Assessment & Evaluation in Higher Education, 2015
Vol. 40, No. 1, 151–164, http://dx.doi.org/10.1080/02602938.2014.893418

Routledge
Taylor & Francis Group

A comparison of peer and tutor feedback

John Hamer[a]*, Helen Purchase[a], Andrew Luxton-Reilly[b] and Paul Denny[b]

[a]*School of Computing Science, University of Glasgow, Glasgow, UK;* [b]*Department of Computer Science, University of Auckland, Auckland, New Zealand*

We report on a study comparing peer feedback with feedback written by tutors on a large, undergraduate software engineering programming class. Feedback generated by peers is generally held to be of lower quality to feedback from experienced tutors, and this study sought to explore the extent and nature of this difference. We looked at how seriously peers undertook the reviewing task, differences in the level of detail in feedback comments and differences with respect to tone (whether comments were positive, negative or neutral, offered advice or addressed the author personally). Peer feedback was also compared by academic standing, and by gender. We found that, while tutors wrote longer comments than peers and gave more specific feedback, in other important respects (such as offering advice) the differences were not significant.

Keywords: peer review; computing; feedback

Introduction

Asking students to provide written feedback on coursework produced by one or more of their peers can have a positive effect on learning. Peer review is particularly attractive in large classes, as a means of providing substantial quantities of feedback quickly.

Nicol (2010) argues that peer feedback needs to be understood in terms of a dialogue, rather than as a direct substitute for instructor feedback. When students participate in a peer review exercise, they take on several distinct roles. The first role is as the author of a piece of work. They then become assessors, reading work produced by one or more peers, forming an opinion on the work and generating feedback. Next, they become receivers of feedback, making choices as to which advice to follow and which to discard. A feedback dialogue with an instructor has a qualitatively different form. It is predominantly one way (a monologue), and does not require students to critically engage to the same degree.

Seeing a variety of good and poor examples can help students become more detached and critical about their own work. Aside from possible concerns about plagiarism, there are generally few objections from instructors or students about this exposure. Forming an opinion and generating feedback are significantly more cognitively demanding, and we should therefore expect this phase of the peer review process to provide the most important opportunity for learning. This is also the phase that most concerns instructors and students. The final phase, receiving feedback, is often perceived as the most important. This is, after all, the substantial (indeed, only)

*Corresponding author. Email: john.hamer@glasgow.ac.uk

phase involved when feedback comes from the instructor. Replacing an expert instructor with novices will reduce the quality of this feedback.

We distinguish here between *quantitative* and *qualitative* feedback: *quantitative* feedback is provided in terms of numerical marks, *qualitative* feedback is provided by way of textual comments. Our focus is on feedback given in first-year computing science programming assignments. We have already demonstrated that, for a first-year programming class, students' quantitative marking does not differ significantly from that of tutors (Hamer et al. 2009); in this paper, we focus on the quality of the textual comments provided in the qualitative feedback, comparing the comments provided by peers with those of tutors. Such analyses have been performed in other domains (typically the social sciences and arts), and are typically concerned with getting feedback on draft essays. Our work differs by considering feedback on computer programmes, and in looking particularly at a cohort of students who, unlike social science and arts students, are more used to working with numbers and programme code than with text.

Related work

There is a substantial body of work on feedback and its relationship to performance and learning. There is little doubt that feedback is a powerful influence (Hattie and Timperley 2007), but complex interactions arise between the type and source of feedback and its effect. Our intuitions about feedback do not always serve us well.

The effect of feedback quality on performance

Strijbos, Narciss, and Dünnebier (2010) investigated the impact of peer feedback contents and reviewer competence on academic writing performance. They categorised the feedback along the dimension of concise-general vs. elaborated-specific, and reviewer competence was self-declared as high or low. Participants in the study were graduate teacher training and psychology students, and the task was revising a piece of academic text that contained various errors. They found that groups with feedback from a low-competent peer outperformed groups with a high-competent peer during a post-test. Students receiving concise general feedback outperformed groups receiving elaborated specific feedback.

Gielen et al. (2010) investigated the impact of peer feedback characteristics on essay quality in a group of Dutch secondary school students. Peer reviewers were prompted to write a positive and a negative comment with an explanation and suggestions for improvement for each paragraph in a draft essay submitted for comment. The authors categorised the feedback according to appropriateness (related to the assessment criteria or not), specificity (explanation of the judgement), justification (reasoning behind the judgement), presence of suggestions for improvement, presence of both positive and negative comments, presence of thought-provoking questions and clarity. The authors found that the presence of justification significantly improved the quality of the final essay, but only for those whose essays were originally poor. Other quality criteria, such as appropriateness, specificity, clear formulation and presence of suggestions did not have a significant impact on essay quality.

Authoritative feedback can have negative as well as positive effects. Yang, Badger, and Yu (2006) compared the responses of students in two writing classes,

one of which received teacher feedback and one peer feedback. Peer feedback was considered to be less trustworthy, so students receiving peer feedback were more active in checking the feedback that they received against grammar books. Although students used teacher feedback more in their revisions than peer feedback, students receiving peer feedback were more actively involved in self-correction. Students receiving peer feedback made more changes to the meaning of their writing, while students receiving teacher feedback made more surface changes.

Comparing experts and peers

Cho, Schunn, and Charney (2006) asked students in psychology and education to review writing submitted by their peers using SwoRD, a web-based peer review system. Reviews were conducted by undergraduate students, graduate students and an expert reviewer. The feedback comments were categorised as directive (suggestion to make a specific change), non-directive (suggestion to make a non-specific change, such as improve spelling), praise, criticism, summary or off-task. They found that the length of the expert comments was greatest, followed by graduate students and undergraduate students in decreasing order. The following significant differences were identified in the analysis of the comment categories:

- the expert produced approximately three times as many directive suggestions as the other groups;
- the expert and graduate students provided more non-directive comments than the undergraduate students;
- undergraduates produced more praise comments than the expert (almost 70% more praise) and graduate students. Graduate students produced the most criticism;
- the expert produced the fewest summary statements.

Cho et al. (2008) showed (in the context of revising draft essays for a social science course) that multiple peer reviews can result in a greater improvement in the quality of an essay than a single expert reviewer or peer reviewer. They suggest that, when the target audience consists of non-experts, non-expert reviewers are just as helpful (or more helpful) than expert reviewers. They also found (Cho and MacArthur 2010) that single and multiple peer feedback contained more non-directive feedback and more praise than expert reviews, concluding that students who received feedback from multiple peers improved their essays more than those receiving expert feedback.

Davies (2006) analysed the comments that third-year students provided when peer reviewing essays about computing. He found that the weaker students in the class (lower two quartiles) tended to provide less critical feedback, while the stronger students (upper two quartiles) were more critical. Students were especially critical of referencing and explanations, and tended to provide holistic positive comments rather than specific positive comments.

Context

The Engineering Computation and Software Development course (ENGGEN 131) is a compulsory course for all first-year engineering students at the University of

Auckland. The 12-week course consists of a MATLAB programming module taught over the first six weeks, and a C programming module taught over the final six weeks. Each module includes a medium-sized programming project contributing 10% toward the final grade. The project submissions are graded by employed tutors (typically engineering graduate students), who complete a rubric prepared by the course instructor. For the C module of the course, once the project submission deadline has passed, each student is allocated four randomly selected projects to review. Students participating in this peer review process complete an identical rubric to that of the tutors, and are awarded 2% towards their final grade for completing all four reviews. Students were given a preparatory exercise using the peer review tool, aimed at familiarising them with the process.

A total of 613 students sat the final examination and 599 made a submission for the project in the C module of the course. We will refer to these 599 students as project authors: 538 of the project authors then went on to complete their allocation of peer reviews. The project itself required students to implement an exhaustive search algorithm, and was graded out of a total of 25 marks. The rubric was divided into three sections: style (6 marks), correctness (17 marks) and efficiency (2 marks).

Submitted code that was well commented, consistently indented and made appropriate use of functions would be awarded the marks in the style section of the rubric. The correctness marks were awarded if the programme compiled without warning and produced the correct output for several input data-sets. The efficiency marks were awarded if the project author had implemented several basic pruning strategies allowing the programme to compute solutions to several larger data-sets in a reasonable amount of time.

For each of the three sections of the rubric, an open-ended comment area was available for providing feedback to the project author. In this paper, our analysis focuses on the extent and type of the feedback written in these comment areas by both tutors and student peer reviewers.

Research questions

We do not consider here the effect of feedback quality on performance, but focus rather on the comparison between tutors' and peers' feedback. Our context is large first-year programming courses, in contrast to prior work which focused on essays. We anticipate our results to differ from the prior studies because of the differing nature of the assessment.

Our research questions are:

- Q1: How seriously did the peers take their responsibilities? We consider this with respect to the number of comment boxes completed (Q1a) and the length of the comments (Q1b).
- Q2: Was there any difference between the peers' and tutors' reviewing? We consider this with respect to the general/specific dimension (Q2a), with respect to the positive/negative/neutral/advice/personal/off-topic dimension (Q2b) and with respect to the marks given (Q2c).
- Q3: Were there any differences with respect to personal characteristics, specifically academic ability (Q3a) and gender (Q3b).

Methodology

We began by making a random selection of 10% of the project authors. Each student in this sample of 59 was marked by one tutor and peer reviewed by up to four students. We collated and classified all of the comments generated by both the tutor marking and peer reviewing processes for all project authors in our sample.

Comments were classified by four coders. The coders initially discussed the coding scheme and classified 10 sample comments together. The comments were then divided in two, and each coder independently classified half the comments. After this initial classification, the coders paired up and compared their results, coming to a consensus decision on any differences. Comments were assigned uniform anonymous identifiers, so the coders were not aware of whether a comment was written by a tutor or a student.

The three primary categories for classification were 'positive', 'negative' and 'advice/action'. A comment was classified as positive if it highlighted something that was done well, and negative if it highlighted something that was done poorly. Advice/action comments gave suggestions for making modifications to the programme. Moreover, these three primary categories were further divided into specific and general. Specific comments targeted particular elements of the code, whereas general comments were either vague or more high-level. We defined two additional categories: 'personal voice' and 'off-topic'. Comments were classified as personal voice if they were written in the second person (i.e. 'you') or included other personal features such as emoticons.

Off-topic comments were unrelated to the project. This gave us a total of eight categories, and each comment could be classified with any number of these.

Examples

Examples of feedback belonging to each category, taken from student comments, follow:

S+: Comments in this category provided positive feedback about a specific element of the code.

- *Good use of functions. There were 4 functions used in total in the programme which makes it easier to read.*
- *The commenting is very descriptive which is good.*

S−: Comments in this category provided specific negative feedback about the functionality, style or correctness of the programme.

- *The title is not printed to screen.*
- *A large section of code in the main function is indented too far.*
- *Indentation: Block of code starting from line 111 has one level of indentation too many.*

S0: Comments in this category were specific, but were not obviously positive or negative in tone.

- *Otherwise comments are there.*
- *Values may or may not be correctly allocated, this is irrelevant to the marking.*

SA: Comments in this category provided specific advice to a student about how to improve their code.

- *Your code should have terminated when the solutions were printed out. This did not happen because you wrote* $y = +1$*, which does not increment y. Instead, you should use* $y = y + 1$ *or* $y{+}{+}$.
- *When you are reading in the data, you need to keep a counter of the #characters, and when that matches the required data-set, the following data can be read in.*
- *To prevent the majority of your code existing within an else statement (for the null file pointer test), you could place a 'return 1'; or exit in the else statement and continue your code outside of the conditional.*

G+: Comments in this category are general comments that are positive. The comments do not relate to a specific element of style or requirement specified in the assignment.

- *Good style.*
- *Codes works well. Well done!*
- *Nicely done.*

G−: Comments in this category are general negative comments. They do not refer to any specific elements of code, but are instead comments directed at the overall quality (summary comments).

- *almost all marks lost here due to incompleteness*
- *this code appears incomplete*
- *does not seem to work very well*

G0: Comments in this category are general comments that do not have either positive or negative connotations. Few comments occurred in this category.

- *I dont know this!!*

GA: Comments in this category provided general advice to peers, but did not refer to specifics within the code.

- *be more careful*

PV: Comments in this category were personal in tone in that they recognised that the comments, although being about a submission, were directed to another person. Many of these were combined with one of the other categories, linked with a general or specific criticism.

- *Honestly, programming is not that hard - I'm sure if you tried, you could do it*
- *I don't get what u are trying to do here.*

OT: Comments in this category were off-topic.

- *off-topic: any comments on my marking or comments my mail is xxxxxx@hotmail.com*

We categorised each comment separately, and in some cases a single comment attracted more than one of the eight categories. As an example, the following comment in the 'correctness' section of the rubric:

> *Good work. However set 4 was not quite correct, missed out the 11 box when printing results.*

was classified as 'general-positive' and 'specific negative'.

Data analysis

Our analysis investigates differences between the feedback provided by peer reviewers and that provided by tutors, both in terms of the quantity and in terms of the categories described previously. We also consider whether certain demographic information, such as gender and ability, has an impact on the comments written by peer reviewers.

Although the rubric asked students to comment on three aspects of the submission (correctness, style and efficiency), the efficiency part of the submission was optional. This meant that several of the efficiency comment boxes were either empty or simply said 'not done'. We have, therefore, removed all the efficiency comments from our analysis. Thus, for each review, we analysed two comment boxes: one for correctness and one for style.

We chose a random 10% of the authors (59). Each author's submission was marked by one tutor, providing a total of 118 tutor-comment boxes; of these, 43 were empty, leaving 75 tutor-comments for analysis. There were 25 tutors. Each author's submission was allocated to four peer reviewers: 24 peer reviews (representing 11%) were not done at all, leaving 208 completed peer reviews, and a total of 416 peer-comment boxes. Of these, 113 comment boxes were empty, providing a total of 283 peer-comments for analysis. As we were looking at the quality of the comment text, it does not make sense to consider empty comments. There were 182 peer reviewers involved in reviewing these 59 authors.

Question 1a

How seriously did the peers take their responsibilities with respect to the number of comment boxes completed? There is no significant difference between the number of comments boxes completed by peers and tutors (Table 1). However, it is interesting to note that some tutors did not write any comments at all, resulting in 14 tutor reviews (of the 59) having no comments. These tutor reviews have been excluded from all of the analyses below. We felt that in these cases the tutor had

Table 1. Comparison of number of comment boxes completed by peer and tutors, showing no statistical significance.

	Peers	Tutors	Mann–Whitney	*p*
Total number of comment boxes completed	283, out of 416 (68%)	75, out of 118 (64%)	$U = 23{,}447$, $n = 534$	0.376
Mean number of comment boxes completed per review [0,2]	1.36 $n = 208$	1.27 $n = 59$	$U = 5696.5$, $n = 267$	0.343
Total number of correctness comment boxes completed	143, out of 208 (69%)	40, out of 59 (68%)	$U = 6077.5$, $n = 267$	0.890
Total number of style comment boxes completed	140, out of 208 (67%)	35, out of 59 (59%)	$U = 5646$, $n = 267$	0.255

underperformed. As our aim is to compare peer reviewing performance against *competent* tutor reviewing, it is appropriate to remove these poor reviews.

We did not similarly remove peer reviews with no comments as, unlike with tutors, instructors are unable to choose which peers should provide reviews. This leaves 253 reviews: 208 peer reviews and 45 tutor reviews for further analysis.

Question 1b

How seriously did the peers take their responsibilities with respect to the length of the comments? The total comment length per review was longer for tutors than peers, as was the length of the comments provided per comment box. This was the case overall, as well as for both correctness and style comments (see Table 2).

Question 2a

Was there any difference between the peers' and tutors' comments with respect to the general/specific dimension? Tutors tended to give more specific feedback than peers overall. This is also the case for correctness comments, but not for style comments, where there is no significant difference (see Table 3).

Table 2. Comparison of length of comments completed by peer and tutors. Tutors wrote significantly longer comments.

	Peers	Tutors	Mann–Whitney	*p*
Total comment length per review (tokens)	25.25 $n = 208$	51.82 $n = 45$	$U = 2526.5, < 0.00$	<0.001
Mean comment length over all comment boxes	12.62 $n = 416$	25.91 $n = 90$	$U = 12882.5$	<0.001
Mean correctness comment length over all comment boxes	12.12 $n = 208$	33.11 $n = 45$	$U = 2688$	<0.001
Mean style comment length over all comment boxes	13.13 $n = 208$	18.71 $n = 45$	$U = 3732$	0.031

Table 3. Comparison of length of comments completed by peer and tutors, broken down by general and specific phrases.

Mean number of …	Peers	Tutors	Mann–Whitney	*p*
General phrases per review	0.66 $n = 208$	0.58 $n = 45$	$U = 4420$	0.531
Specific phrases per review	1.52 $n = 208$	1.89 $n = 45$	$U = 3803$	0.043
General phrases per comment box	0.33 $n = 416$	0.29 $n = 90$	$U = 17{,}982$	0.468
Specific phrases per comment box	0.76 $n = 416$	0.94 $n = 90$	$U = 15912.5$	0.016
General phrases per correctness comment box	0.38 $n = 208$	0.29 $n = 45$	$U = 4264$	0.262
Specific phrases per correctness comment box	0.69 $n = 208$	0.96 $n = 45$	$U = 3777$	0.028
General phrases per style comment box	0.28 $n = 208$	0.29 $n = 45$	$U = 4633$	0.892
Specific phrases per style comment box	0.84 $n = 208$	0.93 $n = 45$	$U = 4182$	0.232

Question 2b

Was there any difference between the peers' and tutors' comments with respect to the positive/negative/neutral/advice/personal/off-topic dimension? Tutors make significantly more negative comments than peers. Peers tend to give more positive comments than tutors, but this difference only approaches significance. Tutors give more advice than peers, but again, this difference only approaches significance (see Table 4).

From the authors' point of view, was there a difference in the total amount of feedback they got from tutors when compared to their peers? This is to be expected, given that each author's submission was reviewed by four peers, and only one tutor, so there is no question that each author would receive more feedback from peers than tutors. It is, however, a crucial question, as we wish to demonstrate the large amount of feedback that can be obtained by peer reviewing. In this analysis, we include those tutor reviews that have no comments (see Table 5).

Table 4. Comparison of length of comments completed by peer and tutors, broken down by positive/negative/neutral/advice/personal/off-topic.

Mean number of …	Peers	Tutors	Mann–Whitney	*p*
Positive phrases per review	1.24 $n = 208$	0.19 $n = 45$	$U = 3948$	0.085
Negative phrases per review	0.74 $n = 208$	1.22 $n = 45$	$U = 3015$	<0.001
Neutral phrases per review	0.01 $n = 208$	0.02 $n = 45$	$U = 4643$	1.000
Advice phrases per review	0.19 $n = 208$	0.31 $n = 45$	$U = 4130$	0.085
Personal voice phrases per review	0.34 $n = 208$	0.42 $n = 45$	$U = 4191$	0.154
Off-topic phrases per review	0.00 $n = 208$	0.00 $n = 45$	$U = 680$	1.000

Table 5. Comparison of total amount of feedback from peers and tutors.

	Peers	Tutors	Mann–Whitney	*p*
Mean number of comment tokens received per author	89.0 $n = 59$	39.5 $n = 59$	$U = 3.872$	<0.001

Question 2c

How did the marks given by peers differ from those given by tutors? As well as giving space for reviewers to provide comments, the marking rubric also asked for marks to be allocated under different categories. The maximum number of marks that could be awarded to any submission was 25.

We have already addressed the question of the difference between tutor-given and peer-given marks (Hamer et al. 2009), where we found a correlation of $R = 0.713$. So as to verify those results, we repeat the analysis here, by determining whether there is a correlation between the tutor-given and peer-given marks for the 59 submissions. The peer-given mark used in this analysis is the mean of all the marks given by the peer reviewers.

We include those tutor reviews with no comments, as, in this analysis, we are only interested in the marks given. There is no significant difference between the marks given by peers and those given by tutors, with, on average, peers giving less than one mark less than tutors (see Table 6). The correlation between the peer-given and tutor-given marks is $r = 0.748$, $p \ll 0.001$, which is similar to previous studies (Hamer et al. 2009). The regression formula is peer-given mark = 5.4649 + 0.7408 × tutor-given mark (see Figure 1).

Table 6. Comparison of mean marks given by peers and tutors.

	Peers	Tutors	Paired sample *t*-test	*p*
Mean mark	18.20, $\sigma = 5.45$, max = 23, min = 1.33, $n = 59$	18.95, $\sigma = 5.40$, max = 23, min = 0, $n = 59$	$t = 1.491$	0.167

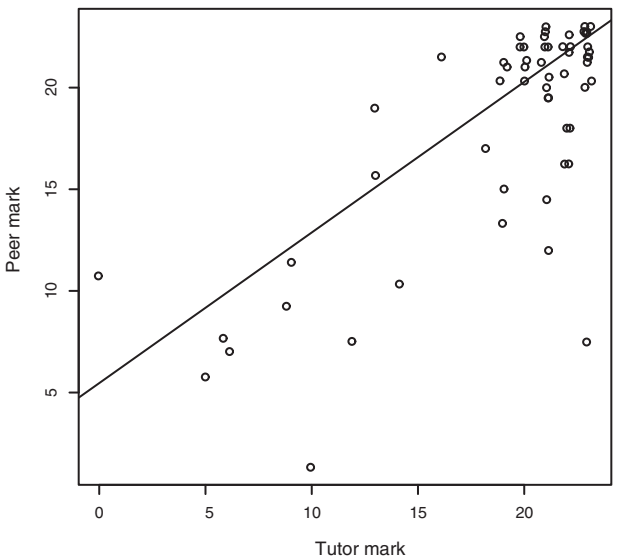

Figure 1. Scatter plot of tutor and peer marks.

Question 3a

Does the comment data differ according to the ability of the student peer reviewer? Based on examination marks after the peer review, we divided the 182 peer reviewers into four ability quartiles:

(Q1) 0–57 (24.2%, $n = 44$)
(Q2) 58–69 (24.7%, $n = 45$)
(Q3) 70–81 (29.1%, $n = 53$)
(Q4) 82–90 (22.0%, $n = 40$)

The mean examination mark over all 182 peer reviewers was 68.29 (min = 15, max = 90).

We analysed the comments given in all 208 peer reviews, with respect to these quartiles, and considering several different aspects of the comments. The higher performing students tended to give more general, and more negative comments than the lower performing students (see Table 7).

Question 3b

Does the comment data differ according to the gender of the student peer reviewer? Of the 182 peer reviewers, 161 were men and 47 were women. The only gender difference found were that women tended to give more general comments than men, and used the personal voice more often (see Table 8).

Analysis summary

Forming an opinion is understood to be cognitively demanding, and our results generally support this. Tutors are better at being specific when commenting on programme correctness and (along with higher performing students) are more prepared to make negative comments. Tutors also write more in each review, although the total quantity of feedback is greater from peers due to the writing of multiple reviews.

In other regards, however, differences between tutor and peer commenting were not significant. Tutors did not provide more advice, and nor did they give more specific comments when commenting on programme style. There were no statistically

Table 7. ANOVA comparison of mean number of peer comment types by peer quartile.

	Mean total comment length	Mean action comments	Mean specific comments	Mean general comment	Mean positive comments	Mean negative comments	Mean personal voice comments
Q1 ($n = 50$)	23.68	0.14	1.16	0.58	1.06	0.52	0.32
Q2 ($n = 55$)	19.18	0.13	1.56	0.42	1.15	0.71	0.29
Q3 ($n = 58$)	25.12	0.24	1.48	0.72	1.21	0.74	0.31
Q4 ($n = 45$)	34.55	0.27	1.93	0.98	1.60	1.02	0.47
ANOVA	$p = 0.253$	$p = 0.296$	$p = 0.079$	*0.003*[a]	$p = 0.121$	*0.030*[b]	$p = 0.523$

[a]Q4 is significantly different from Q2 ($p = 0.02$).
[b]Q4 is significantly different from Q1 ($p < 0.016$).

Table 8. Comparison of mean number of peer-comment types by gender.

	Mean total comment length	Mean action comments	Mean specific comments	Mean general comment	Mean positive comments	Mean negative comments	Mean personal voice comment
M (*n* = 161)	23.65	0.17	1.48	0.59	1.15	0.73	0.27
F (*n* = 47)	30.70	0.25	1.66	0.92	1.53	0.78	0.57
Mann–Whitney *U*	3302	3472	3532	2993	3210	3616	3070
p	0.181	0.186	0.476	*0.016*	0.099	0.618	*0.009*

significant differences in the number of general, positive, neutral, personal voice and off-topic comments between the two groups.

Some minor stylistic differences were observed between female and male students, with women using the personal voice more often and writing more general comments.

Conclusion

Our study has explored the differences and similarities in the feedback given by tutors and student peers under similar conditions. We have not asked whether individual pieces of feedback are 'right' or 'wrong', but looked instead at the nature of the feedback: its specificity and positivity, and whether it is actionable, personal or off-topic.

We expected, and found, that tutors identify more points to comment on than peers, and are able to make more specific comments on technical matters such as correctness. The increased frequency of negative comments in reviews by tutors and by high performing students reflects a confidence in the course material. This pattern conforms to our own anecdotal observation that negativity increases with the initial acquisition of expertise, but then subsequently reduces as teaching experience tempers expectations.

In other respects, the characteristics of feedback we measured were surprisingly similar. Where judgement is being made on matters that are not strictly 'right' or 'wrong', the feedback between tutors and peers showed no significant differences. The reasons for this are not clear. It suggests that even after two or three years of study, tutors are not appreciably more confident about the aesthetics of programme code than novices. Perhaps this is not surprising, as few opportunities are provided elsewhere in the curriculum for students to engage in discussions on programming style.

We, and others, have argued that effective peer review does not depend on the feedback produced by peers being of the same standard as tutors, as its primary value arises from the process of writing a review. Any help elicited by the feedback received can therefore be considered a bonus. The quality of the feedback produced is a reflection of how successfully students accomplished the peer review task. For this particular course, we see evidence of a good match between the challenges of the review task and the abilities of the cohort. The feedback provided by tutors on technical issues may be more specific than that given by peers, but not on more subjective matters.

Of interest, but not investigated directly in this paper, is the question of how students engaged with the feedback they received. In the introduction, we described peer feedback as a dialogue, in contrast to the monologue character of instructor-generated feedback. Yang et al.'s (2006) observation that students respond to teacher feedback in a more passive manner are most pertinent, as this points to a potentially deep and significant benefit of peer review. Further research into student engagement with peer feedback would be of value. We note the design of the peer review activity used in the course we studied does not provide an opportunity for any extended dialogue between students, and we suggest this would be a productive direction to enhance the use of peer reviewing.

On a closing note, we observed in an earlier paper (Hamer et al. 2009) that a policy of employing tutors by considering the quality of their peer reviewing in previous years, rather than simply on their academic grades, appeared effective. The course here did not adopt this policy, and this may have resulted in a somewhat mediocre team of tutors. Improvements may only arise through regular peer review practice across the curriculum, with a focus on developing reviewing skill.

Notes on contributors

John Hamer is a freelance software developer, part-time lecturer at the Universities of Glasgow and Uppsala, and honorary lecturer at the University of Auckland where he held a senior lectureship in Computer Science until 2012. His research interests include contributing student pedagogies and collaborative learning.

Helen Purchase is senior lecturer in Computing Science at the University of Glasgow. She has research interests in information visualisation (in particular in empirical studies of graph drawing comprehension) and in collaborative and contributing educational pedagogies. She is the author of "*Experimental Human Computer Interaction - A Practical Guide with Visual Examples*" (Cambridge University Press, 2012).

Andrew Luxton-Reilly is a senior lecturer in Computer Science at The University of Auckland, where he has been involved in Computer Science Education research since 2004. His research interests include contributing student pedagogies, software tools that support collaborative learning, and the learning and teaching of novice programmers. He has previously published 43 articles in journals and internationally peer-reviewed conference proceedings.

Paul Denny is a senior tutor in Computer Science at the University of Auckland. His interests include developing and using technologies for supporting collaborative learning, particularly involving student-authored resources. His recent publications have examined the effectiveness of virtual achievements for motivating students in online environments, and proposing a taxonomy for classifying the variation between student solutions to programming problems. He is the author of PeerWise, a tool that supports student-authored assessment questions used by over 700 institutions worldwide.

References

Cho, K., and C. MacArthur. 2010. "Student Revision with Peer and Expert Reviewing." *Learning and Instruction* 20 (4): 328–338. doi:10.1016/j.learninstruc.2009.08.006.

Cho, K., C. D. Schunn, and D. Charney. 2006. "Commenting on Writing." *Written Communication* 23 (3): 260–294. doi:10.1177/0741088306289261.

Cho, K., T. R. Chung, W. R. King, and C. Schunn. 2008. "Peer-based Computer Supported Knowledge Refinement." *Communications of the ACM* 51 (3): 83–88. doi:10.1145/1325555.1325571.

Davies, P. 2006. "Peer Assessment: Judging the Quality of Students' Work by Comments Rather than Marks." *Innovations in Education and Teaching International* 43 (1): 69–82. doi:10.1080/14703290500467566.

Gielen, S., E. Peeters, F. Dochy, P. Onghena, and K. Struyven. 2010. "Improving the Effectiveness of Peer Feedback for Learning." *Learning and Instruction* 20 (4): 304–315. doi:10.1016/j.learninstruc.2009.08.007.

Hamer, J., H. C. Purchase, P. Denny, and A. Luxton-Reilly. 2009. "Quality of Peer Assessment in CS1." In *5th International Workshop on Computing Education Research (ICER)*, 27–36.

Hattie, J., and H. Timperley. 2007. "The Power of Feedback." *Review of Educational Research* 77 (1): 81–112.

Nicol, D. 2010. "From Monologue to Dialogue: Improving Written Feedback Processes in Mass Higher Education." *Assessment & Evaluation in Higher Education* 35 (5): 501–517.

Strijbos, J., S. Narciss, and K. Dünnebier. 2010. "Peer Feedback Content and Sender's Competence Level in Academic Writing Revision Tasks: Are they Critical for Feedback Perceptions and Efficiency?" *Learning and Instruction* 20 (4): 291–303. doi:10.1016/j.learninstruc.2009.08.008.

Yang, M., R. Badger, and Z. Yu. 2006. "A Comparative Study of Peer and Teacher Feedback in a Chinese EFL Writing Class." *Journal of Second Language Writing* 15 (3): 179–200. doi:10.1016/j.jslw.2006.09.004.