} ICR_j = \frac{1}{N}\sum_{i=1}^{i=N}\sum_{j=1}^{j=N} s_{ij}} \qquad \forall \boldsymbol{j \in N} \tag{2}$$

The individual contribution factor of an assessee j (ICF_j) is calculated by dividing ICR_j by $\overline{ACR}$ as in Equation (3).

$$\boldsymbol{ICF_j = \frac{ICR_j}{\overline{ACR}} = N \times \frac{\sum_{i=1}^{i=N} s_{ij}}{\sum_{i=1}^{i=N}\sum_{j=1}^{j=N} s_{ij}}} \qquad \forall \boldsymbol{j \in N} \tag{3}$$

Equation (3), originally proposed by Conway et al. (1993), is the fundamental equation to compute individual contribution factor (*ICF*) of an assessee. Equation (3) would result fair individual marks if the individual contribution rating scores are consistent. However, not all self- and peer-assessment scores are consistent and reliable. Several extensions of this original method have been purposed (e.g., Li, 2001; Neus, 2011; Ko, 2014) but none of these methods fully consider an assessor's reliability to an assessee. Method proposed by Ko (2014) considers some sort of overall assessor's reliability but the proposed method requires iteration and highly sensitive to positive evaluation parameter which needs to be specified in advance.

In order to incorporate an assessor's reliability to an assessee, Equation (3) can be modified as Equation (4).

$$\boldsymbol{ICF_j = \frac{ICR_j}{\overline{ACR}} = N \times \frac{\sum_{i=1}^{i=N} w_{ij}\, s_{ij}}{\sum_{i=1}^{i=N}\sum_{j=1}^{j=N} w_{ij}\, s_{ij}}} \qquad \forall \boldsymbol{j \in N} \tag{4}$$

Where w_{ij} is the reliability weight of assessor i to an assessee j and sum of the reliability weights of all assessors to an assessee j can be equated to 1.00 as given by Equation (5).

$$\boldsymbol{\sum_{i=1}^{N} w_{ij} = 1.00} \qquad \forall \boldsymbol{j \in N} \tag{5}$$

Reliability weights (w_{ij}) can be estimated using Equation (6).

$$\boldsymbol{w_{ij} = \frac{r_{ij}}{\sum_{i=1}^{N} r_{ij}}} \qquad \forall \boldsymbol{i, j \in N} \tag{6}$$

Where r_{ij} is the relative relevance of assessor i's score of an assessee j.

Peer- and self-assessment scores can be used in order to estimate the relative relevance (r_{ij}). If an assessor's individual contribution rating score towards an assessee is similar to the average or mean score by all assessors towards the particular assessee, the assessor can be considered as relatively reliable. On the other hand, if an assessor's individual contribution rating score towards an assessee is significantly different than the average or mean score by all assessors towards that particular assessee, the assessor can be considered as relatively unreliable. Hence one of the methods to estimate this relative relevance is using standard normal probability of absolute z-score (a z-score is a standardised distance from the mean or average) using Equation (7).

$$\boldsymbol{r_{ij} = \phi\,(|z_{ij}|)} \qquad \forall \boldsymbol{i, j \in N} \tag{7}$$

Where z-score (z_{ij}) can be computed from Equation (8).

$$z_{ij} = \frac{s_{ij} - \frac{1}{N}\sum_{i=1}^{N} s_{ij}}{\sqrt{\frac{\sum_{i=1}^{i=N}\left(s_{ij} - \frac{1}{N}\sum_{i=1}^{N} s_{ij}\right)^2}{N}}} \qquad \forall\, i, j \in N \tag{8}$$

The relationship between relative relevance (standard normal probability) and absolute z-score is provided in Figure 1. When a z-score of a rating score given by an assessor to an assessee approaches zero, the assessor can be treated as 'reliable'. The larger value of z-score can be treated as 'less reliable' as it indicates divergence from the mean or average.

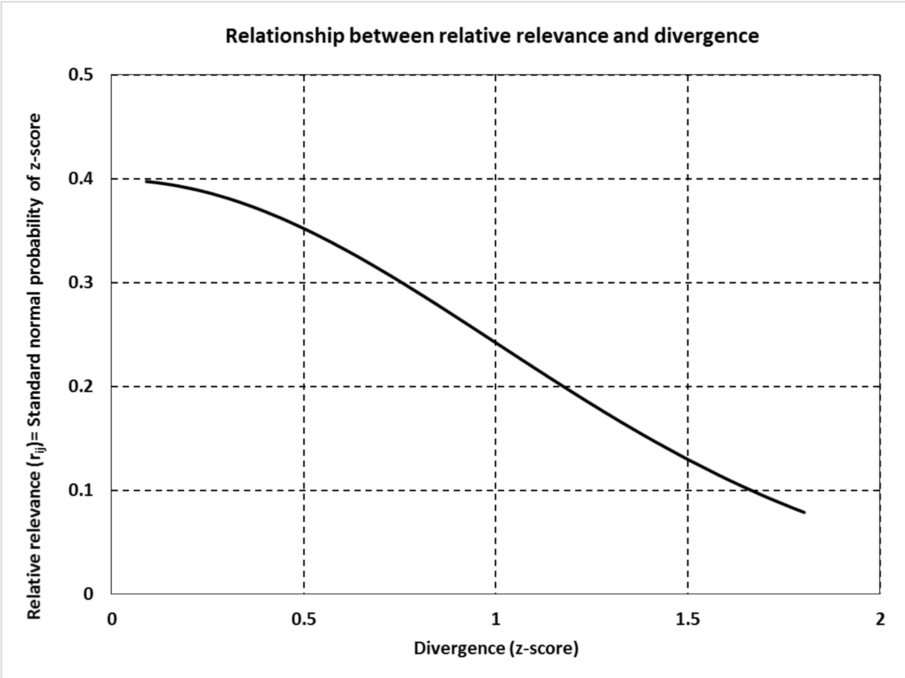

Figure 1: Relationship between relative relevance and z-score

Relative relevance (r_{ij}) varies from 0.40 (when z-score is 0 meaning that an assessor's rating score of an assessee matches with mean rating score) to 0 (when z-score is extremely large meaning that an assessor's rating score of an assessee is significantly different than that of mean score). However, the relative relevance becomes very small when z-score is more than 2. Once the relative relevance (r_{ij}) is estimated, reliability weights (w_{ij}) can be calculated using Equation (6) and proportionally adjusted to make sure that sum of the reliability weights of all assessors to an assessee becomes 1.00. The reliability weight-based individual contribution factors can then be estimated using Equation (4).

This approach is generous for rating scores which are close to the mean or average. This is a good approach to accommodate small divergences from the mean as there is always a subjectivity in peer- and self-assessments and it is not fair to penalise for small divergences. On the other hand, this method also makes sure that minority assessments are never dismissed although relative relevance of them diminish with the increase of divergence. The major limitation of this approach is that it does not address the issue of distorted majority assessments where majority of team members decide to rate themselves and each other in a pre-arranged way.

Example Cases

In this section, two cases of inconsistent or distorted minority peer- and self-assessment scores are discussed in order to see how the assessor's reliability would help to adjust these inconsistencies.

Case I: Over-rated self and under-rated peers' scores

Table 1 shows an example of a selfish student who over-rated self and under-rated peers. Assessor E over-rated himself/herself in the expense of his/her peers. Using Equation (3) which does not consider assessor's reliability, the individual contribution factors are 0.87, 0.87, 1.37, 0.87 and 1.02 (which translate to individual marks of 52, 52, 82, 52, and 61 for a team mark of 60) for assessees A, B, C, D and E respectively. If assessor E's scores are completely dismissed for being inconsistent, the individual contribution factors are 0.90, 0.90, 1.40, 0.90 and 0.90 (which translate to individual marks of 54, 54, 84, 54, and 54 for a team mark of 60) for assessees A, B, C, D and E respectively. Assessor E has successfully inflated his or her individual contribution factor by over-rating self-assessment score in the expense of peers' scores. This is not fair and this type of selfish behaviour should be discouraged. Using Equation (4) which considers assessor's reliability, the individual contribution factors are 0.89, 0.89, 1.39, 0.89 and 0.92 (which translate to individual marks of 53, 53, 83, 53, and 55 for a team mark of 60) for assessees A, B, C, D and E respectively. This is much fairer than the previous individual marks. The adjustment has resulted an increase in peers' individual marks and a decrease in self-mark. Minority assessment scores have also been taken into account albeit of lower weights.

Table 1: Over-rated self and under-rated peers' scores

Team mark = **60**	Individual Contribution Scores	Assessees (j)					Total
	Student	A	B	C	D	E	
Assessors (i)	A	18	18	28	18	18	100
	B	18	18	28	18	18	100
	C	18	18	28	18	18	100
	D	18	18	28	18	18	100
	E	**15**	**15**	**25**	**15**	**30**	100
Total individual contribution rating (ICR_j)		87	87	137	87	102	500
Average contribution rating ($\overline{ACR}$)		100					
Individual contribution factor ($ICF_j = \frac{ICR_j}{\overline{ACR}}$) from Equation (3)		0.87	0.87	1.37	0.87	1.02	
Individual marks from Equation (3)		**52**	**52**	**82**	**52**	**61**	
Individual contribution factor when dismissing E's scores		0.90	0.90	1.40	0.90	0.90	
Individual marks when dismissing E's scores		**54**	**54**	**84**	**54**	**54**	
Individual contribution factor ($ICF_j = \frac{ICR_j}{\overline{ACR}}$) from Equation (4)		0.89	0.89	1.39	0.89	0.92	
Individual marks from Equation (4)		**53**	**53**	**83**	**53**	**55**	

Case II: Under-rated self and over-rated peers' scores

Table 2 shows a rather uncommon but possible case of peer- and self-assessment rating scores where an overgenerous student under-rates himself/herself and over-rates peers.

Table 2: Under-rated self and over-rated peers' scores

Team mark = **60**	Individual Contribution Scores	Assessees (j)					Total
Assessors (i)	Student	A	B	C	D	E	
	A	18	18	28	18	18	100
	B	18	18	28	18	18	100
	C	18	18	28	18	18	100
	D	18	18	28	18	18	100
	E	**20**	**20**	**30**	**20**	**10**	100
Total individual contribution rating (ICR_j)		92	92	142	92	82	500
Average contribution rating ($\overline{ACR}$)		100					
Individual contribution factor ($ICF_j = \frac{ICR_j}{\overline{ACR}}$) from Equation (3)		0.92	0.92	1.42	0.92	0.82	
Individual marks from Equation (3)		**55**	**55**	**85**	**55**	**49**	
Individual contribution factor when dismissing E's scores		0.90	0.90	1.40	0.90	0.90	
Individual marks when dismissing E's scores		**54**	**54**	**84**	**54**	**54**	
Individual contribution factor ($ICF_j = \frac{ICR_j}{\overline{ACR}}$) from Equation (4)		0.90	0.90	1.40	0.90	0.89	
Individual marks from Equation (4)		**54**	**54**	**84**	**54**	**53**	

Using the Equation (3) which does not consider assessor's reliability, the individual contribution factors are 0.92, 0.92, 1.42, 0.92 and 0.82 (which convert to individual marks of 55, 55, 85, 55, and 49 for a team mark of 60) for assessees A, B, C, D and E respectively. If assessor E's scores are completely dismissed for being inconsistent, the individual contribution factors are 0.90, 0.90, 1.40, 0.90 and 0.90 (which convert to individual marks of 54, 54, 84, 54, and 54 for a team mark of 60) for assessees A, B, C, D and E respectively. Assessee E received lower individual mark because of his/her own under-rated self- and over-rated peers' scores. However, using Equation (4) which considers assessor's reliability, the individual contribution factors are 0.90, 0.90, 1.40, 0.90 and 0.89 (which convert to individual marks of 54, 54, 84, 54, and 53 for a team mark of 60) for assessees A, B, C, D and E respectively, which is more reasonable. The adjustment has resulted a decrease in peers' individual marks and an increase in self-mark. Again, minority assessment scores have been taken into account albeit of lower weights.

Discussion

This study presents an extended approach to detect and refine inconsistent peer- and self-assessment scores in a teamwork and to compute realistic values of individual contribution factors for distorted minority peer-and-self assessment scores. Individual contribution factors are commonly multiplied by team mark to convert team mark into individual marks, provided the scores are reliable (the extent to which the scores are consistent). In order to discuss the characteristics of the proposed method, consistent mathematical equations are provided. Two example cases of peer- and self-assessment scores are used in order to discuss the characteristics of the extended approach. The case examples clearly show that the method is very useful to adjust inconsistent peer-and self-assessment scores. Even though the extended approach addresses the issue of distorted minority self-and-peer assessment scores, it does not address the issue of distorted majority assessments where majority of team members decide to rate themselves and each other in a pre-arranged way. Moreover, the approach needs to be implemented in the real teamwork environment in order to identify the impacts of these scoring adjustments in teamwork performance (both process and product).

References

Alias, M., Masek, A., & Salleh, H. H. M. (2015). Self, peer and teacher assessments in problem based learning: are they in agreements? *Procedia-Social and Behavioral Sciences*, 204, 309-317.

Boud, D. (2013). *Enhancing learning through self-assessmen*t. Routledge.

Conway, R., Kember, D., Sivan, A., & Wu, M. (1993). Peer assessment of an individual's contribution to a group project. *Assessment & Evaluation in Higher Education*, 18(1), 45-56.

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

Hansen, R. S. (2006). Benefits and problems with student teams, suggestions for improving team project. *Journal of Education for Business*, 82(1), 11-19.

Ko, S. 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., Wyvll, M., & Farrow, S. (1996). A survey of methods of deriving individual grades from group assessments. *Assessment & Evaluation in Higher Education*, 21(3), 267–280.

Li, L. K. Y. (2001). Some refinements on peer assessment of group projects. *Assessment & Evaluation in Higher Education*, 26(1), 5-18.

Michaelsen, L. K., Knight, A. B., & Fink, L. D. (Eds.) (2002*). Team-based learning: A transformative use of small groups*. Greenwood publishing group.

Nepal, K. P. (2012). A comparative evaluation of analytical methods to allocate individual marks from a team mark. *European Journal of Engineering Education*, 37(4), 397-404.

Nepal, K. P. (2016). Simplified framework for managing team learning in engineering subjects. *International Journal of Engineering Education*, 32(3A), 1182–1193.

Neus, J. L. (2011). Peer assessment accounting for student agreement. *Assessment & Evaluation in Higher Education*, 36(3), 301-314.

Race, P. (2000). *500 Tips on Group Learning*, 2000, London: Kogan Page.

Willey, K., & Gardner, A. (2009). Improving self- and peer assessment processes with technology. Campus-Wide Information Systems, 26 (5), 379–399.

Acknowledgements

This study was partly supported by internal research grants from Deakin School of Engineering. Author would like to thank the school for the support.