International Journal of Engineering Education Vol. 34, No. 4, pp. 1289–1298, 2018
Printed in Great Britain

0949-149X/91 $3.00+0.00

A Relative Relevance Approach to Refine Inconsistent Peer- and Self-Assessment Scores in Teamwork Assessment*

KALI PRASAD NEPAL
School of Engineering, Deakin University, Geelong, Australia. E-mail: knepal@deakin.edu.au

This study proposes an extended approach to refine inconsistent peer- and self-assessment scores in teamwork assessment. These refined scores are commonly used to estimate individual contribution factors, also referred in some literature as individual weighting factors. The individual contribution factors are then multiplied by team mark to convert team mark into individual marks, provided the scores are valid (the degree to which the scores measure the true contributions) and reliable (the extent to which the scores are consistent). However, not all peer- and self-assessment scores are valid and reliable. Although the validity is as equally, if not more, important as reliability, this study focuses on the reliability. Anecdotal and literature evidence suggests that there are several cases of inconsistencies in students' peer- and self-assessment scores. Creative accounting scores (over-rating to self and under-rating to peers) by some minority team members are commonly encountered cases of inconsistencies, which are addressed by the proposed extension. To discuss the characteristics of the extended approach, mathematical equations and computations are presented and discussed with the help of typical inconsistent peer- and self-assessment scores. The analysis clearly shows that relative relevance approach based on standard normal probability can be a viable option in order to refine creative accounting cases of inconsistencies.

Keywords: teamwork; peer-assessment; self-assessment; reliability; relative relevance

1. Introduction

Teamwork has been an important part of contemporary higher education as most employers value teamwork-based graduate attributes in their new employees as highly, if not more highly than their ability to work independently. Teamwork in higher education is the process of students learning and working collaboratively, cohesively and cooperatively on a specific team task, activity or assessment item to achieve teamwork-related and/or associated learning outcomes [1, 2]. Several dimensions of teamwork education—PBL, participants, contents, projects, information technology, and assessment-are elaborated by Fruchter [3]. An achievement on learning outcomes of teamwork demonstrates that the students are able to learn and work with other people from diverse disciplines and backgrounds and in a range of situations. Higher education institutions enforce academic staff to teach, practise and assess teamwork knowledge, teamwork skills and processes, teamwork products or outputs and teamwork experience [1, 4]. Potential benefits from learning through teamwork are obvious, but they also have numerous problems [5]. Li [6], Brewer and Mendelson [7] and Lejk, et al. [8] have specifically emphasised the teamwork assessment as one of the most important issues. Although teamwork is effective to help students learn in a team, the assessment of individual student's contribution towards teamwork is not simple. In an ideal teamwork environment, a team's performance is assessed primarily by the product a team produces collaboratively which is equally contributed by all team members. However, neither all team members are homogeneous nor are they equally motivated, hence resulting unequal contributions. In addition, unfair grades can adversely impact motivation and team performance and encourage free-riding.

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). The advantages and disadvantages of peer- and self-assessments have been extensively discussed in existing literature (for example, [9, 10]) and higher education institutions' guidelines on teamwork. Based on these literature information, this study considers peer- and self-assessments are an important part of teamwork assessment process and argues for their widespread use. Peer- and self-assessment scores can be both holistic (based on a single criteria for overall assessment) as well as categorical (based on a number of assessment criteria). Categorical scores may provide better results as students may not always consider details in their holistic assessments. Peer—and self-assessment scores can also be both norm-based (in comparison to the performance of other team members) and criteria-based (based on mastery of a specific set of skills). Norm-based scores are preferred as indi-

* Accepted 22 October 2017.

vidual team member's contribution towards a teamwork is generally assessed relative to other team members rather than based on some absolute assessment criteria of that particular team member.

There is a common practice to elicit individual contributions towards teamwork by confidential peer- and self-assessments (individual contribution scores) by students themselves and to assess teamwork product (team mark) by academic staff [11, 12]. Individual contribution scores are then converted into individual contribution factors (*ICF*). *ICF* are then multiplied by team mark to award individual marks, provided the peer- and self-assessment scores from which *ICF* are derived are *valid* (the degree to which the scores measure the true contributions) and *reliable* (the extent to which the scores are consistent). However, not all self- and peer-assessment scores are valid and reliable. Some of these anomalies are due to deliberate actions whereas others are due to individual differences in natural abilities (senses, minds, nerves etc.), learned capabilities (knowledge, skills, experience etc.) and psychosocial behaviours (fear, relationship, friendship etc.) [13–18]. Examples of such anomalies include biased scores such as over-generous scores (over-rating to self, peers or both), under-valued scores (under-rating to self, peers or both), creative accounting scores (over-rating to self and under-rating to peers) and discriminatory scores (under-rating to some marginalised team members in some prearranged way) and many others [6, 19]. Although the validity is as equally, if not more, important as reliability, this study focuses on the reliability and proposes a method to refine inconsistent creative accounting scores. Creative accounting scores by some minority team members are commonly encountered cases of inconsistencies in teamwork assessment. To discuss the characteristics of the extended approach, mathematical equations and computations are presented and discussed with the help of typical inconsistent peer- and self-assessment scores.

2. Inconsistent peer- and self-assessment scores

Table 1 shows an example of creative accounting case of inconsistent peer- and self-assessment scores. While observing closely in Table 1, we can see that students A and E provided different total raw scores than the rest. In order to scrutinise and compare, we need to normalise these scores so that relative comparison becomes meaningful. Table 2 shows the normalised peer- and self-assessment scores, where sum of each assessor scores are equated to 100 by proportionally adjusting the raw scores.

Closely looking in Table 2, it is clear that assessors A, B, C and D's evaluation scores are consistent even though raw scores as well as total score given by assessor A were lower. However, assessor E provided inconsistent scores (in fact, over-rated his or her own score and under-rated peers' scores). All four assessors assessed that student E contributed less than others but his or her own scores are the opposite. This is a typical creative accounting case unless there is a clear evidence to suggest that E was discriminatorily assessed by other assessors in a prearranged way. Moreover, E's divergence from others is quite substantial and hence should not be considered as reliable as others' scores. If we ignore this inconsistency, the average individual contribution scores and individual contribution factors of assessee A, B, C, D, E are 18, 19.67, 23.67, 19.67, 19 and 0.90, 0.98, 1.18, 0.98, 0.95 respectively. This is clearly unfair for all assessee but E. One the other hand, if we completely dismiss the inconsistent scores provided by assessor E, the average individual contribution and individual contribution factors of assessee A, B, C, D, E are 20, 20, 25, 20, 15 and 1.00, 1.00, 1.25, 1.00, 0.75 respectively. This may be unfair for E as his or her assessment scores are not incorporated. It is important to take into account assessor E's scores to some extent albeit with lower reliability. The proposed extended relative relevance approach based on standard normal probability addresses this issue of creative accounting inconsistency.

Table 1. Raw peer- and self-assessment scores

Raw Scores		Assessee					
	Student	A	B	C	D	E	Total
Assessors	A	16	16	20	16	12	80
	B	20	20	25	20	15	100
	C	20	20	25	20	15	100
	D	20	20	25	20	15	100
	E	12	22	22	22	42	120

Table 2. Normalised peer- and self-assessment scores

Normalised Scores		Assessee					
	Student	A	B	C	D	E	Total
Assessors	A	20	20	25	20	15	100
	B	20	20	25	20	15	100
	C	20	20	25	20	15	100
	D	20	20	25	20	15	100
	E	10	18.33	18.33	18.33	35	100
Average score (All)		18	19.67	23.67	19.67	19	
Average score (without E)		20	20	25	20	15	

3. Existing approaches to refine inconsistent peer- and self-assessment scores

Previous studies have proposed a number of methods to refine and adjust inconsistent peer- and self-assessment scores in order to calculate individual contribution factors (*ICF*). These methods are discussed using consistent mathematical equations here.

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

$$ICR_j = \sum_{i=1}^{i=N} s'_{ij} \qquad \forall j \in N \tag{1}$$

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 assessee and dividing it by the number of members in a team (*N*) as in Equation (2):

$$\overline{ACR} = \frac{1}{N}\sum_{j=1}^{j=N} ICR_j = \frac{1}{N}\sum_{i=1}^{i=N}\sum_{j=1}^{j=N} s'_{ij} \ \forall j \in N \tag{2}$$

3.1 Method 1: Basic individual contribution factor (ICF)

Conway, et al. [20] proposed a method to estimate individual contribution factor (*ICF*) using peer- and self-assessment scores. *ICF* of an assessee (ICF_j) is obtained by calculating the ratio of total individual contribution rating of an assessee (ICR_j) and average contribution rating ($\overline{ACR}$) using Equation (3).

$$ICF_j = \frac{ICR_j}{\overline{ACR}} = N \times \frac{\sum_{i=1}^{i=N} s'_{ij}}{\sum_{i=1}^{i=N}\sum_{j=1}^{j=N} s'_{ij}} \quad \forall j \in N \tag{3}$$

Equation (3) is the fundamental equation to compute individual contribution factor (ICF_j). Example computation is provided in Table 3

As previously discussed, total peer- and self-assessment scores given by assessor A (= 80) and assessor E (= 120) are different than other assessors B, C and D (= 100). This has impacted *ICF* of all assessee. In fact, under-raters (for example, assessor A) or over-raters (for example, assessor E) have a higher influence on *ICF*, which is not fair. In Table 3, assessor E inflated his or her individual contribution factor close to an average contribution (note that *ICF* = 1.00 for an average contribution) by over-rating his or her self-assessment scores and under-rating peers' scores. Assessor A received the lowest contribution factor because he or she under-rated all assessee and also assessor E excessively under-rated him or her. Hence, the issues of this method include, (i) it neither evaluates the reliability of peer- and self-assessment scores nor adjust them, (ii) not all assessors may use the same rating scale, with some being more or less generous than the others, which may result in unfair *ICF* and, (iii) subjectivity in the judgement (for example, good for some assessors may mean 70 out of 100 whereas good for other assessors may mean 90 out of 100) may result in unfair *ICF*. Hence, Method 1 needs to be avoided.

3.2 Method 2: Normalised individual contribution factor (n_ICF)

Li [6] proposed a methodology to normalise inconsistent peer- and self-assessment scores. The originally proposed normalisation process was somewhat complicated. However, what normalisation process does is that it proportionally adjusts each assessor's total score, sum of which equates to any numerical constant, *C*, (say 1 or 100) as shown in Equation (4). That means,

$$\sum_{j=1}^{j=N} s'_{ij} = C \qquad \forall i \in N \tag{4}$$

Hence, the normalised scores given by an assessor *i* to an assessee *j* for his or her contribution to a teamwork can be computed by Equation (5).

$$s_{ij} = C \times \frac{s'_{ij}}{\sum_{j=1}^{j=N} s'_{ij}} \quad \forall i, j \in N \tag{5}$$

It then follows the same process as in Method 1 to compute normalised individual contribution factors (*n_ICF*) as in Equation (6).

Table 3. Basic *ICF* from peer- and self-assessment scores

Raw Scores		Assessee (*j*)					Total
	Student	A	B	C	D	E	
Assessors (*i*)	A	16	16	20	16	12	80
	B	20	20	25	20	15	100
	C	20	20	25	20	15	100
	D	20	20	25	20	15	100
	E	12	22	22	22	42	120
Total individual contribution rating (ICR_j)		88	98	117	98	99	500
Average contribution rating ($\overline{ACR}$)		100					
Method 1: Individual contribution factor ($ICF_j = \frac{ACR_j}{\overline{ACR}}$)		0.88	0.98	1.17	0.98	0.99	

$$n\_ICF_j = N \times \frac{\sum_{i=1}^{i=N} s_{ij}}{\sum_{i=1}^{i=N}\sum_{j=1}^{j=N} s_{ij}} \quad \forall j \in N \qquad (6)$$

The normalisation process makes all assessors' assessment comparable indicators of relative contributions of each member [21]. If the peer- and self-assessment scores are based on norm-based assessment criteria, where team members are asked to distribute a pool of mark out of a numerical constant C, (say 100 or 1), it is not necessary to conduct a normalisation process as the scores are already normalised. Table 4 summarises the computation of normalised individual contribution factors.

In Table 4, the assessors A and E's scores have been normalised (proportionally adjusted) to make their sum to a constant (100, in this example). Compared with Method 1, the normalisation process has made some adjustments in *ICF* by reducing *ICF* of over-raters (for example, assessor E) and by increasing the *ICF* of under-raters (for example, assessor A) to some extent. Because of the normalisation, the *ICF* of assessee B, C and D are not that much affected. However, as Method 1, Method 2 also does not take into account the reliability of peer- and self-assessment scores and the estimated *n_ICF* are still not fair. This method is also not recommended but clearly better than Method 1.

The remaining methods discussed below can be employed either to raw scores (Equation 3) or to normalised scores (Equation 6). However, it is recommended to use normalised scores (Equation 6) as it pre-adjusts, to some extent, the scores given by over-raters or under-raters and makes the scores comparable to each other.

3.3 Method 3: Scaled individual contribution factor (ϕ_ICF)

In order to adjust the spread of *ICF*, Conway, et al. [20] proposed scaled individual contribution factor (ϕ_*ICF*). The ϕ_*ICF* is calculated using Equation (7).

$$\phi\_ICF_j = n\_ICF_j - \phi\left(1 - n\_ICF_j\right) \quad \forall j \in N \qquad (7)$$

The value of scale factor ϕ can be varied widely but normally from 0.00 (there is no change in the spread of *ICF* computed from Equation 6 and hence the same as Method 2) to 1.00 (there is no spread at all as all *ICF* are equal to 1.00 for all team-members). Fig. 1 shows the influence of scale factor (ϕ) on *ICF*.

Selecting a value of scale factor (ϕ) more than 1.00 rewards below-average contributors and penalises above-average contributors (should therefore be strongly discouraged). Selecting a value of scale factor less than 0.00 rewards above-average contributors excessively and penalises below-average contributors significantly (should therefore be discouraged in order to prevent individualistic or do-it-all behaviours).

The ϕ_*ICF* computed from Equation (7) for a scale factor of 0.5 are provided in Table 5.

Similar to Method 1 and Method 2, Method 3 does not check the reliability of assessors' scores, it just adjusts the spread of contributions. Moreover, there is no basis of choosing an appropriate value of scale factor (ϕ) except balancing the spread as per an academic staff's judgement or wish.

3.4 Method 3.1: Agreement-corrected individual contribution factor (ac_ICF).

Neus [22] proposed an agreement-corrected individual contribution factor (*ac_ICF*), which is similar to ϕ_*ICF* where scale factor (ϕ) is the ratio of standard deviation of an assessee and the maximum standard deviation of all assessee as in Equation (8).

$$ac\_ICF_j = n\_ICF_j - \frac{\sigma_j}{\sigma_{max}}\left(1 - n\_ICF_j\right) \quad \forall j \in N \qquad (8)$$

The *ac_ICF* computed from Equation (8) are provided in Table 6. They are similar to Method 3 and

Table 4. Normalised individual contribution factor (*n_ICF*)

Normalised Scores		Assessee (*j*)					Total
	Student	A	B	C	D	E	
Assessors (*i*)	A	20	20	25	20	15	100
	B	20	20	25	20	15	100
	C	20	20	25	20	15	100
	D	20	20	25	20	15	100
	E	10	18.33	18.33	18.33	35	100
Normalised total individual contribution rating (n_ICR_j)		90	98.33	118.33	98.33	95	500
Normalised average contribution rating ($n\_\overline{ACR}$)		100					
Method 2: Normalised individual contribution factor ($n\_ICF_j = \frac{n\_ICR_j}{n\_\overline{ACR}}$)		0.90	0.98	1.18	0.98	0.95	

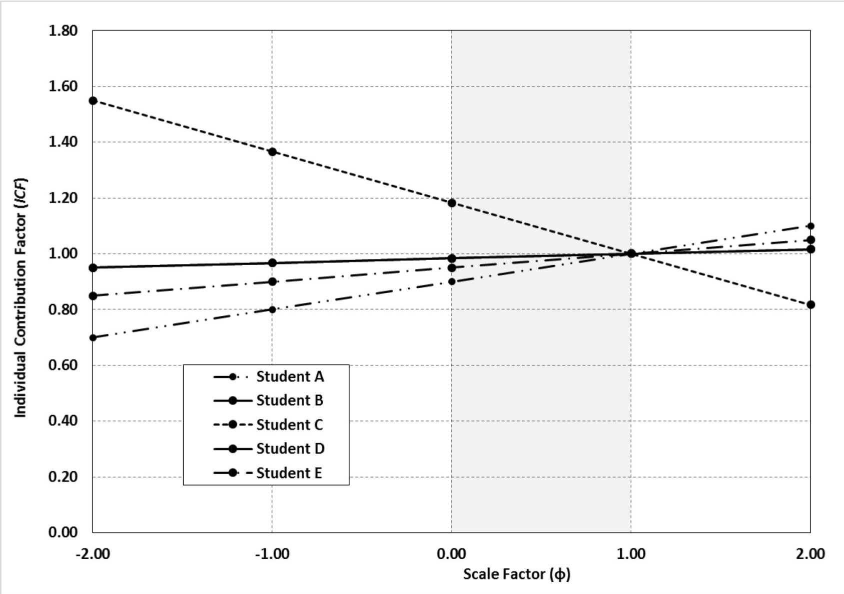

Fig. 1. Influence of scale factor (ϕ) on *ICF.*

Table 5. Scaled *ICF* (ϕ_*ICF*)

Scaled *ICF* (ϕ = 0.5) Conway, et al. [20]	**Assessee (*j*)**				
	A	**B**	**C**	**D**	**E**
Method 3: Scaled *ICF (ϕ_ICF)*	0.95	0.98	1.12	0.98	1.00

Table 6. Agreement-corrected *ICF* (*ac_ICF*)

Agreement corrected ICF Neus [22]	**Assessee**				
	A	**B**	**C**	**D**	**E**
Standard deviation (σ_j)	4.00	0.67	2.67	0.67	8.00
Scale factor (= σ_j/σ_{max})	0.50	0.08	0.33	0.08	1.00
Method 3.1: ac_ICF (agreement corrected *ICF*)	0.95	0.98	1.12	0.98	1.00

hence Method 3.1 has the same issues. Moreover, there is no logic in assigning an *ICF* equal to 1.00 (average contribution) for those assessees whose standard deviation is the maximum (that means there is a little consistency of an assessee' scores given by the peers). Even though assessee A's contribution is increased to some extent as required, assessee E still has higher individual contribution factor due to his or her inflated self-assessment score and deflated peer-assessment scores.

Both Method 3 and Method 3.1 are not recommended.

3.5 Method 4: Reliability weight-based individual contribution factor (it_ICF)

Ko [23] proposed a reliability weight-based *ICF* (*it_ICF*) which assigns differential weights to assessors based on their relative reliabilities. The method of computation is described below.

Iterative individual contribution factor (*it_ICF*) based on reliability weights can be computed by Equations (9–14):

$$it\_ICF_j = \frac{n\_ICR_j}{n\_\overline{ACR}} = N \times \frac{n\_ICR_j}{\sum_{j=1}^{j=N} n\_ICR_j} \quad \forall j \in N \tag{9}$$

where,

$$n\_ICR_j = \sum_{i=1}^{i=N} w_i s_{ij} \qquad \forall j \in N \tag{10}$$

where,

w_i is the reliability weight of assessor *i*, and
s_{ij} is the normalised peer- and self-assessment score given by assessor *i* to an assessee *j*.

The sum of all reliability weights (w_i) of all assessors can be equated to 1.00.

$$\sum_{i=1}^{i=N} w_i = 1 \tag{11}$$

The reliability weights (w_i) are calculated using relative relevance as:

$$w_i = \frac{r_i}{\sum_{j=1}^{j=N} r_j} \tag{12}$$

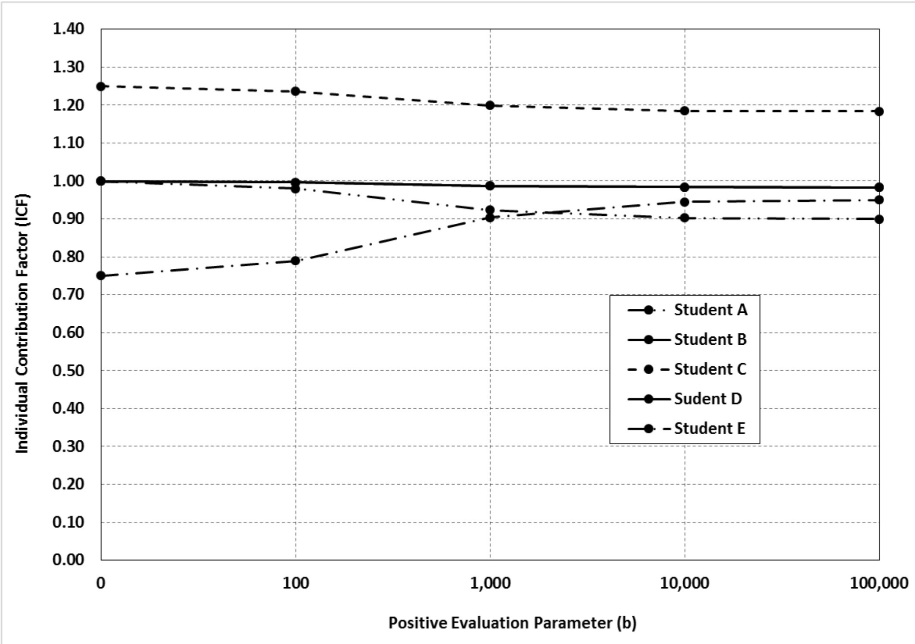

Fig. 2. Influence of positive evaluation parameter (b) on *ICF*.

Table 7. Reliability weight-based *ICF* (*it _ICF*)

Reliability weight-based ICF *(b = 100)* Ko [23]	Assessee (*j*)				
	A	B	C	D	E
Method 4: Reliability weight-based *ICF (it_ICF)*	0.95	0.98	1.12	0.98	1.00

The relative relevance has an inverse relationship with the divergence:

$$r_i = \frac{\sum_{j=1}^{j=N} d_j}{d_i} \tag{13}$$

An assessor's divergence can be estimated using variance as:

$$d_i = b + \frac{1}{N} \times \sum_{j=1}^{j=N} (s_{ij} - ICR_j)^2 \tag{14}$$

where,

b is a positive evaluation parameter.

When the variance of all assessors are equal, the final values of reliability weights of all assessors are equal meaning that all assessors are equally reliable. An iterative computational process is required to identify the final values of reliability weights as *ICR* depends on reliability weights and reliability weights depend on *ICR*. Moreover, reliability weights are highly sensitive to positive evaluation parameter, b and outliers are quickly dismissed when the evaluation parameter is set to a small positive constant. There is no method to accurately identify the realistic value of evaluation parameter. Fig. 2 shows the influence of positive evaluation parameter (b) on *ICF*.

When the evaluation parameter is extremely large (say, ∞), reliability weights of all assessors are equal and hence the Method 4 is similar to Method 2. When the evaluation parameter is zero (0), even a very small divergence from the mean (outliers) are quickly dismissed as shown in Fig. 2. Even though, a good selection of positive evaluation parameter (b) may result fairer individual contribution factors, Method 4 has a number of issues. First, an iterative process is required. Second, the method is highly sensitive to the positive evaluation parameter, b and scale used for the scores. Finally, there is no logic to assume that an assessor will be similarly reliable or unreliable to all assessee (an assessor can be reliable to an assessee but not to another assessee). This method is recommended to use with care.

The *it_ICF* computed from Equation (9) for b = 100 are provided in Table 7.

An extended relative relevance approach is proposed to address these issues in the next section.

4. Method 5: Relative relevance approach

As there are a number of issues in iterative process of estimating reliability weights proposed by Ko [23] in Method 4, we can make use of standard normal probability to estimate relative relevance of scores and hence relative relevance-based *ICF* (*rr_ICF*). The method of computation is described below.

Relative relevance-based *ICF* (*rr_ICF*) can be computed by using Equation (15).

$$rr\_ICF_j = \frac{n\_ICR_j}{n\_\overline{ACR}} = N \times \frac{n\_ICR_j}{\sum_{j=1}^{j=N} n\_ICR_j} \quad \forall j \in N \tag{15}$$

where,

$$n\_ICR_j = \sum_{i=1}^{i=N} w_{ij} s_{ij} \quad \forall j \in N \tag{16}$$

where,

w_{ij} is the reliability weight of assessor i to an assessee j.

Reliability weights can be calculated by estimating relative relevance, which in turn, has an inverse relationship with the divergence (how far a particular score is from the mean). The farther the divergence from the mean the lower the reliability weight. Reliability weights can be calculated using relative relevance as:

$$w_{ij} = \frac{r_{ij}}{\sum_{i=1}^{N} r_{ij}} \quad \forall i, j \in N \tag{17}$$

Relative relevance (r_{ij}) can be directly estimated from standard normal probability of z-score as:

$$r_{ij} = \phi\left(|z_{ij}|\right) \quad \forall\, i, j \in N \tag{18}$$

where,

$$|z_{ij}| = \left| \frac{s_{ij} - \frac{\sum_{i=1}^{i=N} s_{ij}}{N}}{\frac{\sum_{i=1}^{i=N}\left(s_{ij} - \frac{\sum_{i=1}^{i=N} s_{ij}}{N}\right)^2}{N}} \right| \tag{19}$$

The relationship between relative relevance (standard normal probability) and z-score is provided in Fig. 3. When a z-score of a 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.

Relative relevance (r_{ij}) varies from 0.40 (when z-score is 0) to 0 (when z-score is extremely large). However, the relative relevance becomes very small when z-score is more than 2. This approach is a little more generous for scores which are close to the mean. The *rr_ICF* computed from Equation (15) are provided in Table 8.

Method 5 can be a good approach to accommodate small differences as there is always a subjectivity component in peer- and self-assessments and it is not fair to penalise for small divergences. On the

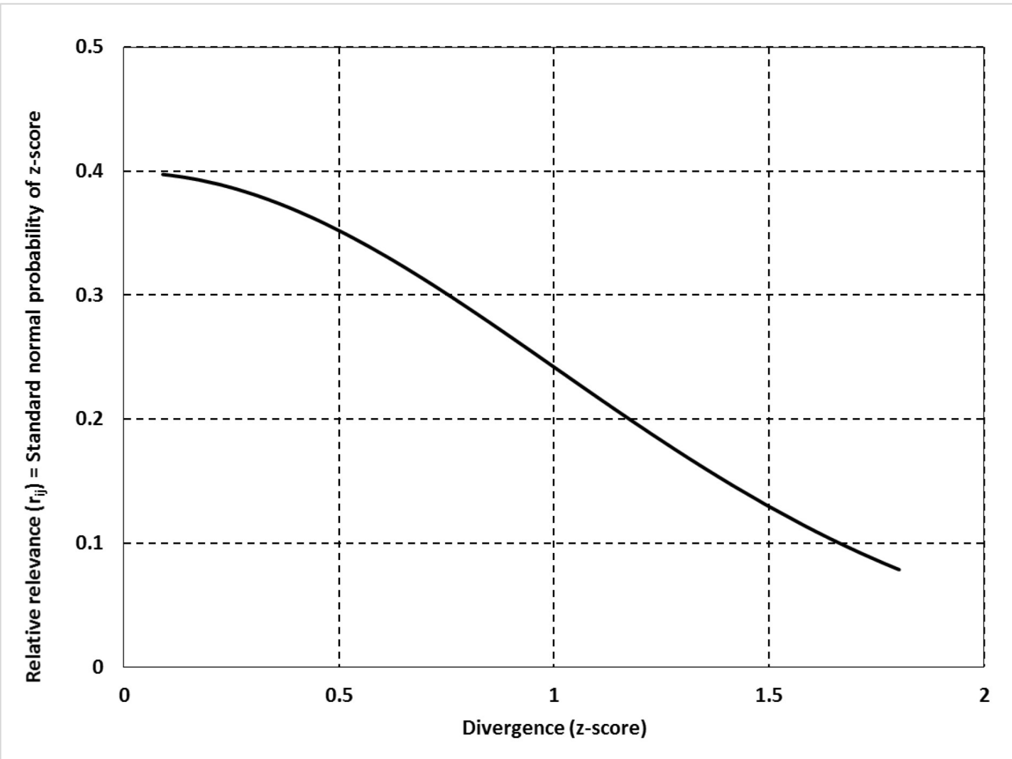

Fig. 3. Relationships between relative relevance (standard normal probability) and z-score.

Table 8. Relative relevance-based *ICF* (*rr_ICF*)

Relative relevance-based *ICF*	Assessee				
	A	B	C	D	E
Method 5: Relative relevance-based *ICF* (*rr_ICF*)	0.98	1.00	1.24	1.00	0.79

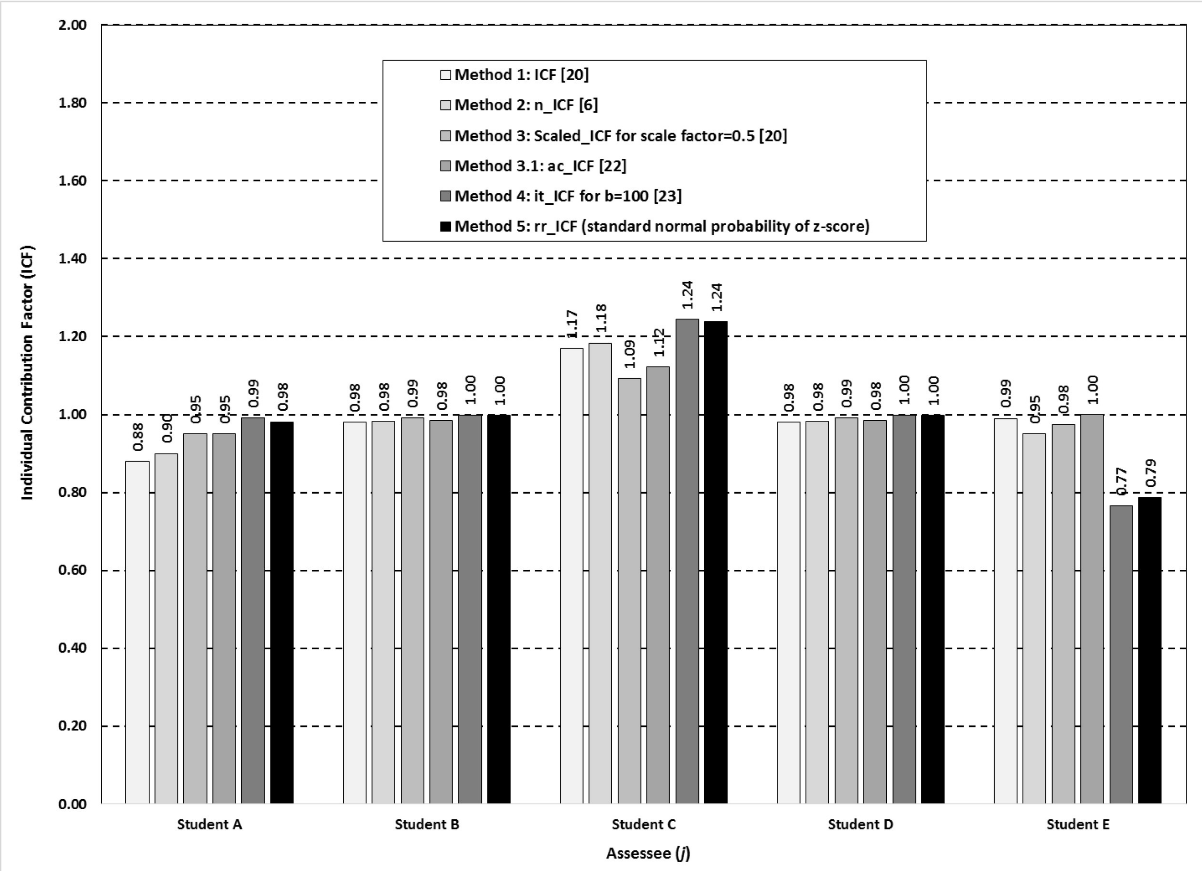

Fig. 4. Comparison of different methods.

other hand, this method also makes sure that minority assessments are never dismissed although relative relevance of them diminish with the increase of divergence (z-score). This method can be the best approach to refine creative accounting inconsistencies in peer- and self-assessment scores and hence individual contribution factors.

5. Comparison of different methods

Fig. 4 shows the comparisons among different methods for inconsistent peer- and self-assessment scores discussed previously in order to estimate individual contribution factor (*ICF*). Reliability-based methods (Method 4 and Method 5) adjust creative accounting cases of inconsistencies automatically and provide fairer individual contribution scores/factors. Method 5 is superior to Method 4 in that it is based on solid statistical theory (standard normal probability), does not require an assumption on positive evaluation parameter (b) and, does not require computer iterations.

6. Discussion

The analysis, evaluation and comparison of several methods to refine creative account cases of inconsistent peer- and self-assessment scores discussed in previous sections provide some interesting observations on their relative merits. Table 9 summarises the advantages and disadvantages of these methods.

It is clear from comparison in Table 9 that reliability and relevance-based methods (Method 4 and Method 5) adjust creative accounting cases of inconsistencies fairly and systematically. Method 5 is superior to Method 4 for a number of reasons—it is based on fundamental statistical method, does not require to use values based on judgment, and does not require computer iteration. However, as do all

Table 9. Advantages and disadvantage of different methods

Methods	Proposed by	Advantages	Disadvantages
Method 1: Basic individual contribution factor (*ICF*)	Conway, et al. [20]	• Easy to implement	• Does not check and refine inconsistencies • Influenced by rating scales used • Unfair
Method 2: Normalised individual contribution factor (*n_ICF*)	Li [6]	• Relatively easy to implement • Is not influenced by rating scales used	• Does not check and refine inconsistencies • Unfair
Method 3: Scaled individual contribution factor (ϕ*_ICF*)	Conway, et al. [20]	• Flexibility in adjusting scores by varying scale factor • Can be fair if appropriate scale factor is used	• Does not check and refine inconsistencies • No basis for choosing appropriate scale factor
Method 3.1: Agreement-corrected individual contribution factor (*ac_ICF*)	Neus [22]	• Removes the ambiguity in selecting appropriate scale factor	• Does not check and refine inconsistencies • No logic to consider assessee with highest standard deviation as a bench mark
Method 4: Reliability weight-based individual contribution factor (*it_ICF*)	Ko [23]	• Checks and refines inconsistencies using reliability weights • Fair if correct evaluation parameter is chosen	• Requires computer iterations • Reliability weights are highly influenced by an evaluation parameter • There is no method to accurately identify the realistic value of evaluation parameter
Method 5: Relative relevance-based individual contribution factor (*rr_ICF*)	This paper	• Checks and refines inconsistencies based on their relative relevance • Bases on sound statistical theory of deviation • Accommodates small differences in scores • No computer iterations required • Fair	• Requires basic knowledge of statistical theory

methods discussed in this paper, it does not address the issue of discriminatory scores where a majority number of team members go against a minority number of team members in a prearranged way.

7. Conclusion

This study presents an extended relative reliability approach to refine inconsistent peer- and self-assessment scores in a teamwork and to estimate individual contribution factors. Individual contribution factors are commonly multiplied by team mark to convert team mark into individual marks, provided the scores are valid (the degree to which the scores measure the true contributions) and reliable (the extent to which the scores are consistent). This study only addresses creative accounting cases of inconsistencies. Using consistent mathematical equations, computations and graphical representations, the proposed extended method is compared with several existing methods. Elaborated discussions on comparative evaluation of different methods are presented with the help of a typical example of inconsistent peer- and self-assessment scores. The comparison clearly shows that the proposed extension has a number of advantages over existing methods as it uses a sound statistical theory to calculate the relative relevance (or relative reliability) of peer- and self-assessment scores of an assessee and refines inconsistencies fairly (by considering scores based on relative relevance, by allowing some leeway for minor differences from the average and by not dismissing inconsistencies completely) and by refining the inconsistencies automatically (no assumptions or computer iterations required). It is important that the proposed method is discussed with the students in advance so that they know how their contributions to a teamwork are assessed and considered in awarding individual marks. This is because the students need to understand the implications of their inconsistent peer- and self-assessment scores so that they behave strategically which can help to minimise teamwork problems. When implemented with such care, the proposed approach can have positive impacts on both teamwork process and teamwork product. Further studies would help to identify the significance of such impacts and validation of the approach. It would also be interesting to look at industry models for team management.

References

1. K. P. Nepal, Simplified Framework for Managing Team Learning in Engineering Subjects, *International Journal of Engineering Education*, **32**(3A), 2016, pp. 1182–1193.
2. L. C. Schmidt, Engineering teams: Individual or group sport?, *International Journal of Engineering Education*, **22**(3), 2007, pp. 659–664.
3. R. Fruchter, Dimensions of teamwork education, *International Journal of Engineering Education*, **17**(4/5), 2001, pp. 426–430.
4. K. P. Nepal, Comparative evaluation of PBL and traditional lecturebased teaching in undergraduate engineering courses : evidence from controlled learning environment, *International Journal of Engineering Education*, **29**(1), pp. 17–22, 2013.
5. R. S. Hansen, Benefits and problems with student teams: Suggestions for improving team projects, *Journal of Education for Business*, **82** (1), 2006, pp. 11–19.
6. L. K. Y. Li, Some refinements on peer assessment of group projects, *Assessment & Evaluation in Higher Education*, **26**(1), 2001, pp. 5–18.
7. W. Brewer and M. I. Mendelson, Methodology and metrics for assessing team effectiveness, *International Journal of Engineering Education*, **19**(6), 2003, pp. 777–787.
8. M. Lejk, M. Wyvill and S. Farrow, A survey of methods of deriving individual grades from group assessments, *Assessment & Evaluation in Higher Education*, **21** (3), 1996, pp. 267–280.
9. G. B. Kilic and M. Cakan, The analysis of the impact of individual weighting factor on individual scores, *Assessment & Evaluation in Higher Education*, **19**(3), 2006, pp. 301–314.
10. D. Boud and N. Falchikov, Aligning assessment with long-term learning, *Assessment & Evaluation in Higher Education*, **31**(4), 2006, pp. 399–413.
11. K. P. Nepal, An approach to assign individual marks from a team mark: the case of Australian grading system at universities, *Assessment & Evaluation in Higher Education*, **37**(5), 2012, pp. 555–562.
12. M. Alias, A. Masek and H. H. M. Sallehm, Self, Peer and teacher assessments in problem based learning: Are they in agreements, *Procedia-Social and Behavioral Science*, **204**, 2015, pp. 309–317.
13. M. D. Lew, W. A. M. Alwis and H. G. Schmidt, Accuracy of students' self-assessment and their beliefs about its utility, *Assessment & Evaluation in Higher Education*, **35**(2), 2010, pp. 135–156.
14. K. Cho, C. D. Schunn and R. W. Wilson, Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives, *Journal of Educational Psychology*, **98** (4), 2006, pp. 891–901.
15. W. Cheng and M. Warren, Having second thoughts: student perceptions before and after a peer assessment exercise, *Studies in Higher Education*, **22**(2),1997, pp. 233–239.
16. D. M. Sluijsmans, G. Moerkerke, J. J. Van Merrienboer and F. J. Dochy, Peer assessment in problem based learning, *Studies in educational evaluation*, **27** (2), 2001, pp. 153–173.
17. K. J. Topping, Trends in peer learning, *Educational Psychology*, **25** (6), 2005, pp. 631–645.
18. K. Pond, Learning to assess students using peer review, *Studies in Educational Evaluation*, **23**(4), 1997, pp. 331–348.
19. K. J. Topping, E. F. Smith, I. Swanson and A. Elliot, Formative peer assessment of academic writing between postgraduate students, *Assessment & Evaluation in Higher Education*, **25**(2), 2000, pp. 149–169.
20. R. Conway, D. Kember, A. Sivan and M. Wu, Peer assessment of an individual's contribution to a group project, *Assessment & Evaluation in Higher Education*, **18**(1), 1993, pp. 45–56.
21. S. Sharp, Deriving individual student marks from a tutor's assessment of group work, *Assessment & Evaluation in Higher Education*, **31** (3), 2006, pp. 329–343.
22. J. L. Neus, Peer assessment accounting for student agreement, *Assessment & Evaluation in Higher Education*, **36**(3), 2001, pp. 301–314.
23. S. S. Ko, Peer assessment in group projects accounting for assessor reliability by an iterative method, *Teaching in Higher Education*, **19** (3), 2014, pp. 301–314.

Kali Prasad Nepal is a senior lecturer at School of Engineering, Deakin University, Australia. His research activities in engineering education concentrate on learning system components and teamwork-based learning in engineering subjects at engineering schools.