Gender parity in peer assessment of team software development projects

Tom Crick
Swansea University
Swansea, UK
thomas.crick@swansea.ac.uk

Tom Prickett
Northumbria University
Newcastle upon Tyne, UK
tom.prickett@northumbria.ac.uk

Jill Bradnum
Northumbria University
Newcastle upon Tyne, UK
jill.bradnum@northumbria.ac.uk

Alan Godfrey
Northumbria University
Newcastle upon Tyne, UK
alan.godfrey@northumbria.ac.uk

ABSTRACT

Development projects in which small teams of learners develop software/digital artefacts are common features of computing-related degree programmes. Within these team projects, it can be problematic ensuring students are fairly recognised and rewarded for the contribution they make to the collective team effort and outputs. Peer assessment is a commonly used approach to promote fairness and due recognition. Maintaining parity within assessment processes is also a critical aspect of fairness. This paper presents the processes employed for the operation of one such team project at a UK higher education institution, using the Team-Q rubric and analysing the impact of the (self-identified) gender of learner marking and the learner being marked on the scores obtained. The results from this institutional sample (N=121) using the Team-Q metric offers evidence of gender parity in this context. This study also makes the case for continued vigilance to ensure Team-Q and other rubrics are used in a manner that supports gender parity in computing.

CCS CONCEPTS

• **Social and professional topics → Computing education**; **Student assessment**.

KEYWORDS

Peer assessment, group projects, team working, gender, diversity

ACM Reference Format:
Tom Crick, Tom Prickett, Jill Bradnum, and Alan Godfrey. 2022. Gender parity in peer assessment of team software development projects. In *Computing Education Practice 2022 (CEP 2022), January 6, 2022, Durham, United Kingdom.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3498343.3498346

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CEP 2022, January 6, 2022, Durham, United Kingdom
© 2022 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-9561-8/22/01…$15.00
https://doi.org/10.1145/3498343.3498346

1 WHAT IS IT?

In the UK and other jurisdictions, computing degrees regularly require teams of learners develop software/digital artefacts. This commonly involves the demonstration of various software engineering competencies; for example, analysis and design; implementation; testing; configuration management and version control; team working; project management/control, etc [9]. Such projects are not universally well-received or well-regarded by students [12]; however there is an employability-related dimension [16, 26] to the development of these competencies, as well as required characteristics for professional body degree accreditation [5].

Within these team projects, all members of the teams are expected to contribute to the development of these software/digital artefacts. Contribution may be structured in some manner by task, by role, etc; in other projects, the teams are more self-managing. Commonly, there will be some expected collective outcomes such as a product (e.g. common look and feel, an integrated product, etc) or task-based (e.g. risk analysis, project plan, team demonstration of the product, etc). In addition to the collective tasks there may be individual tasks which again could be product (e.g. building of sub-system X) or task-based (e.g. testing of the product). Assuring fair contributions from all learners to collective tasks can present challenges [18]. One common approach to address this challenge is peer assessment [12]. It has been reported that peer assessment: is appreciated by learner as a mechanism to hold members accountable and aid fairer marking; supports giving and receiving feedback; prompts personal reflection and improvement; supports supervision; informs project planning and management; facilitates exploring and reshaping group dynamics; improves project outputs; and offers a sense of safety to raise issues without repercussions [16]. Various peer assessment schemes which are based upon peer ranking, holistic view or peer rating, and category based [15, 27] have been proposed. One common form of peer rating/category system is for learners to assess their peers by a given metric, calculating means for individual and teams, and then weighting collective marks accordingly. This can be achieved by online tools such as WebPA [20], BuddyCheck.io [22], or SparkPlus [23] or via the use of online surveying tools. In general terms, the algorithm commonly adopted is then: (*i*) each learner scores each of their peers in their team by a set of metrics; (*ii*) a weighting is then calculated by Mean Peer Score for the learner divided by Mean Peer Score for the whole

team; and (*iii*) the individual learner would then be awarded the mark that team is awarded multiplied by this weighting. Such team projects commonly have individual assessed outcomes as well as team responsible assessed outcomes and it would normally only be the team outcomes that would be subject to the weighting.

Clearly the criteria used has a significant influence on the weighting and hence potentially the grade awarded to the individual learners. Given that all of the learners contribute to the marking, there is more potential for unconscious or conscious bias to influence the marking than if the marking was completed solely by faculty. This paper, building on recent work [7], explores whether a validated approach – Team Q [2] – together with a situated set of practices, supports gender parity in terms of the peer assessment marks awarded. Team Q [2] assesses five components of team working: contributes to team project; facilitates contributions of others; planning and management; fosters a team climate; and manages potential conflict. Each of the components is in turn measures by indication of '*how often does your peer demonstrate the following*' against a set of descriptions. Each description is awarded scores as follows: Never=0; Sometimes=1; Usually=2; Regularly=3; and Always=4. The full set of descriptions can be seen in Table 1. Hence a learner scores each of their peers out of 56 overall. The peer weighting is then calculated using the algorithm indicated previously (with a learner's mean peer score divided by their team's mean peer score).

2 WHY ARE YOU DOING IT?

It is widely recognised that STEM [1] and more specifically computing remain male-dominated disciplines. In the UK, only 1 in 5 Computing, Engineering and Technology students were female in the 2019-2020 academic year [14]. For the computing discipline in the UK, 26,285 out of 105,485 (just less than 20%) identified as female and 210 identified as non-binary [14]. Addressing this imbalance is critical for the disciplines involved, from a social, cultural and economic perspective, to maximise the potential future development of the discipline. This is further reinforced by the ambitions of the United Nations Sustainable Development Goals: *SDG4: Quality Education* and *SDG5: Gender Equality*.

Belonging [29] is recognised as a crucial factor for retaining learners within the computing discipline, yet learners who self-identify as women have been reported to have a lower sense of belonging [17]. The challenges faced by female students related to belonging has also been explored by qualitative studies [31]. Together, this highlights the need to carefully evaluate whether assessment practices promote belonging, support diversity, and gender parity. Whilst there is always the need to assure equity in assessment, in this case when learners are contributing to the assessment processes of their peers the need to assure the processes used exhibits gender parity is particularly strong.

Peer assessment is the main focus of this work; however, it is acknowledged the data is being gathered as part of assessed tasks which may, in some way, influence the outcomes. Additionally, more details of the situated practices adopted are provided as a context.

3 WHERE DOES IT FIT?

The team project is run as part of the final year of an undergraduate Computer Science degree at a UK higher education institution. The study took place in the 2020-2021 academic year between January and April over a 15-week period including a 3-week spring vacation, under the constraints of the COVID-19 pandemic. Institutional ethical approval for the study was obtained. Explicit written consent was obtained as part of the peer assessment learners were asked to indicate "What gender do you identify as?" as an optional free text field; additionally, they were specifically requested to approve their consent to be included in the study. Hence learners have the opportunity to not supply the information and additionally specifically consent to participate. Those who provided a null response have not been included in the study. The size of the cohort was 170. Of this group, 121 learners are included in the study with 108 learners self-identifying as male ('male' or 'man' or 'masculine') and 13 self-identifying as female. A further three learners identified responded as non-binary (responding: 'I don't know'; 'non-binary'; and 'nothing'); these have not been included in the study due to concerns that they may be individually identifiable. The key aspects of the management of the projects are as follows:

Team allocation: The learners are allowed to either self-select their teams or choose to be assigned a team. Teams are normally comprised of five individuals. If learners wish to be assigned a team, these are allocated randomly upon a first-come, first-served basis. In this case, three teams were allocated randomly, these were exclusively male teams.

Project selection: All the projects are 'live' development projects in the sense that teams develop a software/digital artefact for a third-party. Some of the projects are self-sourced by the learners, some are tutor sourced.

Learning Agreements: As part of the establishment of teams, the learners are required to produce a learning agreement which documents key decisions regarding how the team will collaborate to complete the work. Teams are encouraged to reflect upon this agreement as the projects progress. A writing frame is provided posing key questions the learning agreement should address.

Development approach: The teams are required to follow a full-stack development approach with each team member developing a subsystem that they can ultimately demonstrate individually if required. However, the teams are encouraged to demonstrate a fully-integrated working product and are recognised for doing so as part of the marking rubric. If presenting an integrated working product is not possible for reasons beyond an individual learner's control (for example, there is a passenger in the team) adjustments are made so that a learner is not unfairly penalised.

Support: The teams are supported by weekly progress review meetings with a tutor. These follow a stand-up style with each team member asked to identify progress and any road blocks which can then be discussed in more depth. A Microsoft Word and a InVision Freehand template were provided to support this activity. These records were uploaded to the institutional virtual learning environment at the end of meetings. External to the meetings, the supervising tutor attempted to support the teams to resolve any team-related issues; one team had to be reorganised in this delivery.

Assessment: There are three related components of summative assessment: a project proposal (10%), a digital product and related demonstration (50%) and an individual report which critically evaluates the project and the professional, ethical, legal and social issues a finalised and deployed version of the produced prototype would

Table 1: Means of Team Q Score by gender of marked learner and by gender of marker pairing (female marking female, female marking male, male marking female and male marking male)

		Marked Gender		Marker Gender / Marked Gender			
Component	Description	Female	Male	Female-Female	Female-Male	Male-Female	Male-Male
	Mean Team-Q Score	46.94	47.15	50.60	46.41	46.30	47.22
	Number of marks awarded in each category	32	323	5	27	27	296
Contribute to team project	*Participates actively and accepts a fair share of the group work*	3.47	3.49	3.80	3.44	3.41	3.49
	Works skilfully on assigned tasks and completes them on time	3.44	3.37	3.40	3.19	3.44	3.38
	Gives timely, constructive feedback to team members, in the appropriate format	3.19	3.24	3.20	3.26	3.19	3.23
Facilitates contributions of others	*Communicates actively and constructively*	3.28	3.38	3.60	3.37	3.22	3.38
	Encourages all perspectives be considered and acknowledges contributions to others	3.41	3.44	3.80	3.37	3.33	3.44
	Constructively builds on the contributions of others and integrates own work with work of others	3.38	3.36	3.80	3.33	3.30	3.37
Planning and Management	*Takes on an appropriate role in the group (e.g. leader, note take, etc)*	3.31	3.05	4.00	2.93	3.19	3.06
	Clarifies goals and plans the project	3.22	3.20	3.60	3.00	3.15	3.22
	Reports to team on progress	3.38	3.35	3.60	3.30	3.33	3.34
Fosters a team climate	*Ensures consistency between words, tone, facial expressions, and body language*	3.47	3.47	3.60	3.63	3.44	3.46
	Expresses positivity and optimism about team members and project	3.44	3.49	3.40	3.63	3.33	3.48
Manages potential conflict	*Displays appropriate assertiveness: neither dominating, submissive nor passive aggressive*	3.34	3.42	3.60	3.30	3.30	3.43
	Contributes to appropriately healthy debate	3.28	3.45	3.60	3.30	3.22	3.46
	Responds to and manages direct/indirect conflict constructively and effectively	3.44	3.46	3.60	3.37	3.41	3.46

need to mitigate. The team aspects are: 50% of the proposal and 20% of the digital product and related demonstration which are marked as a team and are weighted by peer assessment.

Peer Assessment: The project proposal and the demonstration contain team and individual tasks which are peer assessed. Over various historical deliveries of the course various technologies have been used to administer the peer assessment including paper, virtual learning environment tools and other electronic tools. In this delivery, peer assessment was administered by Microsoft Forms. Two rounds of peer assessment were completed, one formative and one summative. Only the summative round is included in the study.

4 DOES IT WORK?

The responses to the Team-Q rubric were analysed using a combination of Excel (data storage and cleaning) and R (statistical analysis). The Team-Q Score, number of marks awarded, and means for responses to the different descriptions in the Team-Q rubric by the gender of the marked learner and by the gender of the marker pairing are shown in Table 1 above. There is little statistical difference in the mean of Team-Q Score for the marks awarded to female (46.94) and male (47.15) leaners (t-test t=-0.087708, df=35.438, p=0.9306). Analysis of variance (ANOVA) suggests little statistical difference in the mean Team-Q Score awarded between "female-marking male" (46.41), "male-marking female" (46.30) or "male-marking male" (47.22) pairs (markers gender F=0.104, p=0.748 and marked gender F= 0.177, p=0.674). The slightly higher female-to-female marking pairing mean (50.60) is not statistically significantly different to the other pairings (t-test t=0.697, p=0.487). This is a sample size of N=121 learners on one course delivered with a low incidence of female learners (13) but is supportive of gender parity in terms of the gender of the learner being marked and gender of the learner completing the marking.

5 WHO ELSE HAS DONE THIS?

Peer assessment and related web-based peer assessment has been advocated as a mechanism for equitable assessment of contribution to team and team software development projects for a number

of years [2, 4, 10, 12, 15, 18, 21]. When peer assessment is used in a summative context, there can be bias due to affiliation with a group [3], and learners sometimes do not want to award a low mark to their peers (and particularly to their friends) [24]. Bias in peer assessment on the basis of gender has been widely reported [13, 25], and elsewhere bias has not been evidenced [11, 28]. The performance benefits associated with diverse teams are also reported. This mixed picture highlights the need to validate tools employed in different contexts to assure the process exhibits gender parity.

6 WHAT WILL YOU DO NEXT?

The study will be repeated in more typical learning conditions as the COVID-19 pandemic [6, 8, 30] may have influenced the results. The sample size is small, including only a small number of female learners hence growing the sample size may increase confidence in the results. Learners may identify as minorities for many reasons (e.g. gender (including non-binary), ethnicity, neurodiversity, no family history with higher education, etc) and considering all such factors is an area for future work. Statistical parity, is one thing, learner perceiving assessment to be fair is equally critical. Learner perceptions present an urgent set of research work to supplement quantitative with qualitative data via interviews with learners to explore their view of inclusion and whether there are perceptions of micro-aggressions or bias in team activities.

7 WHY ARE YOU TELLING US THIS?

The Team Q peer model produces a team work weighting via a marking scheme synthesised from wider research, and thereby it presents a comprehensive model for what constitutes effective team working. Its usage highlights to learners the broad set of competencies that constitute good quality team working which are not solely technical competencies [5, 16, 19, 26]. The authors contend this provision of a benchmark for good team working practice provides useful formative feedback for the learners as they complete the projects in their teams. Limitations discussed in the previous section not withstanding, it is encouraging that there was statistical evidence of gender parity within the peer assessment scheme in this context. Team-Q is a well-established rubric which explores more dimensions of team working than some of the more standard approaches that are embedded in existing tools [20, 22, 23] by default. Finally, given the low overhead of evaluating the statistical impact of self-identified gender upon peer assessment results, doing so is a practice recommendation.

REFERENCES

[1] Chardie L. Baird. 2018. Male-dominated STEM disciplines: How do we make them more attractive to women? *IEEE Instrumentation Measurement Magazine* 21, 3 (2018), 4–14. https://doi.org/10.1109/MIM.2018.8360911

[2] Emily Britton, Natalie Simper, Andrew Leger, and Jenn Stephenson. 2017. Assessing teamwork in undergraduate education: a measurement tool to evaluate individual teamwork skills. *Assessment & Evaluation in HE* 42, 3 (2017), 378–397. https://doi.org/10.1080/02602938.2015.1116497

[3] Christina M. Cestone, Ruth E. Levine, and Derek R. Lane. 2008. Peer assessment and evaluation in team-based learning. *New Directions for Teaching and Learning* 2008, 116 (2008), 69–78. https://doi.org/10.1002/tl.334

[4] Nicole Clark, Pamela Davies, and Rebecca Skeers. 2005. Self and Peer Assessment in Software Engineering Projects. In *Proc. of ACE'05*. 91–100.

[5] Tom Crick, James H. Davenport, Paul Hanna, Alastair Irons, and Tom Prickett. 2020. Computer Science Degree Accreditation in the UK: A Post-Shadbolt Review Update. In *Proc. of CEP'20*. Article 6. https://doi.org/10.1145/3372356.3372362

[6] Tom Crick, Cathryn Knight, Richard Watermeyer, and Janet Goodall. 2020. The Impact of COVID-19 and "Emergency Remote Teaching" on the UK Computer Science Education Community. In *Proc. of UKICER'20*. ACM, 31–37. https://doi.org/10.1145/3416465.3416472

[7] Tom Crick, Tom Prickett, and Jill Bradnum. 2022. A preliminary study of peer assessment feedback within team software development projects. In *Proc. of SIGCSE'22*. ACM.

[8] Tom Crick, Tom Prickett, and Julie Walters. 2021. A Preliminary Study Exploring the Impact of Learner Resilience under Enforced Online Delivery during the COVID-19 Pandemic. In *Proc. of ITiCSE'21*. https://doi.org/10.1145/3456565.3460050

[9] James H. Davenport, Alan Hayes, Rachid Hourizi, and Tom Crick. 2016. Innovative Pedagogical Practices in the Craft of Computing. In *Proc. of LaTiCE'16*. 115–119. https://doi.org/10.1109/LaTiCE.2016.38

[10] Fabian Fagerholm and Arto Vihavainen. 2013. Peer assessment in experiential learning Assessing tacit and explicit skills in agile software engineering capstone projects. In *Proc. of FIE'13*. 1723–1729. https://doi.org/10.1109/FIE.2013.6685132

[11] Nancy Falchikov and Douglas Magin. 1997. Detecting Gender Bias in Peer Marking of Students' Group Process Work. *Assessment & Evaluation in HE* 22, 4 (1997), 385–396. https://doi.org/10.1080/0260293970220403

[12] Neil Andrew Gordon. 2010. Group working and peer assessment — using WebPA to encourage student engagement and participation. *ITALICS* 9, 1 (2010), 20–31. https://doi.org/10.11120/ital.2010.09010020

[13] Laura Heels and Marie Devlin. 2019. Investigating the Role Choice of Female Students in a Software Engineering Team Project. In *Proc. of CEP'19*. ACM, Article 2. https://doi.org/10.1145/3294016.3294028

[14] HESA. 2021. What do HE students study?: Personal characteristics. https://www.hesa.ac.uk/data-and-analysis/students/what-study/characteristics

[15] Mark Lejk and Michael Wyvill. 2001. Peer Assessment of Contributions to a Group Project: A comparison of holistic and category-based approaches. *Assessment & Evaluation in HE* 26, 1 (2001), 61–72. https://doi.org/10.1080/02602930020022291

[16] Alexander Mitchell, Terry Greer, Warwick New, Joseph Walton-Rivers, Matt Watkins, Douglas Brown, and Michael James Scott. 2021. Student Perspectives on the Purpose of Peer Evaluation During Group Game Development Projects. In *Proc. of UKICER'21*. ACM, Article 7. https://doi.org/10.1145/3481282.3481294

[17] Catherine Mooney and Brett A. Becker. 2020. Sense of Belonging: The Intersectionality of Self-Identified Minority Status and Gender in Undergraduate Computer Science Students. In *Proc. of UKICER'20*. ACM, 24–30. https://doi.org/10.1145/3416465.3416476

[18] Helen Phillips, Wendy Ivins, Tom Prickett, Julie Walters, and Rebecca Strachan. 2021. Using Contributing Student Pedagogy to Enhance Support for Teamworking in Computer Science Projects. In *Proc. of CEP'21*. ACM, 29–32. https://doi.org/10.1145/3437914.3437976

[19] Tom Prickett, Morgan Harvey, Julie Walters, Longzhi Yang, and Tom Crick. 2020. Resilience and Effective Learning in First-Year Undergraduate Computer Science. In *Proc. of ITiCSE'20*. ACM. https://doi.org/10.1145/3341525.3387372

[20] WebPA Project. 2005. WebPA. https://webpa.lboro.ac.uk/login.php

[21] Richard Raban and Andrew Litchfield. 2007. Supporting peer assessment of individual contributions in groupwork. *Australasian Journal of Educational Technology* 23, 1 (Mar. 2007), 34–47. https://doi.org/10.14742/ajet.1272

[22] Shareworks. 2021. Buddy Check. https://www.buddycheck.io/

[23] SparkPlus. 2021. Introduction to SparkPlus. https://sparkplus.com.au/

[24] Baharini Sridharan, Joanna Tai, and David Boud. 2019. Does the use of summative peer assessment in collaborative group work inhibit good judgement? *Higher Education* 77 (2019), 853–870. https://doi.org/10.1007/s10734-018-0305-7

[25] Jacklin Stonewall, Michael Dorneich, Cassandra Dorius, and Jane Rongerude. 2018. A Review of Bias in Peer Assessment. In *Proc. of CoNECD 2018*. 1–9.

[26] Sarah. Thomas and Susan Busby. 2003. Do industry collaborative projects enhance students' learning. *Education + Training* 45, 4 (2003), 226–235. https://doi.org/10.1108/00400910310478157

[27] Yanbin Tu and Minn 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 (2005), 197–206.

[28] Richard Tucker. 2014. Sex does not matter: gender bias and gender differences in peer assessments of contributions to group work. *Assessment & Evaluation in HE* 39, 3 (2014), 293–309. https://doi.org/10.1080/02602938.2013.830282

[29] Nanette Veilleux, Rebecca Bates, Cheryl Allendoerfer, Diane Jones, Joyous Crawford, and Tamara Floyd Smith. 2013. The Relationship between Belonging and Ability in Computer Science. In *Proc. of SIGCSE'13*. ACM, 65–70. https://doi.org/10.1145/2445196.2445220

[30] Richard Watermeyer, Tom Crick, Cathryn Knight, and Janet Goodall. 2021. COVID-19 and digital disruption in UK universities: afflictions and affordances of emergency online migration. *Higher Education* 81 (2021), 623–641. https://doi.org/10.1007/s10734-020-00561-y

[31] Emily Winter, Lisa Thomas, and Lynne Blair. 2021. 'It's a Bit Weird, but It's OK'? How Female Computer Science Students Navigate Being a Minority. In *Proc. of ITiCSE'21*. ACM, 436–442.