Peer assessment in large undergraduate classes: an evaluation of a procedure for marking laboratory reports and a review of related practices

Judy R. Harris

School of Physiology and Pharmacology, University of Bristol, Bristol, United Kingdom

Submitted 8 November 2010; accepted in final form 5 January 2011

Harris JR. Peer assessment in large undergraduate classes: an evaluation of a procedure for marking laboratory reports and a review of related practices. Adv Physiol Educ 35: 178–187, 2011; doi:10.1152/advan.00115.2010.—This study provides evidence that peer marking can be a reliable tool for assessing laboratory reports in large cohorts. It was conducted over a 4-yr period with first-year undergraduates (~180 students/cohorts) taking a mammalian physiology course, but the procedure adopted would be applicable to any other laboratory-based discipline. The process was found to be efficient in staff time, enabling a summative practical report to be marked in <1 h (<5% of the time that had previously been required for staff marking), facilitating rapid feedback to students on their performance. When samples of the peer-assessed reports were marked by a single member of staff, there was excellent correlation between peer and staff marks (r = 0.96–0.98), although peer-awarded marks exceeded staff marks by an average of 2.5–3.0%. The validity of peer marking was independent of both the sex of the marker and the staff score awarded to the marker for the same piece of work. Feedback from students was largely positive; they reported that the procedure adopted was effective in increasing their understanding of the underlying physiology and contributed to their understanding of best practice in presenting a laboratory report. Seventy percent of students agreed that it was acceptable for peer assessment to contribute a small (up to 5%) component of the overall mark for the course. The results are discussed in relation to other reports of peer marking, particularly when used to assess an academic product or process in a scientific discipline.

Peer assessment has been defined (33) as “an arrangement for peers to consider the level, value, worth, quality, or success-fulness of the products or outcomes of learning of others of similar status.” It can be a valuable educational process, providing students with insight into the criteria against which their work is assessed and clarifying the requirements for producing work of a particular standard (5, 6, 28, 33). It also provides students with a degree of ownership of the assessment process and encourages them to reflect on the quality of their work (1, 32) through providing opportunities for good practice to be shared and reinforced as well as deterring poor practices that may be more apparent to a marker than the original writer. The introduction of peer assessment early in the undergraduate curriculum should therefore help first-year students to develop lifelong learning skills from the outset of their degree program (see Refs. 23) as well as encourage the development of generic skills such as critical appraisal and an ability to provide colleagues with objective feedback on their work. The latter skill is valuable in a wide range of careers, particularly for professions in, and allied to, healthcare.

Peer assessment also has the potential to save staff considerable marking time. This is of particular benefit in relation to large student cohorts, which have become more common as the increased numbers of students entering higher education have rarely been associated with a corresponding increase in staff numbers. In these situations, peer marking can offset the inevitable reduction in the ability of staff to give students prompt and individual formative feedback on their work. Such feedback is generally acknowledged to make an important contribution to the learning process (13, 29), and student views on its quality and timeliness are routinely sought in national surveys such as the United Kingdom National Student Satisfaction Survey (19) and the United States National Survey of Student Engagement (20). The former survey repeatedly demonstrates that most undergraduates would appreciate more (and more prompt) feedback on their work. In an experimental discipline such as physiology, an important factor in student learning is the receipt of timely feedback on laboratory work including the quality of data presentation, analysis, and interpretation.

Despite the potential benefits of peer assessment, it is not practiced widely in undergraduate curricula. The initial reaction of many staff and students is that it is inappropriate for students to assess each others’ work, particularly in their first undergraduate year (23), and that peer assessment is not sufficiently reliable to be used in a summative context. There have been several reports (1, 5, 28, 30, 34) showing that students lack confidence in their ability to assess their peers’ work reliably, with many of them regarding assessment as the exclusive preserve of staff. Students express particular concern about potential bias in marking when the peer marker knows the identity of the writer and/or vice versa (5, 6). There have also been reports of sex bias in peer assessment (16) and that students of differing levels of academic ability vary in their reliability as peer assessors (15, 34). Some procedures for peer assessment have also been reported as being very time consum- ing, involving much preparation by staff and requiring extensive training for students (5, 28).

The present study describes and evaluates a procedure for summative peer assessment of laboratory reports in a first year BSc physiology course. The objective was to determine the reliability and effectiveness of peer assessment in a large enrollment class by the following:

1. Comparing the peer and staff marks awarded to a sample of physiology laboratory reports.

2. Evaluating the impact of the marker’s sex on the validity of the peer mark that they awarded.

Address for reprint requests and other correspondence: J. R. Harris, School of Physiology and Pharmacology, Univ. of Bristol, Medical Sciences Bldg., University Walk, Bristol BS8 1TD, UK (e-mail: judy.harris@bris.ac.uk).
3. Evaluating the impact of the peer marker’s performance in the assignment (as determined by the faculty mark awarded to their work) on the validity of the mark that they awarded.

4. Exploring students’ views of peer assessment, both in general terms and the value of the particular procedure adopted.

5. Comparing the efficiency of the process in staff time and the speed with which feedback was provided to students with the previous practice of staff marking.

Preliminary results from this study have been previously reported (12).

METHODS

The study was conducted over a 4-yr period with four cohorts of undergraduates taking a first-year BSc physiology course, registration numbers for which ranged from 170 to 190 students/year. It focused on a renal physiology practical class undertaken at the end of the first semester, in which students worked in pairs to investigate the effect of posture and exercise on water diuresis. Each pair of students was assigned to one of three “posture groups” and recorded urine and sodium output for 2 h after drinking. Class data were then collated and made available to all students. After the class, students completed an individual, precirculated pro forma report structured in the format of a scientific paper. This report, together with the scoring rubric, is provided in the Supplemental Material.1

Completion of the report, other than the Results section, required students to fill in “blanks” (45 in total) with appropriate words, phrases, or occasionally two to three sentences of free text. These parts of the report carried approximately half of the available marks. In the Results section, which carried the remaining marks, students used their own experimental data and the collated class results to construct graphs and perform numerical exercises based on the data. They were also required to select and carry out appropriate statistical tests on their own data and the collated class data.

The design of the pro forma report was modified after the first year of the study when staff marking of a sample of reports showed that almost half of the sampled peer markers (19/40) had incorrectly added up the marks they had awarded for each subsection of the report, with errors ranging from −9% to +6% about the arithmetically correct figure. The report was therefore amended to include a number of boxes (11 in all) in the margin, which indicated the total mark achievable for each subsection, graph, calculation, etc., together with a space for the peer marker to insert the mark achieved. This greatly improved the reliability with which peer markers added up the total marks assigned; in the second year of the study, incorrect additions occurred in only 4 of 28 double-marked scripts, with errors being restricted to a much narrower range (−4.5% to +2% about the correct figure).

The Marking Session

Peer marking of the reports was carried out in a compulsory review session held 2 wk after the laboratory class. Completed reports were handed in 24 h before this. Students were warned that failure to hand in a report on time, without a bona fide explanation, would debar them from the marking/review session and from being awarded credit for their work. The submitted reports were then distributed randomly but not anonymously. Students were asked to exchange any report that had inadvertently been given to the original writer; this was a rare occurrence given the size of the cohort. A member of staff (the author) then guided the peer marking process by completing and explaining an exemplar pro forma report, which was broadcast to the class. A predetermined marking scheme was used, and the rationale for correct/acceptable answers was explained as the session progressed.

For the text-based sections of the report, peer markers were given guidance on the acceptable words/phrases/sentences for completing the blank parts of each section. Since these mostly required students to insert only one or two words, markers had limited scope for making judgments about the amount of credit they should award, although more judgment was required in assigning credit for some parts of the Discussion that consisted of two to three explanatory sentences.

In the Results section, students needed to exercise more judgment in awarding credit, although some guidance was provided, for example, that credit should be given for the correct labeling of axes and inclusion of correct keys in graphs, accurate numerical analysis, and the appropriate use of statistics. Markers were also encouraged to include annotations/corrections on the script to enrich the feedback provided to the writer.

At the end of the marking session, which took 40–45 min, students were asked to sign the report that they had marked before returning it to the writer. The writer was then instructed to review their marked script carefully in relation to the individual marks that had been allocated and the accuracy with which they had been totaled. They were then to check whether or not the marking was fair, or to specify that they wished to appeal against the mark that had been awarded, for which they were required to provide justification.

Each report was marked by only one peer assessor. All marked reports were collected at the end of the session, and the staff member (the author) marked a sample of them. Most of the faculty-marked scripts were selected by drawing scripts at regular intervals from an alphabetical list of the entire student cohort. For example, in the first year of the study, the scripts submitted by every fifth student in this list were selected, independent of the score assigned to them. Some additional scripts were then added to the sample to ensure that it included scripts that represented the entire range of peer-awarded marks. Any reports for which an appeal had been lodged (<5 over the period of the study) were also checked by the member of staff.

Analysis of Peer and Staff Marking Data

Regression analysis was performed on peer-awarded marks (adjusted to be arithmetically correct if required; see above) versus the corresponding staff-awarded marks to determine the correlation between them. Also, the sample mean and SD for peer marks were compared with those of the corresponding staff-awarded marks using a paired t-test to compare the absolute values of the marks awarded as opposed to their rank order. Any sex differences in the quality of students’ marking judgment were assessed by carrying out separate regression analyses between staff-awarded marks and the corresponding marks awarded by either male or female markers.

Sex differences in the accuracy of marking were also assessed using an unpaired t-test to compare the average “peer minus staff” mark for the samples of male versus female peer markers. Sex bias in marking (i.e., any bias in the award of marks to peers of the same sex vs. the opposite sex) was investigated in two ways. First, the data for male and female markers were further subdivided on the basis of the sex of the writer; separate regression analyses were then carried out between staff marks and the corresponding peer marks awarded by each gender of marker to writers of the same sex versus the opposite sex. Second, the average “peer minus staff” mark was calculated for

1 Supplemental Material for this article is available online at the Advances in Physiological Education website.
the four situations of “male marker-male writer,” “male marker-female writer,” “female marker-male writer,” and “female marker-female writer.” Any sex bias within the marks awarded by male and female markers was then determined using an unpaired t-test.

An additional regression analysis was carried out to determine whether peer markers who had achieved a high (staff-awarded) mark for the assignment were more reliable markers than students who had performed less well in the assignment.

Student Feedback

At the end of the marking sessions in the first 2 yr of the study, students were asked to complete a written questionnaire that solicited specific feedback on the effectiveness of the renal peer marking session in increasing their understanding of renal physiology and of how to present a laboratory report. In the third and fourth years of the study, these targeted questions were replaced by questions soliciting more generic feedback on the use of peer assessment, such as the importance of anonymity, whether it was fair and appropriate for peer marking to contribute a summative component to the overall course mark, the frequency with which peer marking should be used, and the extent to which evaluating someone else’s work was helpful to the marker’s own learning.

RESULTS

Validity of Students’ Judgment in Assessing the Work of Their Peers

Figures 1 and 2 show the relationship between peer marks and the corresponding staff marks for student cohorts in the first and second year of the study, respectively. The sample sizes were 40 scripts from a cohort of 172 students (23%) and 28 scripts from a cohort of 185 students (15%), respectively. In both years, there was a very high correlation between the peer and staff marks awarded ($r = 0.97$ and 0.98, respectively). In the first year, the average peer mark for the sampled scripts exceeded the average staff mark by 2.5% [average peer mark: 77.7% (SD 12.5) vs. average staff mark: 75.2% (SD 12.6)], whereas in the second year, the corresponding values were peer “overmarking” by an average of 2.9% [average peer mark: 79.2% (SD 14.4) vs. average staff mark: 76.3% (SD 13.0)]. The difference between the peer marks and staff marks awarded was statistically significant ($P < 0.001$) in both years. Most of the overmarking by peers related to the Results section of the report, in which markers had to exercise more critical judgment in assigning the marks. A t-test showed that the slopes of the regression lines in Figs. 1 and 2 were not significantly different from 1.0 ($P > 0.05$), indicating that any difference between staff and peer marks awarded to a given report was independent of the magnitude of the marks awarded. Smaller numbers of scripts were double marked in the third and fourth years of the study, but a similar profile of results was obtained. Those data are not presented in the same format as Figs. 1 and 2 but are included in the analyses described below.

Validity of students’ marking judgment as a function of their gender. Figures 3 and 4 show all the data from the first 2 yrs of the study plus data obtained from eight cases in the follow-up year plotted as a function of marker gender. Figure 3 shows data from male ($n = 30$) markers, and Fig. 4 shows data from female ($n = 46$) markers. The sample sizes for male versus female markers reflect the sex balance within the total student cohort. The respective correlation coefficients between the peer and staff marks for male and female peer markers were 0.96 and 0.98. The mean mark generated by the male peer markers was 75.2% (SD 14.1) versus a corresponding mean staff mark of 72.3% (SD 12.7). The mean mark generated by the female peer markers was 81.6% (SD 11.6) versus a corresponding
mean staff mark of 79.0% (SD 11.8). There was no significant difference in the average “peer mark minus staff mark” for male versus female markers (P = 0.74, unpaired t-test).

This shows that, on average, the validity of peer marking was independent of sex and that both sexes overmarked to the same extent in relation to staff marking. Furthermore, although the variance of the parameter “peer mark minus staff mark” was somewhat larger for male than female markers (see also the frequency histograms in Fig. 5), an F-test revealed no significant differences (P > 0.05) between the variances of the marks awarded by the staff member compared with those awarded by either sex of peer markers.

Sex bias in peer marking. This analysis was taken a stage further by investigating whether peer markers favor, or discriminate against, either their own or the opposite sex. Regression analysis between peer and staff marks in the situations of “male marker-male writer” and “male marker-female writer” (Fig. 3) showed that the respective correlation coefficients were both 0.96. Similar regression analysis in the situations of “female marker-male writer” and “female marker-female writer” (Fig. 4) showed that the respective correlation coefficients were both 0.97. There was therefore no evidence for any sex bias in the ranking of the marks assigned by either male or female markers. There was also no significant difference between the average “peer mark minus staff mark” for male students marking their male peers versus male students marking their female peers (P = 0.31) or for female students marking their male peers versus female students marking their female peers (P = 0.70).

Quality of students’ marking judgment as a function of the marker’s own score for the assignment. Figure 6 shows the differences between peer-awarded and corresponding staff-awarded marks (expressed as “peer mark minus staff mark”) for each double-marked report plotted as a function of the peer marker’s own score as assigned by the staff marker. These data were obtained for the 31 markers across all 4 yr of the study for whom the sampling process had provided a staff mark both for the marker’s report and for the report that they had marked. As the results clearly show, there was no correlation (r = 0.01) between a student’s reliability as a marker and his/her performance in the same piece of work.

Numerical Impact of Using Peer-Awarded Marks As a Summative Component of the Final Course Mark

The average difference of 2.5–3.0% between peer and staff marks would have translated into a difference of approximately +0.05% in the overall course mark (to which the renal practical report contributes 2%) as a result of using peer, rather than staff, marking. Disregarding arithmetical errors in totaling marks, which were largely eliminated by the redesign of the pro forma report after the first year of the study, the largest discrepancy for an individual student’s mark was 13% (peer mark > staff mark), which would have translated into a difference of +0.25% in the final course mark. Discrepancies of >5% between peer marks and staff marks were identified in just under a quarter (17 of 76) of the staff-marked reports, but only 3% of the peer marks deviated by >7% from the corresponding staff mark.

Fig. 3. Scatterplot of peer marks awarded by male students in relation to the corresponding staff marks. Data were obtained in years 1–3 of the study. The solid line represents the linear regression line; the dashed line represents the line of theoretical perfect correspondence (n = 30; correlation coefficient, r = 0.96; slope of regression line = 1.07). The average peer- and staff-awarded marks were 75.2% and 72.3%, respectively. Open circles denote marks awarded to female writers (n = 20; r = 0.96; regression line not shown); solid circles denote marks awarded to male writers (n = 10; r = 0.96; regression line not shown).

Fig. 4. Scatterplot of peer marks awarded by female students in relation to the corresponding staff marks. Data were obtained in years 1–3 of the study. The solid line represents the linear regression line; the dashed line represents the line of theoretical perfect correspondence (n = 46; r = 0.98; slope of regression line = 0.96). The average peer- and staff-awarded marks were 81.6% and 79.0%, respectively. Open circles denote marks awarded to female writers (n = 31; r = 0.97; regression line not shown). Solid circles denote marks awarded to male writers (n = 15; r = 0.97; regression line not shown).
Feedback From Students

Student feedback is shown in Tables 1–3. Table 1 shows feedback from the first 2 yr of the study, in which students’ views were sought on specific outcomes of the peer marking session, whereas Tables 2 and 3 show feedback gathered in the third and fourth years of the study, respectively, in which students’ views on more generic aspects of peer assessment were solicited. It can be seen from the results shown in Table 1 that the majority of students reported that the process of completing the renal pro forma report and the subsequent peer marking session had been helpful in increasing their understanding of renal physiology and of how to present a laboratory report. Most students also reported that the process had been effective in encouraging them to take care in presenting their work. The written questionnaires gave students the opportunity to make individual comments, and a number of students reported that preparing for the peer assessment session had increased their motivation in completing the report: “It increased the effort I put into completing the pro forma” and “I usually forget about practicals once I’ve done them but this makes sure that I follow them up and actually understand them.” Students also appreciated the interactive nature of the session and thought it “was much better than simply reading out the answers—easier to concentrate.” On the negative side, some students commented that they felt “unqualified to mark another’s work.” Free text feedback from the participating students also showed that the summative aspect of the mark awarded provided strong motivation to submit high-quality work, with frequent comments such as “Made me work harder thinking it was going to be marked.”

In relation to more generic aspects of the use of peer assessment, Table 2 shows that most students (70%) in the third year of the study were comfortable for peer-assigned marks to make a modest (up to 5%) contribution to the overall course mark. There was also a clear student view from this cohort that, in the absence of summative marking by a member of staff, peer assessment is fairer (i.e., less subject to bias) than self-assessment. Students’ confidence in the fairness of the process was confirmed by the results shown in Table 3 (obtained from a different student cohort from the results displayed in Table 2), as only 11% of the cohort considered that peer marking had not given them fair feedback on their work. In terms of their own learning, 45% of students in the fourth year of the study reported that evaluating someone else’s work had been generally helpful for their own learning, with only 21% of students disagreeing with that statement. There were, however, mixed views from this cohort as to whether peer assessment should be used more often, with approximately a third of students in favor, a third opposed, and the remainder adopting a neutral stance.

Additional Benefits of the Peer Marking Process

The procedure adopted was very efficient in staff time compared with the previous practice of staff marking, when ~180 laboratory reports took ~30 staff hours to mark and several weeks had elapsed before the marked reports were returned. In contrast, the peer marking session was completed in <1 h, and students received formative feedback only a few days after submitting their reports, in time for it to be of use in writing up subsequent laboratory reports. Since a single member of staff led the peer assessment session, all students received the same feedback, and any individual queries could be discussed with the whole class. The procedure also ensured that all students who attended the session received feedback on their work, whereas previously students had often failed to collect reports that had been marked and annotated by staff. The success of the peer assessment process adopted for the renal physiology practical report has resulted in the same procedure being adopted for two other laboratory sessions in the same course.
DISCUSSION

This study describes how peer assessment can provide prompt formative and summative feedback on a laboratory report for a large, first-year student cohort. Some care was needed in designing the pro forma report to minimize students’ arithmetical errors when totaling the marks awarded, but, once optimized, the procedure yielded valid peer marks, unbiased by the sex or level of achievement of the marker. These findings are discussed in the context of other reports of peer assessment, particularly those that relate to a structured academic product (e.g., a report or poster) or process (e.g., an oral presentation) in a scientific discipline. Some of the benefits and caveats of peer marks contributing to summative, as well as formative, assessment are also considered.

Correlation Between Peer and Staff Marks

The demonstrated correlation between peer and staff marks \( (r = 0.96–0.98) \) was higher than others reported in the literature. Values of \( r \) between 0.62 and 0.89 have been obtained in a range of biomedical studies in which the peer-assessed assignment has consisted of a practical report in biochemistry (32), an oral presentation in pharmacology (14) or biology (10, 16), or a poster presentation in biology (2) or comparative animal physiology (25). The relatively high correlation between staff and peer marks in the present study is likely to derive from the structured nature of the marking task, the fairly constrained marking schedule that was used, and the systematic, real-time guidance that it was possible to provide students as they assigned marks to each section of the report. It may be significant that the next highest correlation \( (r = 0.89) \) was, as in the present study, obtained from an analysis of peer assessment of laboratory reports (32).

Comparison Between Mean Values for Peer and Staff Marks

A high correlation between peer and staff marks implies good agreement in the ranking of the submitted reports but does not necessarily signify correspondence between the absolute values of the marks assigned to the double-marked reports. In fact, nearly 80% of students in the present study were more generous than the staff marker, with the average peer mark exceeding the average staff mark by 2.5–3.0%, although a difference of \( >6–7\% \) was rarely found. More generous marking by students (on average, by \( 5\% \)) was also reported for second-year biology oral presentations (16), whereas Billington (2) reported a difference of 11% between the average peer mark awarded to a biology poster assignment, compared with the average staff mark, despite there being a 0.80 correlation between the two sets of marks. Interestingly, students do not always mark their peers more generously than do staff. For example, Hughes and Large (14) reported that students were less generous markers than staff by \( \approx 3\% \) on pharmacology oral presentations, and English et al. (6) found that peer marks awarded by first-year medical students for an “open book” data interpretation, in-course assessment were less than those awarded by staff by \( \approx 5\% \). Similarly, in a study of peer assessment of short-answer questions based on bioscience laboratory practicals (9), staff moderation generally resulted in peer marks being raised (by up to 10%).

Although the numerical difference between peer and staff marks in the present study was statistically significant, the

<table>
<thead>
<tr>
<th>Feedback Score</th>
<th>Summative Peer Assessment Is Acceptable</th>
<th>Anonymity Between Marker and Writer Is Important</th>
<th>Summative Peer Assessment Is Fairer Than Self-Assessment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly agree</td>
<td>24</td>
<td>46</td>
<td>38</td>
</tr>
<tr>
<td>Agree</td>
<td>46</td>
<td>24</td>
<td>37</td>
</tr>
<tr>
<td>Neutral</td>
<td>11</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>Disagree</td>
<td>14</td>
<td>7</td>
<td>12</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>5</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

Values are expressed in %; \( n = 182 \) students (96% of the cohort).

Table 2. Feedback obtained in an e-voting session in the third year of the study in relation to peer assessment in a summative context

<table>
<thead>
<tr>
<th>Peer Marking Gave Me Fair Feedback</th>
<th>How Effective Was the Peer Marking Session in Increasing Your Understanding of the Kidney?</th>
<th>How Effective Was the Peer Marking Session in Encouraging You to Take Care in Presenting Your Work?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly agree</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>Agree</td>
<td>48</td>
<td>36</td>
</tr>
<tr>
<td>Neutral</td>
<td>30</td>
<td>34</td>
</tr>
<tr>
<td>Disagree</td>
<td>10</td>
<td>18</td>
</tr>
<tr>
<td>Strongly disagree</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

Values are expressed in %; \( n = 156 \) students (84% of the cohort).

Table 3. Feedback obtained via written questionnaires in the fourth year of the study in relation to generic aspects of peer assessment (its fairness, value for learning, and preferred frequency)
discrepancies were no greater than intermarker variations between staff-awarded marks that have been reported in the literature. For example, very wide ranges of marks were awarded by different staff markers to the same essays submitted by dental and psychology students, respectively (15, 22), whereas Magin and Helmore (18) reported that the individual “rater reliability” for teachers assessing engineering students’ oral presentations, although being superior to that for students, was still “moderately low.”

**Comparison Between Variances of Peer and Staff Mark Distributions**

The SD of peer marks awarded in the present study was very similar to that of the corresponding staff marks. This contrasts with the findings of some studies (14, 18, 32), which have reported more “bunching” of peer compared with staff marks, implying that students may have lacked confidence in using a wide marking range. In contrast, Langan et al. (16) found that the range of marks awarded by students was much greater (as well as being more generous) than those awarded by staff. The similar variances between staff and peer marks in the present study probably reflect the structured nature of the marking task and the use of detailed marking criteria, which meant that students felt sufficiently confident to award marks over a reasonably, but not excessively, wide range.

In conclusion, the mark profiles generated by the peer assessors in the present study compare well with the corresponding profiles generated by the staff marker, as measured via a range of parameters, providing confidence in the robustness of the procedure adopted. The degree of correspondence between the profiles of peer marks and staff marks was also as good as, and in several respects better than, the correspondence reported in other studies.

**Influence of Gender on Peer Marking Validity**

The influence of gender on peer marking validity can be considered from two perspectives: comparisons between the profile of marks awarded by male versus female markers regardless of the sex of the recipient (considered as sex differences) and the extent to which the mark awarded was influenced by the gender of the recipient (considered as sex bias). In relation to gender differences, no significant differences were reported between the average peer marks awarded by male versus female students for biology oral presentations (10), although the validity of marking (determined by the average “peer mark minus staff mark”) for each sex was not compared. A range of findings related to sex bias in assessment carried out by staff has been reported in the literature (4, 8, 21). There are also mixed reports for sex bias in peer assessment carried out by undergraduates. For example, Langan et al. (16) found that male students consistently assigned overgenerous marks for biology oral presentations delivered by other males, by ~2–3% compared with staff marking, whereas no such sex bias was apparent for female peer markers. In contrast, sex bias was not apparent for first-year male or female students in allocating marks to their peers for their contributions to group work in either medicine or science and technology (8).

The lack of evidence for either sex differences or sex bias in the present results (see Figs. 3 and 4) adds further weight to the conclusion that gender issues are unlikely to be a significant cause for concern in adopting peer assessment more widely, at least in the format used in the present study.

**Influence of Marker Achievement on Peer Marking Validity**

The impact of students’ academic ability on their accuracy as markers has been investigated in a number of studies. It was concluded on the basis of self- and peer assessment of essays by dental students (15) that, whereas students across the entire ability range tended to overmark both themselves and their peers, the average discrepancy between peer- and staff-awarded marks was much greater (approximately +18%) for markers in the lower ability range. Another study (34) reported that the final grade of first-year engineering students was a strong (positive) indicator of their accuracy in both peer and self-assessment. A different pattern however, emerged in the peer assessment study by Stefani (32), which indicated that peer markers who had performed least well in the assignment (as determined by their staff-awarded mark) were more stringent than the tutor in awarding marks to their peers (by ~3%), whereas all other peer markers were slightly more generous (by ~2%) than the tutor.

In the present study, the marking judgment displayed by students was found to be independent of the quality of their own assignment (see Fig. 6), an outcome similar to that reported by Ferguson et al. (9). Similarly, in a study (14) of the peer marks awarded to pharmacology oral presentations by students throughout the ability range, there was no systematic correlation with the marks that markers had themselves been awarded for the same assignment. Note, however, that this study (14) did not strictly compare students’ achievement on the assignment with their marking validity, which can only be determined by the closeness with which the mark they awarded approximated to that of the corresponding staff-awarded mark.

In view of the range of findings in the literature, the impact of students’ overall academic achievement on their accuracy as peer markers may be worthy of further study, particularly if peer assessment becomes more widely adopted.

**Benefits and Caveats of Peer Marks Contributing to Summative Assessment**

The present results provide justification that peer assessment can make a modest contribution to summative assessment, although caution should clearly be exercised in this respect. For example, although the average discrepancy between the peer and staff marks awarded was small, the discrepancy relating to an individual student’s mark was occasionally as large as 13%, owing to a more significant lack of judgment by the marker in awarding credit. Furthermore, even though a peer marker may bring excellent judgment to bear in assessing another’s work, this will be undermined if they fail to add up the constituent marks correctly. Even after redesign of the pro forma report to minimize errors in totaling the score, some markers continued to make, albeit small, arithmetical mistakes, and it is difficult to see how such errors could be completely eliminated. Since the renal practical report contributed only 2% to the overall course mark, the discrepancies incurred were considered to be tolerable, particularly in view of the very high correlation between staff and peer marks and the motivational benefits (confirmed by student feedback) of the small summative component attached to the mark awarded. It should also be
remembered that students were given clear instructions to review their marked script at the end of the peer marking session and were given the opportunity to lodge a substantiated appeal against the peer mark awarded if they felt this was justified on the grounds of arithmetical error or poor marking judgement. In fact, very few students (<5 students over the 4-yr duration of the study) took up this opportunity.

**Student Feedback**

In common with findings of several other studies (2, 5, 6, 11, 28, 30), some students questioned their ability to assess the work of their peers and expressed concern about the fairness of the process. In agreement with the findings of Smith et al. (30), this concern has reduced over the years of the study, probably because it has been possible to provide successive cohorts of students with evidence of their predecessors’ reliability in the same task. Furthermore, most students were content for peer-awarded marks to make a modest contribution to their overall course mark. It may be significant in this context that the peer marks awarded for the renal laboratory report in the present study (which averaged 79% over the 4 yr of the study) were typically much higher than the corresponding average mark (of 57%) that was obtained by the same students in their end of year examinations. One could argue that a summative contribution to the final overall course mark from a high-scoring piece of coursework would be to the advantage of most students, thereby predisposing them to favor a summative aspect to the process. However, the majority view that summative peer assessment is fairer than summative self-assessment in the absence of staff marking (Table 2) implies that most students adopted a responsible approach to the process.

**Benefits of Peer Assessment in Developing Students’ Lifelong Learning Skills**

The ability to assess, reflect on, and improve one’s own standard of work is a valuable skill across a wide range of activities and employment, as is the ability to provide colleagues with reliable and objective feedback. It has been suggested (3) that the assessment process in higher education, including the provision of opportunities for students to assess their own and each others’ work, provides a powerful route through which to develop important workplace skills. A number of studies (1, 11, 25, 31, 32, 35) of peer assessment in a biomedical context have reported that, although students find peer assessment challenging, their feedback shows that it can encourage them to “think more,” “learn more,” and “reflect more on their work.” It is interesting that, whereas at least two-thirds of students in the first 2 yr of the study reported that the peer marking session had been valuable in increasing their understanding of the material covered (Table 1), only 45% of the students in the fourth year of the study responded positively when they were asked, in more general terms, whether evaluating another’s work had helped their own learning (Table 3). Since the format of the peer marking session was the same on each occasion, it is unlikely that this discrepancy reflects a significant difference in student satisfaction with the procedure adopted. It is more probable that undergraduates, particularly those in their first year, find it easier to assess their attainment of specific learning outcomes (e.g., their understanding of renal physiology) than their acquisition of more generic learning skills.

Only a third of the students in the fourth year of the study (Table 3) considered that peer assessment should be used more often throughout the course, and a third of them expressed the view that all summative marking should be carried out by staff. This is a very similar profile of responses to that obtained in a detailed survey of feedback from sports science students (35) on the value of formative peer assessment, which concluded that, although more than half of the students found the process useful for increasing their subject knowledge and understanding, it was not a useful strategy for all students, with 20% of the cohort not favoring its more widespread use.

A range of student preferences for the frequency of peer assessed assignments has been reported in the literature. On the basis of student feedback, Ballantyne et al. (1) cautioned against using peer assessment for more than one assignment per year per student cohort, whereas O’Moore and Baldock (24) found that peer assessment sessions for problem-based learning and tutorial assignments were so popular with third-year engineering students that the number per year was increased at the students’ request. Several studies (1, 5, 11, 28, 32, 35) have reported that most students find peer assessment challenging. It is still relatively novel, and student reactions to it are likely to be influenced by a number of factors. For example, it is unlikely that they will welcome peer assessment if they perceive it simply as a potentially unreliable replacement for staff marking. However, evidence for the validity and educational value of peer marking is growing. Furthermore, the resulting reduced turnaround time in providing students with feedback on their performance is significant, especially for large enrollment courses in which academic staff resources are limited. As the practice becomes more common, and the evidence for its benefits increases, it is to be hoped that most students will see the value of such “active” versus “passive” assessment practices and this will overcome any understandable initial reluctance for them to engage with the process.

There are mixed reports in the literature as to whether peer assessment improves learning to the extent of impacting on subsequent examination performance. No significant differences were reported between the examination performances of a medical student group that participated in peer assessment compared with a matched group that did not (6), whereas examination results for physiology topics that had been taught via problem-based writing with peer review were found to be superior to those taught via didactic lectures (27). It would have been interesting to evaluate the impact of peer marking on the quality of subsequent laboratory reports submitted by students participating in the current study. That analysis was beyond the scope of the study reported here but could form the basis of future studies.

**Efficiency of Peer Assessment in Relation to Time and Staff Resources**

Now that the process has been validated, reliable quantitative and qualitative feedback on the renal practical report can be provided to students in <5% of the time that was required for staff marking of the same assignment. That time saving has significant benefits for staff in reducing faculty workload, and students derive important benefits from the resulting rapid
turnaround time in receiving feedback on their performance. Those benefits have now been extended by applying peer assessment to two other first-year physiology laboratory reports, and peer assessment of second-year physiology laboratory reports is currently being trialed. The initial time taken in validating the process (e.g., double-marking reports and analyzing peer versus staff marks) should not, however, be underestimated. Other authors (5, 28) have commented that peer assessment processes can be time consuming to set up and manage, especially at the outset. The experience of the present study has been that the initial investment in time was justified by the educational benefits of the process and the time savings achieved once the process was established.

It could be argued that even greater time savings for staff would be achieved by replacing both staff and peer marking of practical reports with a series of questions in a format suitable for machine marking. It would, however, be difficult to design a computer-marked assignment that reproduces the style of a correctly formatted experimental report with the data presentation, analysis, and interpretation that this involves. Automated marking would also eliminate the educational benefits of peer assessment in terms of students’ learning and their development of generic skills.

**Student Involvement in Constructing Marking Criteria**

A number of studies (see Ref. 7) have suggested that the validity of peer marking improves when students are involved in constructing the marking criteria for the assignment. When this was evaluated quantitatively (16), it was found that students who had participated in the development of assessment criteria, although not achieving higher grades for their own assignments, were slightly more reliable markers than those who had not. However, the situation is not straightforward. For example, it has also been reported (26) that, when first-year biology students peer marked histology posters, “discussion between tutor and student leading to joint construction of marking criteria” was not effective in increasing students’ understanding of those marking criteria unless they were also provided with exemplars of a range of previously graded work. Given the size of the student group in the present study (~180 students/session), it was considered that student involvement in constructing the marking criteria would have been very unwieldy, a conclusion also reached by Hanrahan and Isaacs (11) in their large cohort study, and would not have been justified by the potential benefit.

**Anonymity of Markers/ Writers**

It was difficult in this study to guarantee that markers did not know the identity of the person whose work they were assessing. Additionally, markers were asked to identify themselves on the reports that they had marked to guard against, and if necessary follow up, any irresponsible marking. Despite the size and heterogeneity of the cohort (which was drawn from at least 7 degree programs), it is possible that some students may have known the person whose work they marked and/or their marker. Student feedback in the third year of the study (Table 3) showed that ~70% of the cohort felt it was important that the peer marker should not know the identity of the person whose work they were marking. Similar student views about the need to preserve anonymity of the writer and/or marker have been reported in a number of other studies (2, 6, 30, 32).

Interestingly, “secret” peer assessment (17), in which an individual did not know the identity of their marker, provided a greater spread of marks and more distinction between individuals than when the identity of the marker was known. Coding all reports using, for example, students’ examination candidate numbers, would ensure anonymity between markers and writers, but this would be a complex and time-consuming process with such large student cohorts (see also Ref. 11) and might also overemphasize the assessment aspect of the exercise, which was intended primarily as a learning experience for students. It was therefore decided that, on balance in the present study, the potential costs of preserving anonymity outweighed its potential benefits. Furthermore, the ability to provide (and, if necessary, justify) fair, objective feedback to one’s peers is an important life skill, and it could be argued that developing this skill should be encouraged from the outset of an undergraduate career.

**ACKNOWLEDGMENTS**

The author thanks Stephen J. W. Lisney for comments on the manuscript and the students who participated in the study.

**DISCLOSURES**

No conflicts of interest, financial or otherwise, are declared by the author(s).

**REFERENCES**

16. Langan MA, Wheater PC, Shaw EM, Haines BJ, Cullen RW, Boyle JC, Penney D, Oddekop JA, Ashcroft C, Lockey L, Preziosi RF. Peer...


