

Monte Carlo evaluations of methods of grade distribution in group projects: simpler is better

Sebastián G. Guzmán

To cite this article: Sebastián G. Guzmán (2018) Monte Carlo evaluations of methods of grade distribution in group projects: simpler is better, *Assessment & Evaluation in Higher Education*, 43:6, 893-907, DOI: [10.1080/02602938.2017.1416457](https://doi.org/10.1080/02602938.2017.1416457)

To link to this article: <https://doi.org/10.1080/02602938.2017.1416457>



View supplementary material [↗](#)



Published online: 24 Jan 2018.



Submit your article to this journal [↗](#)



Article views: 336



View related articles [↗](#)



View Crossmark data [↗](#)



Monte Carlo evaluations of methods of grade distribution in group projects: simpler is better

Sebastián G. Guzmán^{a,b} 

^aDepartment of Anthropology and Sociology, West Chester University of Pennsylvania, West Chester, PA, USA;

^bUniversidad Andres Bello, Facultad de Educación, Santiago, Chile

ABSTRACT

Group projects are widely used in higher education, but they can be problematic if all group members are given the same grade for a project to which they might not have contributed equally. Most scholars recommend addressing these problems by awarding individual grades, computing some kind of individual weighting factor (IWF) from peer and (sometimes) self-assessments, which is then multiplied by the group grade to generate an individual grade. Several variants of the IWF method have been proposed, sometimes with complex algorithms. However, theory suggests they are inaccurate and their accuracy has not been evaluated. This article uses Monte Carlo experiments to assess the accuracy of the original IWF method and variants proposed in the past decade. Findings show that the earlier, simpler methods work best and that self-assessments should definitely be avoided.


KEYWORDS


Peer assessment; self-assessment; group projects; individual weighting factor; Monte Carlo

Introduction

Group projects are a common type of assignment in higher education today. They can help students develop teamwork skills and confront the challenges of undertaking large and complex assignments they could not complete individually in a semester, and they require less grading effort from instructors, among other well-known benefits (see Jaques and Salmon 2007). However, there are common problems with these assignments, most notably free-riding and unfair grades, which can have severe implications on students' motivation to work and the group learning processes. To avoid these problems, several scholars have suggested ways of computing individual grades for individual contributions to group projects (e.g. Goldfinch and Raeside 1990; Conway et al. 1993; Goldfinch 1994; Lejk and Wyvill 2001a, 2001b; Li 2001; Tu and Lu 2005; Sharp 2006; Zhang and Ohland 2009; Neus 2011; Ko 2014; Spatar et al. 2015). One of the most common techniques is the computation of some variety of individual weighting factor (IWF), which is multiplied by the group grade to calculate each member's individual grade. Recent developments of the technique propose variations of the IWF formula that aim to avoid the unfairness of biased evaluations and other distortions of the final grade. To this end, scholars have replaced the original simple average of peer- and self-assessments by more sophisticated algorithms (e.g. Li 2001; Bushell 2006; Sharp 2006; Neus 2011; Ko 2014; Spatar et al. 2015).

One of the limitations of these developments is that we do not know how accurate they are because the bars against which to measure the final grade distortions introduced by each approach are imprecise. To evaluate each approach, scholars typically focus on special cases in which distortions are evident,

CONTACT Sebastián G. Guzmán  sguzman@wcupa.edu, sebastian.guzman.r@unab.cl

 The supplemental data for this article is available online at <https://doi.org/10.1080/02602938.2017.1416457>.

see if there are correlations between individuals' grades in individual work and group work, or simply check if students like the system (e.g. Baker 2007; Zhang and Ohland 2009; Jin 2012). However, these methods are not the best way to test algorithms. In several scholars' case analyses it is unclear which assessments are distorted, and it is therefore difficult to claim that one method is better than the other. Additionally, the correlation between the students' real and estimated work may be affected by an error that is correlated with the IWF or its variants' estimation. Given the number of assessments involved in computing IWF variants and the large number of variables intervening in the outcome, it is difficult to predict the result in many scenarios.

Zhang and Ohland (2009) tackle these issues by evaluating older variants of the IWF with Monte Carlo experiments. The experiments consist of simulating a large sample of real contributions of students and of errors in their peer and self-assessments, and then comparing IWF variants to each student's real contribution. In this article, I use the same approach to evaluate classic and more recent and sophisticated methods, adding two important methodological improvements. First, my analysis focuses not only on the average distortions each method produces, but also on whether distortions tend to favour students who work less and those who distort assessments to boost their grades. Second, my experiments are run in different scenarios of group size, dispersion of contributions, assessment errors and self-assessment inflation, making their results more robust.

Findings show that the original IWF and the normalised IWF (NIWF) introduced by Sharp (2006), when used excluding self-assessments, are more accurate than the more complex recent methods. I also find that their distortions are small enough that the methods can be safely used to estimate individual contributions.

Peer assessment of group projects

It is well-known that group projects have several benefits and are common today at all levels of education and across disciplines. Unfortunately, there are also major challenges associated with group projects, notably free-riding and unfair grades. Free-riding occurs when some students decide to expend less effort in the group project because they can rely on their peers' effort to obtain nearly the same grade. This may translate into some students learning less than they would otherwise, high workloads for other students, problems in group management and interpersonal relationships, and students' dislike for group assignments (Webb 1995). Unfair grades exist because some group members work and learn more than others, but will be assessed for the group's results. This can translate into a disincentive to work more, dislike for group assignments, and students being passed without actually learning the skills or contents of the course. Studies show that free-riding and unfair grades are students' main concerns regarding group projects (Feichtner and Davis 1984; Macfarlane 2016), and scholars tend to agree these are serious problem that need to be addressed if we are to use group assignments. The vast majority of scholars addressing these problems suggest that the best solution is to derive individual grades for each member, combining instructor assessments of group performance and peer assessments of individual participation. As I explain below, these scholars offer several alternative methods to compute such individual grades.

Since the mid-1990s, most approaches to assigning individual grades for group projects involve variants of Goldfinch and Raeside's (1990) IWF method (for exceptions, see Tu and Lu 2005; Dommeyer 2012). The core of the method consists of computing an estimator of individual contributions (IWF) for each member of the group, which is then multiplied by the group's grade to obtain an individual grade. The different variants of the IWF for student *A* are a (sometimes weighted) index resulting from *A*'s peers' assessments of *A*'s work and sometimes *A*'s self-assessment. The IWF's variants attempt to address several potential problems of the original method, and of peer and self-assessments in general, most notably those resulting from biased assessments.

Bias is a serious concern. Lejk and Wyvill (2001a) find that students tend to favour themselves, with those who work less inflating their self-assessments more than those who contribute more. This leads to a pattern of boosting the grades of those who work less to the detriment of the rest. Additionally,

there can also be bias in favour or against some peers, and different bars against which to measure contributions, leading to over- or under-marking by some group members.

One way of addressing validity concerns is through methods that encourage members to make more valid peer assessments, that is, improving the raw data before an index is computed. There are four techniques along this line. First, assessments should be made confidential so that students can freely express criticism of peers that they would not be willing to express publicly (Lejk and Wyvill 2001a). This recommendation is broadly accepted and uncontroversial today. Second, students may be asked to offer qualitative comments justifying their scores, so that they have to reflect on them and they are less arbitrary, although it is time-consuming for instructors to consider the comments (Loddington et al. 2009). Third, students may be asked to assess each peer in several categories that reflect actual behaviour or to assess each peer holistically with only one assessment or score. Unfortunately, it is so not clear which method is best (Lejk and Wyvill 2001b; Sharp 2006; Ohland et al. 2012). Fourth, students can receive an incentive or penalty for providing valid or invalid assessments (Tu and Lu 2005); yet we do now know whether instructors and students will find such a policy acceptable.

Another important aspect, although neglected by the literature, is that studies often use scales starting from 1. This inflates the grades of all those who worked less. If *A* contributed about half of what *B* contributed and *B*'s contribution is the maximum of 5, the middle point in the 1–5 scale is 3, which is 60% of *B*'s contribution, not half. If the average contribution is 4, *A* would receive 1.25 and *B* would receive .75. By adapting the scale to go from 0 to 4, the middle point is 2, exactly 50% of *B*'s contribution. The average contribution would be 3, giving *A* an IWF of 1.33 and *B* only .67. In other words, with the wrong scale, *B*'s grade was artificially increased .08 times the group's grade and 12% of what should be her real grade, to the detriment of *A*.

A different but complementary approach to issues of validity involves improving the final computation of a member's mark with algorithms that weight and scale raw assessments in different ways or may even exclude self-assessments, which tend to be biased. This approach typically leads to variations of the IWF method.

The IWF method and its variants

Conway and colleagues' (1993) basic IWF for each student is computed as follows. First, add the assessments of member *A*'s contribution assigned for that student by her peers and herself. This can be done with whatever scale is used. We call this the individual effort rating (IER) for each student. Next, compute the average effort rating (AER) of the group, which is the average of all the IERs. Finally, for each student, divide the IER by the AER to obtain the IWF (see example in Table 1). We may express the IWF of student *j* assessed by assessors' *i*'s raw assessments (RA_{ij}) as:

$$IWF_j = \frac{IER_j}{AER} = \frac{\sum_{i=1}^n RA_{ij}}{AER} = \frac{\sum_{i=1}^n RA_{ij}}{\sum_{i=1}^n IER_i / n} = \frac{\sum_{i=1}^n RA_{ij}}{\sum_{i=1}^n (\sum_{k=1}^n RA_{ki}) / n}$$

Each student's IWF may then be multiplied by the group grade to compute an individual grade. Or the final individual grade may be a weighted average of the IWF-based grade and the group grade, with the weights set arbitrarily, usually 50% each.

One of the problems with this method is that some assessors are more generous than others, so they do not all use the same rating scale. This has two implications. If we exclude self-assessments, under-raters would artificially decrease their peers' IWF-esa (Neus 2011; -esa indicating 'excluding self assessments'). For instance, if everyone in a group of four did the same amount of work and all students mark their peers with 25, except for a student who rates them with 20, her IWF-esa would be 1.05, while everyone else would receive .98. And if self-assessments are included, over-raters have a higher influence in the final grade than their peers (Neus 2011). The best solution for this problem is to normalise each assessors' ratings so that all her assessments average (or add) one (or 100 or *N*) before computing the IER. This makes all assessors' evaluations comparable as indicators of relative contributions of each

Table 1. Example of IWF variants' results using Spatar et al.'s (2015, Table 5 data).

	Assessee (j)				Row sum	
	A	B	C	D		
Original raw assessments (RA)						
Assessor (i)						
A	20	20	20	20	80	AER [=ΣIER/n]
B	4	16	17	15	52	
C	4	15	18	14	51	
D	4	15	17	15	51	
IERs [=ΣRA]	32	66	72	64		58.5
IWF [=IER/AER]	.55	1.13	1.23	1.09		
IERs-esa (excl. self-assmnt.) [=ΣRA-esa]	12	50	54	49		41.25
IWF-esa [=IER-esa/AER-esa]	.29	1.21	1.31	1.19		
Normalised assessments (NAs) [RA/ΣRA _i]						
A	.25	.25	.25	.25	1.00	σ _i .00
B	.08	.31	.33	.29	1.00	.10
C	.08	.29	.35	.27	1.00	.10
D	.08	.29	.33	.29	1.00	.10
NIWF [=ΣNA _j /ΣNA _i]	.48	1.15	1.26	1.11		
ac-IWF computation using NAs						
Assessee standard deviation [s _j]	.09	.03	.05	.02		max(s _j) .07
IAF [=1 - s _{j(k)} /max(s _j)]	.00	.71	.47	.77		
ac-IWF [=IAF-esa × (NIWF -esa - 1) + 1]	1.00	1.10	1.12	1.08	4.31	
ASNIWF computations using NAs-esa						
SAIF [=1 - s _{j(k)} /(2 × max(s _j))]	.50	.85	.74	.89		
ASNIWF [=SAIF × (NIWF - 1) + 1]	.74	1.12	1.19	1.09	4.16	
Normalised assmnts., excl. self-assmnt. (NA-esa)						
A		0.33	0.33	0.33	1.00	σ _i .00
B	.11		0.47	0.42	1.00	.16
C	.12	0.45		0.42	1.00	.15
D	.11	0.42	0.47		1.00	.16
NIWF-esa [=ΣNA-esa/ΣNA _i]	.34	1.20	1.28	1.17	4.00	
ac-IWF-esa computation using NAs-esa						
Assessee standard deviation [s _j]	.00	.06	.08	.05		max(s _j) .08
IAF-esa [=1 - s _{j(k)} /max(s _j)]	.93	.23	.00	.37		
ac-IWF-esa [=IAF-esa × (NIWF-esa - 1) + 1]	.39	1.05	1.00	1.06	3.50	
ASNIWF-esa computations using NAs-esa						
SAIF-esa [=1 - s _{j(k)} /(2 × max(s _j))]	.96	.61	.50	.69		
ASNIWF-esa [=SAIF-esa × (NIWF-esa - 1) + 1]	.37	1.13	1.14	1.12	3.75	
it-IWF*	.33	1.19	1.34	1.14	4.00	
it-IWF using NAs*	.48	1.15	1.27	1.11	4.00	
it-IWF using b = μ(σ _i ²)*	.39	1.17	1.31	1.13	4.00	
it-IWF using NAs and b = μ(σ _i ²)*	.42	1.17	1.30	1.12	4.00	
it-IWF using NAs and b = μ(σ _i ²)/10 (i.e. it-IWF2)*	.34	1.19	1.34	1.13	4.00	
it-IWF-esa*	.31	1.19	1.33	1.16	4.00	
it-IWF2-esa*	.32	1.20	1.31	1.17	4.00	

*See Ko (2014) for computation details.

member (Sharp 2006). This is achieved simply by dividing each individual assessment by the sum of all the assessments made by the same assessor (see Table 1), or by asking students to split a set number of points, say 100. Spatar and colleagues (2015) call the IWF computed from normalised assessments the normalised IWF or NIWF (see Table 1). Thus, student *i*'s normalised assessment of *j*'s work can be computed as:

$$NA_{ij} = \frac{RA_{ij}}{\sum_{k=1}^n RA_{ik}}$$

Normalised assessments could also be amplified simply by multiplying the result in the formula by 100, n , $100/n$, or another constant. If this formula is used – as opposed, for instance, from asking students to split a given number of points – the (normalised) AER will be equal to n , the number of assessors. Replacing the (normalised) AER by n , the NIWF becomes simply the average of all NAs of a given student's work:

$$\text{NIWF}_j = \frac{\sum_{i=1}^n \text{NA}_{ij}}{n}$$

Following Lejk and Wyvill's (2001a) general recommendation of excluding self-assessments due to the tendency to inflate them, Sharp (2006) recommends excluding self-assessments in the normalisation process – note that in this case n , the number of assessments for a given student is the group size minus 1. It should be noted, however, that there is no strong agreement about whether to include self-assessments, and multiple scholars continue to use self-assessments (Zhang and Ohland 2009; Ohland et al. 2012; Carvalho 2013; Ko 2014).

Ko argues that self-assessment is better because excluding it involves the assumption that the assessor should be self-assessed as an average, favouring low-contribution students to the detriment of high-contribution students. However, Bushell (2006) notes a serious problem when the normalisation process excludes self-assessments: it involves the assumption that the assessor should be self-assessed as an average contributor, favouring low-contribution students to the detriment of high-contribution students. This distortion produced by the normalisation process is especially noticeable in smaller groups. Bushell proposes a solution that requires the instructor to make decisions on a case-by-case basis. Unfortunately, this solution is impractical because it cannot be easily automated as an algorithm.

Li (2001) had an idea similar to Sharp's, but Li's method leads to producing the same grade for all students if they make correct assessments. It is based on correcting the assessments if an assessor gives to others fewer points than what everyone receives on average – without considering self-assessments. This is avoided if the instructor manually corrects the model when she notices an assessor is right in giving fewer points to others. Yet, even after such case-by-case analysis, which is often difficult, everyone's grade comes closer to the group grade, favouring those who worked less. Unfortunately, it is difficult to notice this in the examples discussed by Li because opinions of students are inconsistent, with no agreed-upon ranking of contributors.

Several authors have offered alternatives to deal with the problem of inconsistent assessments. Sharp (2006) proposed a statistical test to decide whether the difference in estimated contributions is large enough and agreed-upon sufficiently to warrant any difference in final grades. Simplifying this proposal, Neus (2011) suggests using an 'agreement-corrected IWF' (ac-IWF) based on scaling each assessee's NIWF according to the level of agreement among assessments for that assessee. Only if there is significant agreement in the assessments of one assessee does that assessee's grade change. The ac-IWF for a given student can be computed as $\text{IAF} \times (\text{IWF} - 1) + 1$, where IAF is an individual agreement factor for that student. The IAF for a student k is calculated as $1 - s_{j(k)}/\max(s_j)$, where s_j is the (sample) standard deviation of assessments of each assessee, $\max(s_j)$ is the highest s_j , and $s_{j(k)}$ is the standard deviation of assessments of the assessee k 's work (see Table 1).

While it is reasonable to consider assessments as more valid if there is agreement among assessors, the idea that without agreement the grade should be the average is problematic: why are we to assume that in such a case the group average is the best estimate of that assessee's contribution? The peer assessments of that assessee's work may still be a better approximation than the average. Two of the main problems with Neus's logic are evidenced by a paradox: if everyone agrees that A did more work than the rest, A receives a smaller grade than the rest if they disagree about how much more work A did (Ko 2014, 304, 305). Additionally, if all but one member agrees on how much an assessee worked, the single divergent opinion, likely to be biased, can unfairly bring the assessee's grade close to the average (see Table 2 of Ko 2014; and Table 1 in this article).

Spatar et al. (2015) attempt to moderate the ac-IWF's sensitivity to biased assessments of free riders by damping the effect of the IAF, replacing it by a scaled IAF (SIAF) computed as $1 - s_{j(k)}/[2 \times \max(s_j)]$. Yet while the problems of the ac-IWF will be smaller in this case, they are likely to still occur because its logic is faulty. Beyond the ASNIWF, Spatar and colleagues also adopt a pre-existing method to deal

with grades above the permitted maximum, but this is unrelated to the question addressed here about estimating the proportion of the group work done by each student.

Ko (2014) addresses Neus's problems with a formula that weights each assessor's assessments by the assessor's reliability. The logic is that the existence of disagreements about an assessee does not imply all assessments of that assessee are invalid; some are more valid than others, and we should automate the process of identifying their validity as much as possible to avoid time-consuming case-by-case checks proposed by other methods.

Unfortunately, Ko's proposal still has some limitations, stemming from the formula for his 'iterative IWF' (it-IWF), which must be explained to grasp the problem. Ko's reliability weight factor depends on the distance between the assessor's assessments for each assessee and the mean of all assessments for each corresponding assessee. However, Ko notes that the mean is artificially affected by unreliable assessments. Thus, Ko proposes an iterative method: first, we compute a weighted average of assessments for each assessee assuming all assessors are equally reliable; next, we define the reliability of each assessor; then we repeat the process, but using an updated weighted average of assessments for each assessee, where the reliability factor defines the weight. We can then repeat the process, computing new reliability factors and again updated weighted averages, until convergence.

One problem of Ko's it-IWF is that it produces different results depending on the size of the scale used. For instance, using the data in Table 1, in which each assessment could range between 0 and 20, A's it-IWF is .33. But A receives .48 if we use normalise the assessments or ask students to indicate the proportion of the group's work done. The problem of inconsistency could be avoided by simply starting from normalised assessments and setting what a normal assessment would be beforehand; whether 1, $1/n$, 100 or $100/n$. But how are we to determine what is the proper scale, if all scales return different results? The more serious problem lies in how to set parameter b of Ko's algorithm, which defines the level of discrimination of outliers. Ko sets it as the mean of standard deviations of each assessor's assessments, $\bar{x}(\sigma_i)$. Since the rest of the formula is based on variance (σ_i^2) rather than standard deviation, there is a problem of scales, as one is the square of the other. One solution to this issue is to set b as the mean of assessors' variances instead: i.e. $\bar{x}(\sigma_i^2)$. This it-IWF with corrected b and based on normalised assessments produces consistent results regardless of the scale of normalised assessments. However, Ko provides no rationale to identify an optimum b , as the it-IWF formula is not grounded in statistical theory. To prevent inflation of the it-IWF through self-assessment inflation, it is better to set b at a proportion of the mean of variances, such as a tenth of it; a number suggested by some Monte Carlo tests I ran. I here call this variant it-IWF2.

Yet the it-IWF has another problem, which the it-IWF2 does not solve. Although the it-IWF gives less weight to inflated self-assessments, it still allows inflated self-assessments to inflate the final grade. There are two reasons for this. First, because each assessment is weighted by the average reliability of the assessor, a student who inflates only her own assessment but is consistent with peers' opinion in the rest of her assessments can have a high reliability. This is especially true when the rest of the peers are not completely consistent in their opinions and when self-assessment inflation is not extremely exaggerated. The high reliability of the student inflating her self-assessment will increase the weight of the biased self-assessment throughout all iterations. In Table 1, for instance, if student A had been a bit more honest and assessed her peers similarly to what they did and inflated her self-assessment less, she would have inflated her it-IWF considerably more. If she gave ratings of 9, 15, 17 and 15, respectively, her it-IWF would be .37 while her peers agree she deserved .31. The effect would be even higher if the peers disagreed more about B, C and D's contributions.

There is a second reason why the it-IWF and it-IWF2 allow for grade self-inflation. When peers are not consistent in assessing a student, such student's mean contribution will be boosted by an inflated self-assessment, and the boosted mean contribution will sometimes remain through iterations. For instance, assume all three members worked the same and agree on everything but on A's work, with B saying A contributed 110 and C saying A contributed 90. If A said she contributed 120 instead of 100, she would increase her it-IWF to 1.07 and her it-IWF2 to 1.06. Thus, iterations are not effective at dealing with biased assessments distorting the mean contribution from which we calculate assessments'

reliability when there is assessor disagreement. Furthermore, the it-IWF method's treatment of all disagreements as equal is problematic given the pattern of students inflating their self-assessments (Lejk and Wyvill 2001a).

The intuitive solution to the problem of self-assessment inflation is to eliminate self-assessments. However, this is still problematic. If assessments are not normalised, students have the incentive of deflating their peers' assessment to effectively inflate their own grade. Since we know students tend to inflate their self-assessments when allowed, we would expect them to deflate their peers if self-assessments are excluded. Thus, the it-IWF-esa would still be computed on the basis of distorted raw assessments. Those by low-contribution students would tend to be more biased, but the it-IWF-esa cannot identify this pattern and will tend to treat those moderately biased as more accurate than the unbiased assessments of high contribution students. If, on the other hand, assessments are normalised, we have the abovementioned problem noted by Bushell. Thus, there is no clear solution to this issue.

Beyond the mathematical problems of the it-IWF approach, its iterative method makes it somewhat impractical. Although the iteration is automated with an Excel macro, it requires some advanced knowledge of spreadsheets and some adjustments in the spreadsheet or even the macro programme code if dealing with multiple groups of different sizes. This, along with the complexity of the algorithm, which students may not understand, makes it unlikely that the method will be used as much as simpler methods, especially in the less mathematically oriented disciplines. To be worthwhile, the method would need to be substantially more accurate than its best alternative.

In summary, no method to date deals with self-assessment inflation and assessor disagreement properly, but it is unclear which produces the smallest distortions.

Methods

In real cases in which assessors disagree, it is impossible to know what the real contribution of each student was. This makes the evaluation of IWF variants' accuracy difficult, as there is no good bar against which to measure the IWF variants' estimate of the students' real contribution. Monte Carlo experiments circumvent this situation, providing a more robust assessment of each IWF variant. The method consists of estimating results from a large sample of machine-generated random data that follows a patterned distribution. We can do this starting from randomly generated 'real contributions' from which assessments deviate, which is what happens in real life if we assume there is something like a real contribution. This data allows the calculation of each IWF variant's 'estimate error', that is, how much it deviates from the real contribution. From this sample we can calculate statistics for each IWF variant's estimate error. Specifically, I look at the median and the 5th and 95th percentiles for each level of contribution between 0 and 1.55 rounded to .1, where a contribution of 1 (100%) indicates a fair share of the group's work and average contribution. Lejk and Wyvill (2001a, 2001b) show there are very few cases of more extremely high or low contributions, with 99.3% of the cases having a contribution between .55 and 1.45 in my main sample (see Supplementary Appendix 2).

Additionally, I look at how the error is distributed across levels of self-assessment inflation, to identify which methods favour students who inflate their self-assessments and how much. Finally, I also evaluate the Root Mean Square Error (RMSE), computed as $\sqrt{\sum E^2 / N}$, where E is the error and N the number of cases. While this statistic does not specify which students receive a larger error, it is a good summary of the overall error for the sample.

Because results can vary depending on the dispersion of the contributions and the assessments, I first generated data with a distribution that would closely match the dispersion of the real data reported by Lejk and Wyvill (2001a, 2001b, hereafter, L&W distribution) for groups of five: a standard deviation of contributions of 11.64% and a standard deviation of peer assessments of the same assessee of 2.88%, with a tendency to inflate self-assessments or deflate peer-assessments. This deviation of contributions produced an $\mu(s_{\text{Grp.IWF-esa}})$ of 9.56%. This is 1.77% higher than the category-based $\mu(s_{\text{Grp.IWF-esa}})$ of 7.79% reported by Lejk and Wyvill (2001a, 558), because holistic assessments produce a standard deviation

higher than category-based assessments of about 1.76%, based on estimations from Lejk and Wyvill (2001b) reported IWF data. The standard deviation of peer-assessment agreement I estimated from Lejk and Wyvill (2001b) holistic assessments data is 2.93%.

Considering an average contribution as 1 or 100% – with assessments varying in 5% increments – contributions are randomly assigned with a normal distribution capped at 0 and 200%, and later scaled so that the average for each group is 100%. The between-group heterogeneity of contributions within a group, the within-group contributions, and peer assessments of a single peer have a normal distribution.

The inflation of self-assessments tends to be greater among those contributing less, as reported by Lejk and Wyvill (2001a, 558), producing a difference between $\mu(s_{\text{IWF-esa[Grp]}})$ and $\mu(s_{\text{IWF[Grp]}})$ of about 1.28% in groups of five: $\mu(s_{\text{IWF-esa[Grp]}})$ and $\mu(s_{\text{IWF[Grp]}})$ indicate the between-group mean of within-group standard deviation of IWF-esa scores and of IWF scores, respectively. In theory, several distributions could match this requirement. With the formula I use (see Supplementary Appendix 1), the average self-inflation is 21.45% and the standard deviation of self-assessment inflation is 9.14% in the L&W distribution. Graphs and tables with more details about the distributions can be found in the online Supplementary Appendices 2–4 (all supplementary appendixes are available through the link tiny.cc/MonteCarlo).

I do most of the analysis with groups of four students with a similar distribution for three reasons: distortions are more evident in smaller groups, the literature commonly illustrates analyses in groups of four, and Monson (2017) finds that group of four or more are more effective than groups of three. Nonetheless, I repeat the analysis for smaller and larger group sizes.

Because the dispersion of the data may differ in each class, I repeat the test for three levels of agreement in the assessments, three levels of average dispersion of contributions, and three levels of self-assessment inflation. This should provide more robust findings. I vary the levels of agreement, dispersion of contributions, and self-assessment inflation to about half and to about double than the L&W distribution; but in the case of self-assessment inflation, to 0 and about 1.5 times the mean of the L&W distribution, to show more relevant and realistic cases. These factors are approximate, as the resulting distribution also depends on the combinations of other distribution variables and group sizes. The results for each IWF variant are rather similar across scenarios, so I only present some variations here with the most important differences. The rest can be replicated automatically by running the Stata code available in the online Supplementary Appendix 5.

Even if most classes were to approximate Lejk and Wyvill's data, it is possible that some of these other distributions are more realistic for another reason. The dispersion of contributions and the inflation of self-assessments were computed in relation to a 'real contribution' in Lejk and Wyvill's data, but on their IWF and IWF-esa results respectively. So if, for example, my results show that the IWF diminishes the dispersion of estimated contributions, then a sample with a higher dispersion of contributions would better represent the real dispersion of contributions in Lejk and Wyvill's data.

One limitation of the sample is that it does not distinguish disagreement among assessors about who contributed more from disagreement based on someone overrating or underrating everyone. Therefore, if in real courses much of the disagreement comes from some individuals generally underrating or overrating everyone, the distribution of disagreements may not be normal, as the L&W distribution assumes. Consequently, my results may underestimate the distortions produced by the methods that do not normalise assessments. With these methods, one case of overrating in a group can generate large distortions.

I included the IWF variants from the last decade (the ac-IWF, ASNIWF and it-IWF), plus the original IWF and its simplest and possibly most influential variant, the NIWF. I considered each variant in a version with self-assessments and one excluding them (indicated with the suffix -esa). In the case of the it-IWF, I also analysed the possible corrections suggested above, namely, using normalised assessments and a corrected b in the computation (it-IWF2).

To compute the IWF-esa and the it-IWF-esa, which use non-normalised peer-assessments, excluding self-assessments, I used 'deflated peer-assessments' to simulate the effect analogous to inflating self-assessments allowed by the lack of normalisation. Deflated peer-assessments are the same as the

normalised peer-assessments when self-assessments are considered in the normalisation, rounded to the nearest 5%.

Results

RMSE analysis reported in Table 2 show that the IWF-esa and the it-IWF-esa produce the smallest error in all scenarios and on average among all scenarios (average RMSE = .010), with the former performing slightly better in smaller groups. They are followed closely by the it-IWF2-esa and the NIWF-esa (average RMSEs of .012 and .013, respectively). The it-IWF2, it-IWF, IWF and NIWF follow closely. The rest of the variants have substantially larger RMSEs, ranging between .032 and .064. As we would expect, all tend to have higher RMSE in cases of high assessor disagreement. They also tend to have a larger error in groups of three and smaller in groups of seven, although the IWF-esa performs well in small groups. More importantly, the IWF, NIWF and it-IWF have large errors in cases of high dispersion of contributions (>.031), indicating that they are less stable in certain contexts. Since it is especially important to differentiate between contributions when they are large, these high errors make these estimators problematic.

Note that the it-IWF2-esa's already complex formula had to be adjusted. This was because in cases of low differences in contribution and low disagreement within a group, the iteration formula works with extremely small numbers, and by rounding it can generate a division by 0 that generates distorted it-IWF2-esa. Without this correction, the it-IWF2-esa generated an RMSE of .050 in the L&W distribution and an RMSE of .115 in the sample with low disagreement. Such cases would be easily identified by instructors, but would entail additional work. This should warn us about the potential disadvantages of complex algorithms.

Figure 1 shows the estimate error of several IWF variants by level of contribution for the main sample, that is, groups of four students and L&W distribution. It shows that most variants that include self-assessments, as well as the ac-IWF and ASNIWF, produce a broadly spread error or systematically favour those who contribute less. The IWF-esa, it-IWF-esa and it-IWF2-esa perform almost identically, favouring low contribution students about .04 more than high contribution students, with a 5th–95th percentile error range of ± 0.02 at each level of contribution. The NIWF-esa error range is slightly narrower than that of the IWF-esa, it-IWF-esa and it-IWF2-esa, but it penalises high contribution students substantially more.

While their error's range is narrow, the IWF-esa and the it-IWF-esa allow students to inflate their grade by deflating the assessments of their peers' work. Figure 2 reports the error of the estimate for the IWF-esa and it-IWF-esa, which are nearly identical. With both estimators, students tend to increase their weighting factor by slightly more than .01 if they deflate their peer assessments by .04. Note that only about 1.25% of the students would do more than that in the L&W sample. This is significantly less than their similar variants that include self-assessments. Nonetheless, in this regard the NIWF-esa and it-IWF2-esa are better estimates, as they are immune to inflation of self-assessments.

As the RMSE analysis suggests, the patterns are rather consistent across scenarios, especially for the best estimators (graphs can be replicated with the code in Supplementary Appendix 5). Thus, one of the oldest and simplest methods, the IWF-esa, has the best performance. The sophistication of the it-IWF-esa, supposedly aimed at increasing accuracy, produces no better results, indicating that the simpler IWF-esa should be preferred. However, both are slightly sensitive to biased deflation of peer assessments that could increase the assessor's grade. The it-IWF2-esa performs almost as well, but is immune to deflation of peer assessments. However, it performs nearly as well in best case scenario distributions, while its error is 56 and 60% larger in the L&W distribution and the high dispersion of contribution samples, respectively. Another old method, the NIWF-esa, performs almost as well as the IWF-esa but, while it prevents self-assessment inflation, it favours low-contribution students to the detriment of those who contribute more at a more substantial rate. The minimal loss in accuracy might be compensated by the benefit of not favouring efforts to boost one's grade by deflating peer assessments, especially for larger groups, where it performs just as well as the IWF-esa. Other it-IWF, ac-IWF and ASNIWF variants actually increase the distortions produced, which should caution against their use.

Table 2. RMSE of twelve IWF variants in nine scenarios.

Scenario	IWF	IWF-esa	NIWF	NIWF-esa	it-IWF	it-IWF-esa	it-IWF2	it-IWF2-esa	ac-IWF	ac-IWF-esa	ASNIWF	ASNIWF-esa
L&W distribution	.017	.009	.017	.011	.016	.009	.014	.014	.061	.050	.038	.030
Dispersion of contrib.												
Low	.011	.008	.011	.008	.010	.008	.015	.008	.031	.024	.020	.015
High	.033	.010	.033	.021	.031	.010	.016	.016	.124	.104	.078	.062
Disagreement												
Low	.017	.009	.017	.011	.016	.008	.013	.009	.061	.050	.038	.030
High	.022	.020	.022	.021	.021	.020	.030	.024	.061	.049	.038	.030
Self-asmnt. Inflation												
Low	.006	.007	.006	.011	.006	.007	.007	.009	.055	.049	.028	.030
High	.023	.010	.023	.011	.021	.010	.017	.009	.062	.050	.041	.030
Group size												
3	.020	.011	.020	.020	.020	.014	.023	.015	.062	.041	.040	.032
7	.011	.006	.011	.006	.011	.007	.007	.006	.058	.062	.034	.032
Mean	.018	.010	.018	.013	.017	.010	.016	.012	.064	.053	.040	.032

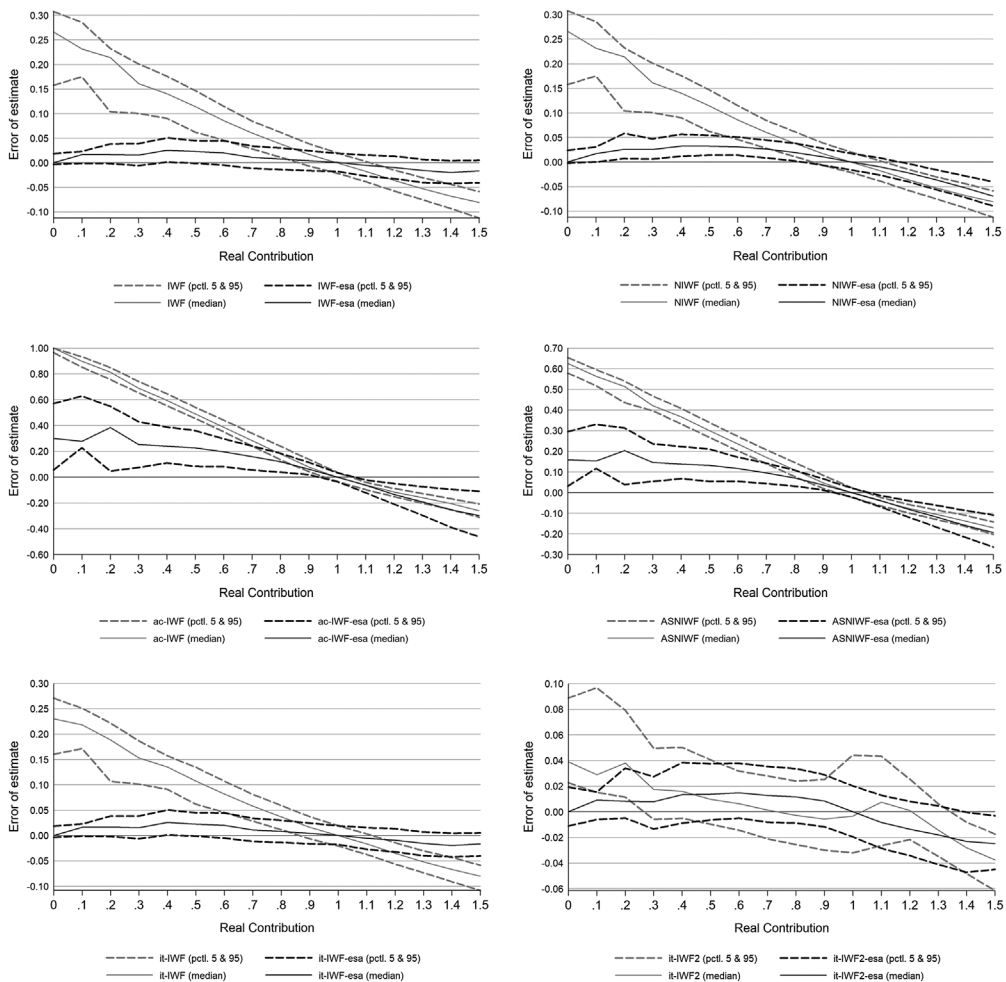


Figure 1. Estimate error's median and 5th and 95th percentiles by level of contribution (rounded to .1) in groups of four members, for twelve IWF variants.

While the errors of the IWF-esa, NIWF-esa and it-IWF-esa are fairly small, graphs in Figures 3–5 comparing them can better illuminate which one should be preferred. In groups of four that approximate the L&W distribution, a student who deflates her peers' assessments by .4 (total) to inflate her grade will likely increase her grade by about .01. Fair assessors are likely to be penalised about .01 (Figure 2). In groups of five, fair assessors are penalised .01 but other students tend to boost their grade less than .01. The effects are negligible for larger groups (Figure 3). Low contribution students' grade will likely be inflated – in the median case – but the inflation will be near 0 for near-average contributors and for complete free-riders, and about .02 for students contributing half of their share. High contribution students are likely to be penalised less than .02 (Figure 2). Both of these figures are reduced to about .01 in groups of five students, and are smaller for larger groups (Figure 4). Additionally, in 90% of the cases, the error introduced by assessor disagreement is less than ± 0.02 in groups of four (Figure 2) or slightly wider than ± 0.01 in groups of six (Figure 4).

Given that the it-IWF2-esa's advantage over the IWF-esa is that it is immune to the effect of deflating peer assessments, but this effect is so small, the difference is unlikely to be worthwhile for instructors. Additionally, much of that advantage is counteracted by the it-IWF2-esa's slightly higher overestimation

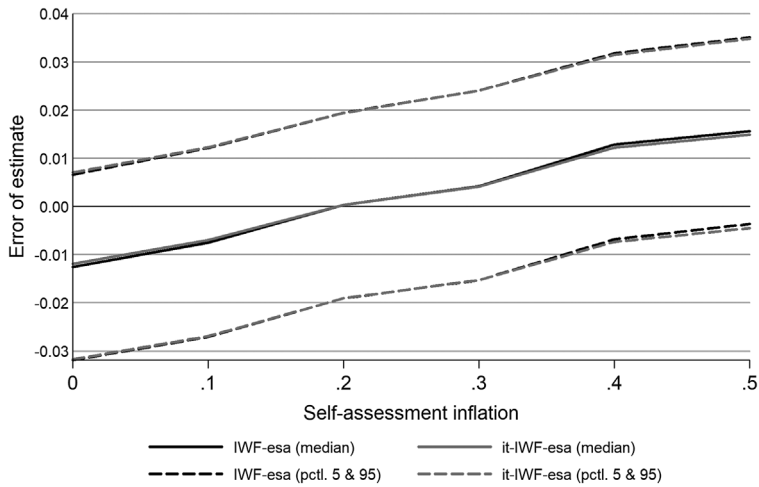


Figure 2. Estimate error's median and 5th and 95th percentiles by level of self-assessment inflation in groups of four members for IWF-esa and it-IWF-esa.

of the work of low-contribution students to the detriment of grade of high-contribution students (Figure 5, left).

The main differences between the IWF-esa and the NIWF-esa are four. First, the NIWF-esa does not allow students to boost their grades by deflating their peers' assessments. Second, the IWF-esa is highly sensitive to the distortions created by over and under-raters. Third, the NIWF-esa is likely to penalise a student who contributes 1.4 times his share in about .04 in groups of four, .3 in groups of 5, and .02 in groups of 6 (Figures 2 and 4). As a consequence of the third difference, the NIWF-esa's error is larger in cases of high dispersion of contributions, which is when grade distribution becomes more important. The NIWF-esa's RMSE in this context more than doubles the IWF-esa's RMSE, at .021 and .010 respectively. Students contributing about 50% of their share tend to be favoured more, with a median .04 extra points, whereas students contributing 1.4 of their share are penalised about .05 (Figure 5, right). Since the IWF-esa slightly underestimates the dispersion of contributions because it tends to favour low

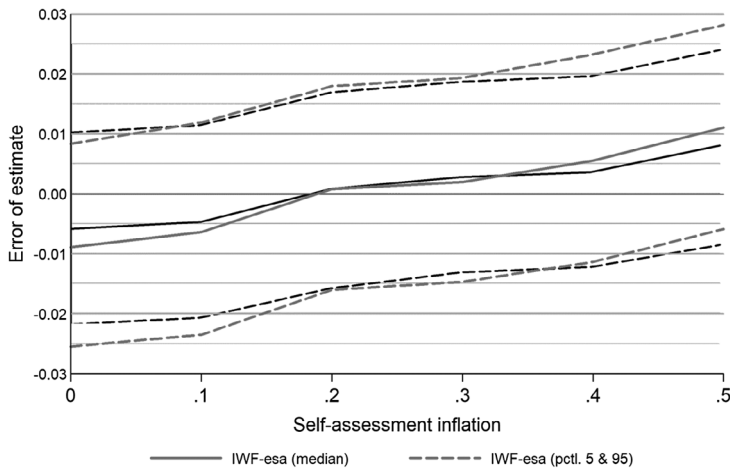


Figure 3. Estimate error's median and 5th and 95th percentiles by level of self-assessment inflation (rounded to .1) for the IWF-esa in groups of five (grey) and six (black).

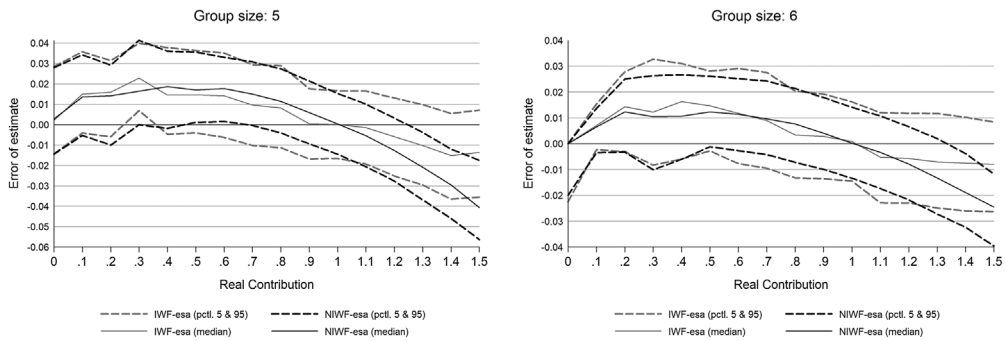


Figure 4. Estimate error's median and 5th and 95th percentiles by level of contribution (rounded to .1) for the IWF-esa and NIWF-esa in groups of five (left) and six (right).

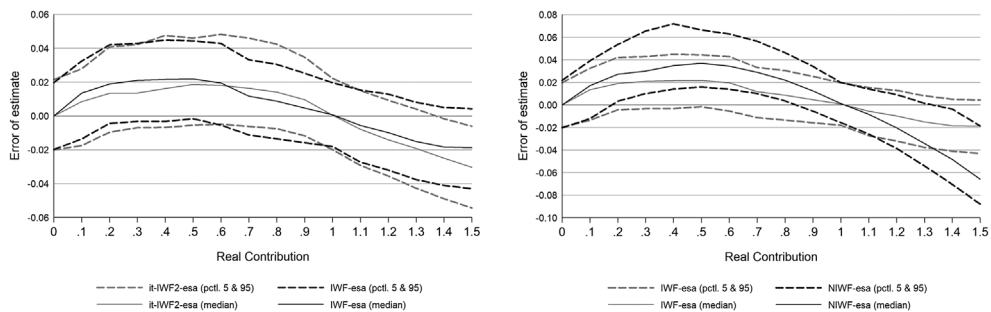


Figure 5. Estimate error's median and 5th and 95th percentiles by level of contribution (rounded to .1) for the it-IWF2-esa, IWF-esa, and NIWF-esa in sample with high dispersion of contribution and groups of four.

contribution students and penalise high contribution students, a dispersion slightly higher than that of the L&W distribution may be more realistic. This makes the behaviour of the estimates in samples with larger dispersion of contributions a critical condition. The fourth difference may seem to make the IWF-esa a better choice. However, this depends on whether disagreement is normally distributed or skewed by a few over or under-raters. If the latter is true, the IWF-esa's distortions from under and overrating may be less acceptable than the still moderate distortions that the NIWF-esa produces in cases of high dispersion of contributions.

Concluding remarks

Several scholars have proposed ways of computing individual grades in group projects to prevent free-riding and unfair grades. However, all methods have their drawbacks and the methods' accuracy has not been adequately tested. Some problems are common to all methods. There are some agreements in the literature about how to improve the validity of peer assessments, but there are also disagreements about some alternatives. A new policy identified here to address the validity of peer assessment instruments is to use scales that start from 0. Starting from 1, as is often done, unjustifiably benefits students who contribute less to the detriment of those who contribute more. Other problems, such as that of collusion, cannot be simply solved with a formal instruction or algorithm, and have to be addressed on a case-by-case basis by instructors. Yet the topic that has generated most debate is how to compute

an individual grade after obtaining the best peer assessments we can. This article evaluated the most common and recent computation methods through Monte Carlo experiments.

Results show that newer methods produce large median estimate errors or broadly spread estimate errors, whereas the older IWF-esa produces fairly small distortions and this is rather consistent across multiple scenarios. Paradoxically, most methods designed to address issues of bias usually produced the largest errors, including Ko's version of the iterative method, the it-IWF, with self-assessments. The best methods were the simplest ones, the IWF-esa and NIWF-esa, along with the about equally accurate but substantially more complex – and thus not worth the trouble – it-IWF-esa and it-IWF2-esa.

However, no method is perfect. The IWF-esa produces low distortions overall, but is sensitive to distortions produced by over-raters and under-raters. Until studies identify the real distribution of disagreements, a more conservative approach would be to prefer the NIWF-esa, at least when there seems to be disagreement due to overrating or underrating, although this is sometimes difficult to identify. My personal experience is that, in every large class, at least one student clearly overrates everyone else, which would create a large distortion for that group if I used the IWF-esa. Even when the average level of distortions in the course is smaller with the IWF-esa, if they are large within one group, instructors may prefer the more moderate distortions of the NIWF-esa.

Either way, all methods will tend to generate some distortions, even if they are usually small ones. Students and instructors need to acknowledge that this is part of the process of team work, where it is nearly impossible to exactly quantify how much each member contributed to the final result. Outside of academia individual contributions to team work are also imperfectly assessed, if they are at all, and students should be prepared for this.

To summarise, Monte Carlo simulations have shown that sometimes the complex algorithms designed to increase accuracy actually produce large distortions. The size of these distortions should serve as a cautionary tale against testing methods only by their correlations with grades, a few ideal cases, and students' opinions. Future variants of the IWF method should always compare the results they would produce against the 'real contribution' of students in simulated data and should pay attention to where the distortions concentrate. Scholars and instructors should also consider this article's results as a reminder that sometimes simpler is better, and usually no method is perfect.

Acknowledgements

I thank Robert Zarrow, Lila Elman and the anonymous reviewers for their helpful comments on previous drafts.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by Universidad Andres Bello [grant number DI-798-15/JM].

Notes on contributor

Sebastián G. Guzmán is an assistant professor of Sociology at West Chester University of Pennsylvania and research professor at Universidad Andrés Bello. His research interests include social movements, political sociology, social theory, research methods, and education.

ORCID

Sebastián G. Guzmán  <http://orcid.org/0000-0002-6463-5844>

References

- Baker, D. F. 2007. "Peer Assessment in Small Groups: A Comparison of Methods." *Journal of Management Education* 32 (2): 183–209.
- Bushell, G. 2006. "Moderation of Peer Assessment in Group Projects." *Assessment & Evaluation in Higher Education* 31 (1): 91–108.
- Carvalho, A. 2013. "Students' Perceptions of Fairness in Peer Assessment: Evidence from a Problem-based Learning Course." *Teaching in Higher Education* 18 (5): 491–505.
- Conway, R., D. Kember, A. Sivan, and M. Wu. 1993. "Peer Assessment of an Individual's Contribution to a Group Project." *Assessment & Evaluation in Higher Education* 18 (1): 45–56.
- Dommeier, C. J. 2012. "A New Strategy for Dealing with Social Loafers on the Group Project: The Segment Manager Method." *Journal of Marketing Education* 34 (2): 113–127.
- Feichtner, S. B., and E. A. Davis. 1984. "Why Some Groups Fail: A Survey of Students' Experiences with Learning Groups." *Organizational Behavior Teaching Review* 9 (4): 58–73.
- Goldfinch, J. 1994. "Further Developments in Peer Assessment of Group Projects." *Assessment & Evaluation in Higher Education* 19 (1): 29–35.
- Goldfinch, J., and R. Raeside. 1990. "Development of a Peer Assessment Technique for Obtaining Individual Marks on a Group Project." *Assessment & Evaluation in Higher Education* 15 (3): 210–231.
- Jaques, D., and G. Salmon. 2007. *Learning in Groups: A Handbook for Face-to-Face and Online Environments*. Abingdon: Routledge.
- Jin, X.-H. 2012. "A Comparative Study of Effectiveness of Peer Assessment of Individuals' Contributions to Group Projects in Undergraduate Construction Management Core Units." *Assessment & Evaluation in Higher Education* 37 (5): 577–589.
- 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., and M. Wyvill. 2001a. "The Effect of the Inclusion of Selfassessment with Peer Assessment of Contributions to a Group Project: A Quantitative Study of Secret and Agreed Assessments." *Assessment & Evaluation in Higher Education* 26 (6): 551–561.
- Lejk, M., and M. Wyvill. 2001b. "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.
- Li, L. K. Y. 2001. "Some Refinements on Peer Assessment of Group Projects." *Assessment & Evaluation in Higher Education* 26 (1): 5–18.
- Loddington, S., K. Pond, N. Wilkinson, and P. Willmot. 2009. "A Case Study of the Development of WebPA: An Online Peer-Moderated Marking Tool." *British Journal of Educational Technology* 40 (2): 329–341.
- Macfarlane, B. 2016. "The Performative Turn in the Assessment of Student Learning: A Rights Perspective." *Teaching in Higher Education* 21 (7): 839–853.
- Monson, R. 2017. "Groups That Work: Student Achievement in Group Research Projects and Effects on Individual Learning." *Teaching Sociology* 45 (3): 240–251.
- Neus, J. L. 2011. "Peer Assessment Accounting for Student Agreement." *Assessment & Evaluation in Higher Education* 36 (3): 301–314.
- Ohland, M. W., M. L. Loughry, D. J. Woehr, L. G. Bullard, R. M. Felder, C. J. Finelli, R. A. Layton, H. R. Pomeranz, and D. G. Schmucker. 2012. "The Comprehensive Assessment of Team Member Effectiveness: Development of a Behaviorally Anchored Rating Scale for Self- and Peer Evaluation." *Academy of Management Learning & Education* 11 (4): 609–630.
- Sharp, S. 2006. "Deriving Individual Student Marks from a Tutor's Assessment of Group Work." *Assessment & Evaluation in Higher Education* 31 (3): 329–343.
- 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.
- Tu, Y., and M. Lu. 2005. "Peer-and-Self Assessment to Reveal the Ranking of Each Individual's Contribution to a Group Project." *Journal of Information Systems Education* 16 (2): 197–205.
- Webb, N. M. 1995. "Group Collaboration in Assessment: Multiple Objectives, Processes, and Outcomes." *Educational Evaluation and Policy Analysis* 17 (2): 239–261.
- Zhang, B., and M. W. Ohland. 2009. "How to Assign Individualized Scores on a Group Project: An Empirical Evaluation." *Applied Measurement in Education* 22 (3): 290–308.