Peer assessment: is it fair?

Nicole Herbert, School of Computing, University of Tasmania, Australia
Nicole.Herbert@utas.edu.au

***Abstract**: A common method of peer assessment is to have team members quantify their team members' contribution. Opponents to this method of peer assessment are concerned about: the impact on minorities within teams (such as female or international students), how often individuals overly reward themselves, and how often team members indicate equal contribution. A web-based peer assessment system has been in use in a capstone project course for over two years with over 25 teams each year, offering an opportunity to investigate suggestions from the literature that this method of peer assessment has undesirable distribution patterns.*

Introduction

Team work can impact greatly on student learning and commitment, but team work presents numerous challenges for performing a fair and accurate assessment. Providing a student with an individual grade is even more difficult in a project course because the work products vary from project to project, and an individual's contribution can be hard to identify. The assessment scheme must include an approach for measuring the individual input of each team member to ensure they all contribute. Team members are often the best source of meaningful data about individual contribution.

Software Engineering Project is a 26-week capstone program, divided into two 13-week consecutive units; the students get two results. The students undertake real projects with local businesses in teams of four to five students. Table 1 summarises the team and student data from 2004 and 2005. Students form their own teams but are encouraged to get a mix of genders and nationality. The majority of international students are Chinese (71%) and all are temporarily in Australia to study and the majority are only here to complete the final 1.5 – 2 years of their degree. A student's result is not based totally on the successful implementation of a working product; the grade for each semester is determined by six main components: design, software, documentation, presentation, professionalism, and personal software process. Components are assessed using a variety of methods, by a number of different people, including self and peer assessment (Clark 2005).

Four tools are used to allow students to indicate their own and team members' individual contribution; all of these are described in Clark, Davies and Skeers (2005). One of the tools is the Work Product Pay Packet (WPPP) – each team member gives a quantitative opinion of how much each team member contributed to a major work product. This approach is similar to that described by Hayes, Lethbridge and Port (2003) and Kennedy (2005). Clark et al. (2005) gives a detailed description of the evolution of the tool and its use in the assessment process to determine a mark. In first semester 2004 each student distributed twenty marks amongst their team members, since second semester 2004 the students have distributed a virtual $100 dollars, see Figure 1. They had to distribute the amount between team members based on the quality and quantity of work produced by each individual. In 2004 the WPPPs were used seven times and in 2005 they were used 10 times.

Brown (1995), Kaufmann, Felder and Fuller (2000), Layton and Ohland (2001) and Hayes et al. (2003) reported success when they used a process of having students quantify the contribution of team members, but they also expressed some concerns. Kaufmann (2000) raised concerns about individuals inflating their own performance and team members agreeing to give equal amounts to avoid conflict. Kaufmann et al. (2000) and Layton and Ohland (2001) both raised concerns about gender and ethnicity bias. Kennedy (2005) asked whether peer assessment was worth the effort, concluding that marks awarded to individuals based on peer assessment differ only slightly from equal allocation.

Table 1. Team and student data from 2004 and 2005

Teams	*2004*	*2005*	*Students*	*2004*	*2005*
Total Teams	27	25	Total Students	129	118
Teams with gender mix	11	10	Females (all in mixed gender teams)	13	17
Teams with more than 2 of each sex	2	2	Males in mixed gender teams	41	30
Teams of all domestic students	20	14	Males not in mixed gender teams	75	71
Teams of all international students	3	5	Domestics in mixed domicile teams	11	13
Teams with domicile mix	4	6	Domestics not in mixed domicile teams	96	65
Teams more than 2 of each domicile	1	3	Internationals in mixed domicile teams	8	16
			Internationals not in mixed domicile teams	14	24

	Pay	Comments
Pat	$30	I have given myself the highest amount as I felt I undertook the more difficult parts of the implementation.
Clive	$25	
Tom	$15	
John	$10	
Jeff	$20	
Total	**$100**	

	Pat	Clive	Tom	John	Jeff	Pay
Pat	30	25	30	30	25	**140**
Clive	25	20	25	25	25	**120**
Tom	15	20	15	10	15	**75**
John	10	15	15	10	15	**65**
Jeff	20	20	15	25	20	**100**

Figure 1. Work Product Pay Packet and Work Product Pay Packet Team Summary

This paper focuses on the data contained in the WPPPs from 2004 and 2005, and investigates the following assertions: team members distribute amounts equally, team members give themselves the highest amount, team members are unwilling to give themselves the lowest amount, and team members in minority groups (female and international students) are discriminated against.

Issue 1: Team members distribute amounts equally

As shown in Table 2, in 2004 the students gave equal amounts to all their team members 453 times out of a possible 903 times (number of students 129 x 7 tests) which is 50% of the time. There was a drop of 15% between first and second semester in 2004 in the number of times students gave equal amounts to all their team members. In 2005 there was a drop of 8% between semesters. As shown in Table 3, in 2004 an entire team gave equal amounts to all their team members 42 times out of a 189 team tests (number of teams 27 x 7 tests each), 22% of the time. In 2005 it was 15% of the time – a significant reduction on 2004. There are four main reasons that could explain the drop:

1. In 2004 a change was made to the amount the students had to distribute between semesters.
2. Students receive formative results during first semester and a summative result at the end of first semester and therefore see the impact of peer assessment.
3. Students were advised in course materials and in lectures and in regular meetings with the lecturer about the consequences of carrying team members and the impact on their own marks and not to make a pact to distribute amounts equally.
4. In first semester students have to distribute work equally for each work product (eg design report), but in second semester they have to distribute work equally over all work products (allowing a student to do more on implementation and less on design).

Table 2. Individuals who gave equal amounts to all team members

	2004			*2005*		
	N	**Possible**	**% of Possible**	**N**	**Possible**	**% of Possible**
Females	34	91	37	92	170	54
Males	419	812	52	415	1010	41
Domestics	345	749	46	269	780	35
Internationals	108	154	70	238	400	60
Semester 1	228	387	59	279	590	47
Semester 2	225	516	44	229	590	39
Total	453	903	50	508	1180	43

Table 3. Entire team gave equal amounts to all team members

	2004			*2005*		
	N	**Possible**	**% of Possible**	**N**	**Possible**	**% of Possible**
All international	16	21	76	29	50	58
All domestic	26	140	19	7	140	5
Mixed domicile	0	28	0	1	60	2
Total	42	189	22	37	250	15

There is no indication that males or females are more likely to distribute amounts equally, but international students gave equal amounts considerably more often than domestic students in both years. In 2004, teams consisting of entirely international students gave equal amounts 76% of the time. In 2005, teams consisting of entirely international students gave equal amounts 58% of the time – eleven times more often than the teams consisting of entirely domestic students, but a significant reduction on 2004. The most interesting revelation of 2004 and 2005 is that mixed domicile teams (teams of both international and domestic students) never (or nearly never) gave equal amounts to everyone. In 2005 there were three international teams that gave the same amount nine, eight and seven times each, the rest of the teams all did it four or less times, five teams only did it once, with 14 teams never doing it. The most an all domestic team rewarded evenly was three times.

Issue 2: Students give themselves the highest amount

In 2004 students gave themselves the highest amount 191 times. In 134 cases they shared the equal amount with at least one other member of their team, but not the entire team. From Table 4 you can see that in both 2004 and 2005 students gave themselves the highest amount just over 20% of the time. International students gave themselves the highest amount less often than the other categories (around 15%), but this should be considered with the evidence of the earlier section that demonstrated that international students were much more likely to give equal amounts to everyone.

Are they giving themselves the highest amount too often? In 2004, of the 191 cases where a student gave themselves the highest amount other team members gave them their highest amount 142 times (74%); in 2005, 203 of the 257 cases had agreement (79%). Figure 2 shows the distribution of how often other team members agreed that the student should have had the highest amount, indicating that giving yourself the highest amount is supported in the majority of cases. Figure 2 also shows how often the people in agreement gave the student the same, more, or less money. 24% of the time they weren't willing to give them as much money as they gave themselves.

Are they giving themselves the highest amount enough? In 2005, at least two members agreed that another member should be given the highest amount 222 times. In only 60% of these cases did the individual agree, and of the remaining 22% of the individuals gave equal amounts to all. In 2005, four people agreed that another member should get the highest amount 29 times and in 21 cases the member agreed (72%), with 5 others giving equal to all. So in conclusion, it seems some individuals can be reluctant to acknowledge their own contribution.

Table 4. Individuals that gave themselves their highest/lowest amount (but ***NOT*** equal amount to all)

		Possible	*Highest to themselves*	*Others equal high*	*% of possible*	*Lowest to themselves*	*Others equal low*	*% of possible*
2004	Females	91	2	23	27	9	7	18
	Males	812	55	111	20	29	39	8
	Domestics	749	51	117	22	29	42	9
	Internationals	154	6	17	15	9	4	8
	Total	903	57	134	21	38	46	9
2005	Females	170	11	20	18	4	28	19
	Males	1010	92	134	22	65	123	19
	Domestics	780	74	127	26	61	93	20
	Internationals	400	29	27	14	8	58	17
	Total	1180	103	154	22	69	151	19

Issue 3: Students will not give themselves the lowest amount

In 2004, students gave themselves the lowest amount 84 times. In 46 cases they shared an equally low amount with at least one other member of their team, but not the entire team. From Table 4 it is evident that from 2004 to 2005 there was a significant increase in the number of students who gave themselves the lowest amount, but in 2004 many more students distributed equal amounts. In 2004 female students were more likely to give themselves the lowest amount, but not in 2005.

Are they giving themselves the lowest amount too often? Of the 84 cases in 2004 where a student gave themselves the lowest amount other team members gave them their lowest amount 66 times (78%). In 2005, 175 of the 220 cases had agreement (80%). This indicates that giving yourself the lowest amount is supported in the majority of cases, but sometimes some students do it unnecessarily.

Are they giving themselves the lowest amount enough? In 2005, two or three team members agreed 222 times that another member should be given the lowest amount, 94 of these members agreed (47%) and of the remaining 32% of the individuals gave equal amounts to all. In 2005, four people agreed that another member should get the lowest amount 51 times and in 35 cases the

member agreed (69%), with 12 others giving equal to all (24%). So in conclusion, it is not enough and an individual tries to disguise a low contribution by distributing equal amounts to all.

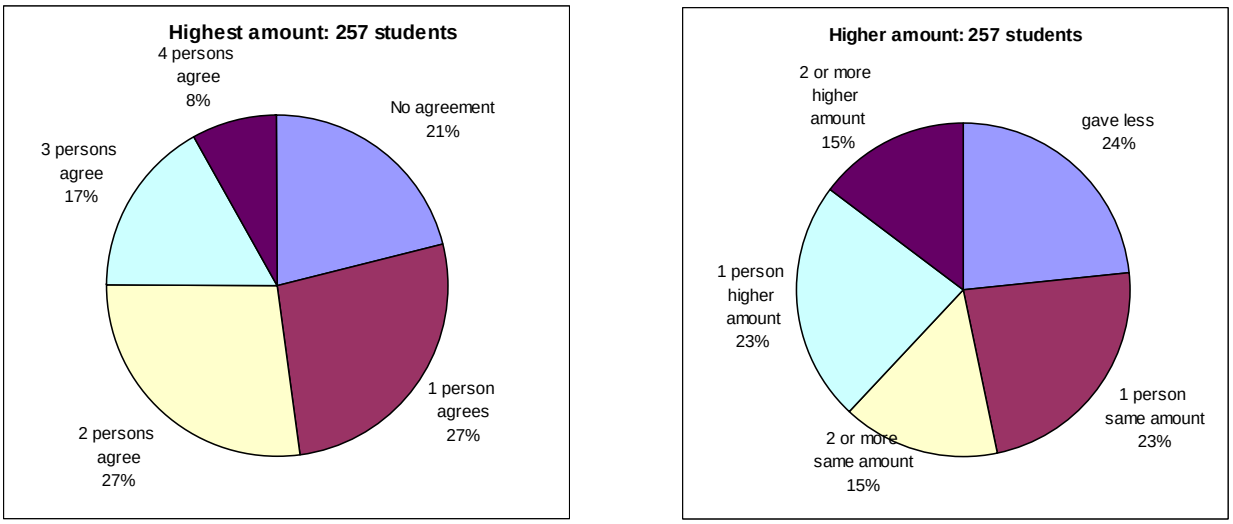

Figure 2. 2005 Highest amount agreement and actual amount agreement

Issue 4: Minorities are discriminated against

All levels of significance are determined using a Mann-Whitney nonparametric test for significance between the distributions of two independent samples, with statistical significance defined by $p<0.05$. Figure 3, in combination with Table 5, indicates that there is no evidence of gender bias, certainly not against the minority female group, and the tests show that there is no significant bias against males.

Figure 3 indicates that domestics in both mixed and non-mixed teams average around the same amount for each other. Domestic students in 2005 gave international students significantly lower amounts than they gave fellow domestic students (Table 5, row 5). Domestic students gave international students significantly lower amounts than international students gave each other (Table 5, row 7). International students gave domestic students significantly more than domestic students gave domestics students (Table 5, row 6), interestingly this was the reverse in 2004! Internationals students gave higher amounts to fellow international students than fellow domestic students in 2004, it was reversed in 2005 and the difference in 2005 was significant (Table 5, row 8). Internationals in mixed teams gave significantly lower amounts to each other than internationals in non-mixed teams (Table 5, row 9). This analysis indicates that there maybe a bias against international students.

Before concluding that there is domicile bias it is necessary to consider the final grades of students, as the peer amounts may just reflect actual performance. An analysis of the 2005 data, shown in Table 6, does show the difference in semester 2 individual final marks between domestic and international students in mixed teams is significant. But the difference between domestic students in mixed teams and domestics in non-mixed teams is not significant and the difference between international students in mixed teams and internationals in non-mixed teams is also not significant. The final marks are calculated using the WPPPs and so are not totally independent. Each team has a design report assessed in semester 2 and internationals in mixed teams did achieve significantly higher results than internationals in non-mixed teams. Similarly with the semester 2 team software marks, international students in mixed teams did average significantly more than international students in non-mixed teams. This evidence confirms that internationals in mixed teams have an advantage over internationals in non-mixed teams and that the peer assessments are just correcting that advantage – indicating there is no domicile bias.

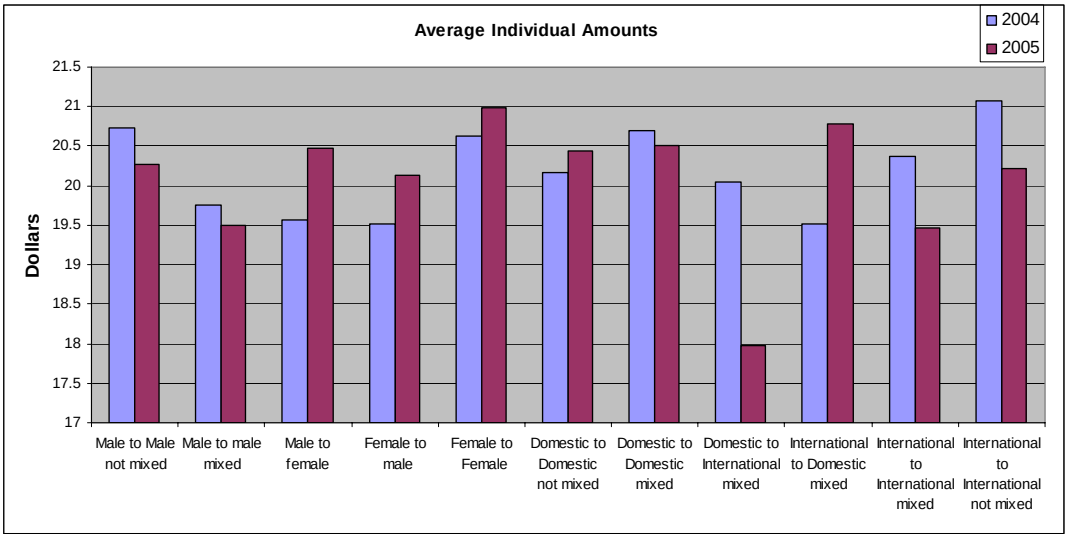

Figure 3. Average individual amounts

Table 5. Significance of differences in average individual allocations for 2005

		N	*Average amounts*	*p*
1	Male to Female Male to Male (mix)	377 932	20.48 19.5	0.1515
2	Male to Female Female to Female	377 160	20.48 20.98	0.0735
3	Female to Male Female to Female	375 160	20.13 20.98	0.4483
4	Female to Male Male to Male (mix)	375 932	20.13 19.5	0.2483
5	Domestic to Domestic (mix) Domestic to International	220 290	20.51 17.98	**<.0001**
6	International to Domestic Domestic to Domestic (mix)	290 220	20.78 20.51	**0.0143**
7	International to International (mix) Domestic to International	320 290	19.47 17.98	**0.002**
8	International to International (mix) International to Domestic	320 290	19.47 20.78	**<0.0001**
9	International to International (mix) International to International (not mixed)	320 920	19.47 20.22	**<0.0001**

Table 6. Significance of differences in 2005 semester 2 marks

		N	*Average Mark*	*p*
Domestic (mix) International (mix)	Semester 2 individual final mark	13 16	73.85 67.31	**0.0314**
	Design report team mark (max 10)	13 16	9 8.13	0.102
	Software team mark (max 20)	13 16	15.77 15.13	0.0869
International (not mixed) International (mix)	Semester 2 individual final mark	24 16	65.75 67.31	0.496
	Design report team mark	24 16	7 8.13	**0.0244**
	Software team mark	24 16	14.42 15.13	**0.0322**
Domestic (not mixed) Domestic (mix)	Semester 2 individual final mark	65 13	76.75 73.85	0.1131
	Design report team mark	65 13	9.2 9	0.4325
	Software team mark	65 13	15.42 15.77	0.1814

Conclusion and implications for practice

This paper investigated four concerns about a method of peer assessment in which team members quantify their team members' contribution. The evidence suggests that equal distribution rarely happens on a team wide basis, though individual students will distribute amounts equally, particularly international students. The paper suggested four factors that were influential in reducing equal distribution. The evidence suggests that students are likely to give themselves their highest amount both too often and not enough. Similarly, the evidence suggests that students give themselves the lowest amount both too often and not enough. This evidence suggests that the tool should not be used in isolation and that it is necessary to have other forms of evaluating the contribution of individuals to confirm the quantitative opinions of the students. Two other methods utilized are timesheets and individual contribution reports written by a team member and agreed to by other team members, these are further described in Clark et al. (2005). Finally, the evidence suggests that there is no gender or domicile bias in the peer assessment, and that peer assessments are reflecting actual (or at least perceived) performance.

References

Brown, R. (1995) Autorating: Getting Individual Marks from Team Marks and Enhancing Teamwork. *Proceedings of Frontiers in Education Conference*.

Clark, N. (2005) Evaluating student teams developing unique industry projects. *Proceedings of the seventh Australasian Conference on Computing Education*. Newcastle, Australia.

Clark, N., Davies, P. and Skeers, R. (2005) Self and Peer Assessment in Software Engineering Project. *Proceedings of the seventh Australasian Conference on Computing Education*. Newcastle, Australia.

Hayes, J.H., Lethbridge, T.C. and Port, D. (2003) Evaluating Individual Contribution Toward Group Software Engineering Projects. *Proceedings of International Conference on Software Engineering, Portland*, Oregon. 622–627

Kaufmann, D.B., Felder, R.M. and Fuller, H. (2000) Accounting for individual effort in cooperative learning teams. *Journal of Engineering Education*. **89** (2), 133–140.

Kennedy G.J. (2005) Peer-assessment in Group Projects: Is It Worth It? *Proceedings of the seventh Australasian Conference on Computing Education*. Newcastle, Australia.

Layton, R.A. and Ohland, M.W. (2001) Peer Ratings Revisited: Focus on Teamwork, Not Ability. *Proceedings of American Society for Engineering Education Annual Conference*, Albuquerque.