Peer-review in practice: Eight years of Aropä

Helen C. Purchase[a]* and John Hamer[a]

[a]*School of Computing Science, University of Glasgow, Glasgow, UK*

Correspondence details:

School of Computing Science

University of Glasgow

18 Lilybank Gardens

Glasgow G12 8QQ

UK

Email for corresponding author: helen.purchase@glasgow.ac.uk

Biographical notes:

Dr Helen Purchase started her interest in e-learning with a PhD in Intelligent Tutoring Systems, and has several publications relating to the use of technology in education. As well as co-creator of Aropä, she leads the GUSTTO project – a social media system for practicing academics to share 'Teaching Tips' with their colleagues.

Dr John Hamer received his PhD in Computer Science from the University of Auckland in 1990, where he served as Senior Lecturer until 2010. He initially created Aropä to support his own teaching. He currently works as a commercial software developer, and has an honorary affiliation with the University of Glasgow.

Peer-review in practice: Eight years of Aropä

Articles discussing and analysing student peer-review activities proliferate the educational literature, typically describing one or more class exercises where students provide feedback on each other's work. These papers usually focus on a peer-review activity designed as a scholarly study, and make conclusions about its success or otherwise. There is not one standard model for 'peer-review' and information on the many different assessment designs used is distributed over an increasing number of publications and websites. This paper provides a meta-review of peer-review activities as they are implemented in practice –using configuration data from over a thousand assignments conducted using an online peer-review system during an eight-year period. We present data on the wide variety of assignment designs and the parameters that comprise them, their rubrics, and comparisons between subject areas. Information on the norms and range of all decisions to be made will encourage instructors (both new to and experienced in conducting peer-review activities) to reflect on and justify the choices they make.

Keywords: peer-review, Aropä, assessment configuration

Introduction

Student peer-review is a well-established pedagogical activity whereby students provide feedback on their peers' work, (unlike exercises where students provide information on fellow team-members' contributions to a team project, which we distinguish by the term 'peer assessment'). Not only does peer-review ensure that student authors receive a range of feedback on their work in a timely manner, it also fosters skills of critical thinking (Bhalerao & Ward (2001)), metacognitive self-awareness (e.g. Topping (2005, 2009), Nicol (2010)), self-reflection (e.g. Mulder et al (2014b), Harland et al (2017)), judgement making (e.g. Topping (1998), Nulty (2011)), skills of "giving and accepting criticism" (Mulder et al (2014b)), as well as helping in demystifying the marking process. (Mulder & Pearce (2007), Topping (2009)).

In designing a peer-review activity, the instructor needs to make several decisions. For example, how many peers' submissions should each student review? What form should the submissions take? How long should students be given to write reviews? What aspects of the submission should reviewers comment on? There are also contextual aspects: if students are working in groups, should each individual submit a report for review, or should there be one report per group? Should students be rewarded for taking the exercise seriously? Should the marks given by peers count towards summative assessment? Existing research publications describing peer-review activities specify decisions made for particular instances only, in the form of case-studies - with such dispersed information, it is impossible to get an overview of the universal practice of peer-review without access to a consolidated data source.

This paper provides a meta-review of peer-review activities as implemented in practice – that is, activities not conducted for the purpose of researching the process and educational effects of peer-review, but part of the learning activities in a class. Such day-to-day activities are rarely reported on: they are not published in the literature, and, while instructors' and students' experiences may be shared with colleagues, this is typically done informally, and internal to an institution. While commercial Learning Management Systems (e.g. Moodle, Blackboard) have access to extensive meta-data about peer-review activities implemented in their systems, this information is not disseminated. Our exclusive access to the data in the non-commercial Aropä peer-review system allows us to report valuable summary statistics on peer-review activities as they happen in practice.

We report on data from over one thousand peer-review activities, include case studies from instructors, and analyse configuration parameters according to subject area. This is exploratory data analysis of eight years of naturally occurring data: there is no

hypothesis to be proven or otherwise. The aim is to report on the form of peer-review activities in practice, as designed by instructors who are using them in their classes for learning purposes (rather than as a vehicle to address educational or cognitive research questions).

The structure of this paper is: we review studies of peer-review activities, and describe configuration parameters in the context of the Aropä system. We present summary data for these parameters, and describe a typical peer-review activity. Five case studies are presented, and comparisons are made between subject areas. We then relate our results to similar ones discussed in the literature.

Studies of peer-review

Several research publications address the educational and cognitive benefits of peer-review, with excellent summaries provided by Nicol et al (2014), Mulder et al (2014b), and Çevik (2015). Publications reporting the outcomes of peer-review activities (including these three) typically describe processes used by a particular instructor in a particular class. They tend to report on one or two activities only, firmly embedded within a specified context, and while details are usually given about the nature of the activity, in many cases these details are incomplete.

For example, Mulder et al (2014a) describe a peer-review activity in a zoology class of 60: each student was required to assess three of their peers' submissions on topics different from the one that they themselves had written on. The students wrote reviews within a week, using a rubric consisting of five open comments, and seven yes/no options with associated textual comments. The online student peer-review management tool (PRAZE, Mulder & Pearce (2007)) was used for this exercise. Anonymous reviewing is assumed, but this detail is not specified.

November and Day (2012) describe peer-review activities using Aropä (Purchase & Hamer (2017)) in the areas of writing skills in general education music and health classes. 121 music students reviewed three peer submissions in a fortnight; 43 health students reviewed two in one week. In both cases, the submissions were written articles, and the rubric comprised giving suggestions for improved writing, as well as specific comments on "clarity and coherence, accuracy, and use of evidence".

Simpson and Clifton (2016) describe students working in groups to create a report of approximately 7000 words; in a class of 37 students, eight reports were produced. Each student wrote an individual review of the report written by one other group. No details are given about the length of time allocated for students to write their reviews, and the only information given about the reviewing rubric is that it contained "Likert-type statements for students to indicate their level of agreement with". Submissions and reviews were distributed anonymously by email by the lecturer (who removed any inappropriate review comments). The lecturer then marked the students' reviews according to quality, clarity and validity.

Cheng et al (2015) conducted a peer-review activity in a biology class of 47 students, each of whom wrote a report on a topic of their choice. An online system distributed the reports anonymously and randomly, five reports per student. The students scored five aspects of the report (knowledge, suitability, correctness, creativity and overall) on a seven-point scale, and commented on the overall report. This process was repeated three times, allowing students to revise their report between each submission. No information is given on how long each review period was.

While the many case study articles on peer-review (see also Li et al (2010), Weaver et al (2016), Brill (2016)) provide interesting information and analysis of individual cases that have led to scholarly publication, there is no comprehensive

analysis of the design of peer-review activities in practice. While there have been analyses of peer-review activities as published in the literature (Falchikov & Goldfinch (2000) compared numerical marks awarded by peers and instructors in 48 published studies, Topping (1998) analysed 109 articles), only Pearce et al (2009) report on unpublished case studies: four peer-review exercises that used their PRAZE system.

A meta-review of peer-review activities in practice is useful because it provides an overview of what is actually happening in classes – across subject areas, disciplines, and institutions. Knowing what others are doing helps in both designing a new activity or reflecting on an existing one: it highlights the range of options available, and reveals the common norms.

The Aropä system

The source of our data is the Aropä peer-review system, an online system that has been provided free, worldwide, continually since 2009. It was voluntarily created by two Computing Science academics (the authors of this paper) who are the sole designers, developers and maintainers of the system. We are therefore in a unique position to report on the range and scope of peer-review activity in practice.

Up to 30th June 2017, 1,103 successful peer-review assignments devised by 116 instructors have been supported at 24 institutions across the world, in 10 countries, in 36 subject areas. Over 37,000 students have written reviews on their peers' work using the system.

System functionality and configurations

Aropä supports the principal peer-review activity: anonymous, randomly allocated peer-reviewing, based on a rubric devised by the instructor, with an interface that allows students to upload their submissions before the submission deadline, write reviews of

peers' submissions allocated to them before the review deadline, and then view the feedback given on their own submission by other students. The instructor specifies the dates, the rubric and the author/reviewer allocation method; a pairing between a submission and a reviewer is known as an 'allocation'.

Based on our analysis of existing systems (see Purchase & Hamer, 2017), the following basic flexibility in assignment specification is important:

- P1: Submission methods. Any type and any number of artefacts can be required for submission.
- P2: Reviewer workload. The number of reviews each student needs to complete can be specified.
- P3: Duration. The number of days allowed for students to do their reviews can be specified.
- P4: Rubric. The marking rubric to be used by reviewers can be of any length, and have any combination of 'closed' (choosing one option from a list), and 'open' responses (writing text), in any order.

Aropä provides additional flexibility in the following aspects of the assignment, by providing options for a range of parameters. Many of these additional features were implemented on request from Aropä users (see Purchase & Hamer, 2017).

- P5: Anonymity. Authors should not know who their reviewers are, but the flexibility for reviewers to know who the authors are is occasionally useful.

- P6: Reviewer Allocation. Student reviewers can be everyone in the class, or only those who uploaded a submission.

- P7: Submission categories: If students are working in groups, they can submit their work as a group, with one submission associated with all group members. Students may associate a topic tag (taken from a pre-defined set) with their submission.

- P8: Allocation method: If submissions are tagged by topic, the instructor can specify that students only review on the topic relating to their own submission, or only on different topic submissions. Students may be asked to work in a group to write a collaborative review.

- P9: Adjustments: Allocations can be made manually, or manually adjusted after having been automatically and randomly created.

- P10: Self-review. Students can be asked to review their own work – an additional self-review allocation is added to the initial allocation list.

- P11: Feedback to authors. The instructor can specify that students see both comments and marks in their reviews, or only the comments – useful if an instructor wishes authors to focus on qualitative responses rather than numeric ones.

- P12: Tutor marking. Instructors and tutors can review authors' submissions, as part of the review process.

- P13: Mark weighting. Different weightings can be associated with the different options in a set of closed responses (represented as radio buttons in the rubric).

- P14: Restricted feedback. Instructors can indicate that reviewers can only see their own feedback if they have done at least one (or all) their allocated reviews.

- P15: Second-level activity. Reviews can be marked. A secondary assignment can be created that takes as input the reviews from the primary assignment, and allows students (or tutors) to mark these reviews. A variation allows authors to provide a response to their own reviewers' comments.

These 15 parameters (designated P1-P15) represent the diversity of assignments that can be conducted in Aropä. Topping (2009, p21) highlights the variety of peer-review models, and the choices that need to be made: "…activities can vary in a number of ways, operating in different curriculum areas or subjects. A wide variety of products or outputs can be peer assessed, including writing, portfolios, oral presentations, test performance, and other skilled behaviours. The participant constellation can vary: The assessors and the assessed may be pairs or groups." Pearce et al (2009) identify the most important parameters as: whether students work individually or in groups; whether groups submit individual or group work; how many reviews each student should write; whether tutors review in addition to students; and whether reviewer-author pairings are constrained by topic or group.

In this paper, we first analyse the configuration of over a thousand Aropä peer-review assignments with respect to these 15 parameters. We then provide some context for the analysis by considering differences according to subject area, and five case studies. The results are then discussed with respect to prior literature.

Peer-review configurations

The Aropä data

There are 3,248 assignments registered in Aropä for the period 1st January 2009 to 30 June 2017. Of these, 1,335 are deemed 'successful'. Many of the remainder (1,688) are assignments created for testing or demonstration purposes; other assignments where fewer than half of the review allocations were completed (225, 7%) are considered unsuccessful. While the reason for this lack of success is unknown to us, discussions with instructors (and our own experience) suggest that these are typically peer-review assignments where students have not been given extrinsic incentives to participate.

Of the 1,335 successful assignments, 194 were based on submissions written by instructors rather than students (using the system for giving students practice in reviewing), and a further 22 utilised tutor-markers only (using the system simply for online marking). There were 76 second-level review assignments – entailing students (60) or tutor-markers (16) providing feedback on the reviews that students had written in a primary peer-review activity.

There are therefore 1,043 successful peer-review assignments and 60 successful peer- review marking assignments for analysis: a total of 1,103. These assignments are contained within a total of 476 'classes'. A class is typically one semester or term duration, and is associated with an academic subject and an instructor, and with a list of enrolled students.

Key performance indicators

To demonstrate the extent of system use and its value as a source of peer-review activity data, we present our Key Performance Indicators (KPIs). Cho & Cho (2011) and Nicol et al (2014) claim that the greatest learning benefit in a peer-review activity comes from

writing reviews, and so one of our main KPIs is the number of unique students who have written reviews. The number of instructors who have returned for repeated use is an indication of prior successful use, and, since our intention is to support large classes in particular, class size is important. Table 1 summarises the KPIs.

Number of successful peer-review assignments	Since 2009	1,103
	in the last two years	519
Number of unique students (worldwide) who have used the system to write at least one peer-review	Since 2009	37,042
	in the last two years	13,360
Number of repeat instructors: that is, instructors who have used the system for more than one peer-review activity	Since 2009	89
Number of higher education institutions with at least one successful assignment	Since 2009	24
Largest class size for one assignment	August 2015, (Commercial Law)	948 submissions 2,716 reviews
Largest number of reviews written in one assignment	October 2012, (Software Engineering)	701 submissions 3,284 reviews
Total number of completed reviews	Since 2009	271,321
	In the calendar year 2016	43,804

Table 1: Aropä KPI usage data (at 30th June 2017)

Assignment configuration

Summary statistics for the configurations of the 1,043 primary peer-review activities are shown in table 2, showing class context information as well as values assigned to the numerical parameters.

1,043 peer-review assignments	minimum	maximum	median	mean	mode
The class					
Number of students enrolled in the class	2	1,057	77	131.7	102
Number of peer-review assignments conducted during the class	1	18	1	2.2	1
The assignments					
Number of students (or student groups) submitting for an assignment	1	948	24	87.5	2
Number of reviews written in an assignment	1	3,284	84	254.7	18
The submissions					
Number of items for each student to submit (P1)	1	16	1	1.2	1
The allocations					
Number of submissions for each student to review (P2)	1	6	2	2.3	2
Number of days between submission deadline and review deadline (excluding outliers of 71 and 118 days; there are two instances of 59 days) (P3)	0	59	5	5.9	7

Table 2: Summary peer-review assignment parameter data.

Options described above (P5-P15) as represented in the 1,043 peer-review assignments are in Table 3. Default values are in bold italics. Of these features, three have been introduced since January 2009 (denoted by *): between-tag allocations (P8, August 2010), applying weights to radio buttons (P13, October 2010) and restricting feedback (P14, March 2011). Percentages in Table 3 relate to the number of assignments for which choice was available for these three features at the time.

Type of submission required (P1)	***PDF/Word (80%),*** plain text (12%), Excel (0%), PowerPoint (1%), ZIP archive (2%), Other (5%)

Anonymity (P5)	***Author identity is not revealed to reviewers (99%),*** Reviewers are shown the name of the Author (1%)
Reviewer allocation (P6)	***Only those who submit are allocated submissions to review (78%),*** Everyone in the class is allocated submissions to review (22%)
Submission (P7)	***Individual (75%),*** group (23%), tagged (2%)
Allocation method (P8) *	***Individual (95%),*** group (2.5%), within-tags (0.5%), between-tags (2%)
Adjustments (P9)	***Random allocation (92%),*** manual adjustment (8%)
Self-review (P10)	***No self-review (96%),*** Authors will review their own work in addition to reviewing peers' submissions (4%)
Show marks in feedback (P11)	***Authors see comments and marks in the feedback from the reviewers (73%),*** Only comments are shown in the feedback from the reviewers (27%)
Tutor marking (P12)	***None (91%),*** Tutors also review submissions (9%)
Mark weighting in rubric (P13) *	***Options in a set of n radio buttons have consecutive weights of 1, 2, 3…n (70%),*** Different weights are associated with the radio button options (30%)
Restrict feedback (P14) *	***All authors see feedback from reviewers (80%),*** Only authors who have completed at least one review see feedback (8%); Only authors who have completed all their reviews see feedback (12%)
Review Marking (P15)	***Reviews are not marked (94%),*** Reviews are marked using a separate Review Marking assignment (6%)

Table 3: Summary peer-review assignment configuration data.

The rubric

Rubric design is arguably the most important aspect of a peer-review activity, since here the instructor decides both what feedback will be useful to authors, as well as the nature of the critical skills to be fostered in the reviewer.

An Aropä peer-review rubric can comprise any number of items, with each item being a set of radio buttons or a comment box. A radio button set can contain any number of labelled options. A typical rubric is shown in Figure 1.

[Figure 1 goes here]

Summary data of the rubrics from the 1,043 successful primary peer-review assignments is shown in Table 4.

1,043 peer-review assignments	minimum	maximum	median	mean	mode
The rubric					
Total number of input items in the rubric	1	59	6	8.7	3
Number of radio button items in the rubric (P4)	0	44	3	4.6	2
Number of comment box items in the rubric (P4)	0	26	4	4.1	1
Proportion of radio button items in the rubric	0	1	0.50	0.46	0.50

Table 4. Summary peer-review rubric design data

The profile is different for the 60 second-level activity rubrics (that is, rubrics used for marking reviews) (Table 5). These rubrics are shorter, and tend to include a greater proportion of radio buttons – indeed, the majority of these rubrics (34, 57%) contain only radio buttons, with no comment boxes.

60 review-marking assignments	minimum	maximum	median	mean	mode
The rubric					

Total number of input items in the rubric	1	5	5	4.7	5
Number of radio button items in the rubric (P4)	1	5	5	3.9	5
Number of comment boxes in the rubric (P4)	0	3	0	0.8	0
Proportion of radio button items in the rubric	0.40	1	1	0.82	1

Table 5. Summary review marking rubric design data

The typical peer-review activity can therefore be described as follows:

> Students submit a single pdf or Word document which represents their own work. After the submission deadline, they are randomly allocated two of their peers' submissions to review anonymously (the authors' names are not revealed). Only those students who have submitted a document are allowed to take part in the reviewing process. The students are given a week to write their reviews, using a rubric that comprises two sets of radio buttons, and one comment box. The values associated with the items in each set of radio buttons are increasing and sequential, and start at 1. After the reviewing deadline, all students who made a submission can see all the reviews that have been written on their work, seeing both the responses given to the radio buttons as well as the comments. The students do not know the identity of their reviewers. Tutors do not review the submissions. This is the only peer-review assignment that the students undertake for this class.

Peer-review in context

This quantitative analysis of Aropä configurations is, of course, based on de-contexualised data: there is no information in the system about learning gains, integration of the activity with other class activities, development of transferrable skills, or student responses. We cannot comment on whether the students found the feedback useful, whether they were asked to use this feedback in further assignments or to discuss it in class, or whether the instructors used the marks awarded for any summative assessment purposes – this information is simply not available to us.

There are three ways we add context to this data: through case studies from frequent users of the system, analysis of activities with respect to subject area, and the results of a survey on incentivisation.

Case studies

The following five case studies, provided by long-term users of Aropä, demonstrate the diversity of assessment designs that have been implemented

Lisa Hau (who teaches Classics at the University of Glasgow) has used Aropä every year since 2011, for eight second year classes. Students are given three essay topics to choose from (e.g. Tragedy, Funeral Speeches, Democracy) and they 'tag' their draft essay according to their topic on submission. Aropä allocates reviewers to essays with a different tag from their own. Students review two of their peers' essay drafts under six categories (e.g. Knowledge and analysis, Argument, Use of primary evidence); for each, a set of radio buttons and a comment box is provided. Each review is then marked by a tutor with respect to Fairness, Comprehensiveness and Helpfulness (using the Aropä 'review marking' feature). Marks are given to the students for the quality of their reviews.

An instructor from the University of Auckland Business School has used Aropä consistently every semester from Semester 2 2010 to Semester 1 2017 (excepting one semester in 2013). He has arranged for 23 different Commercial Law classes (both at first and second year levels of study) to use the system, with two or three peer-review assignments per class. Students submit a report giving legal advice on a problem scenario involving various legal issues, and are asked to review three of their peers' submissions. There is a standard rubric for all assignments, including questions like 'Are the legal principles correctly identified?' and 'Are the arguments developed logically?', with a set of four radio buttons as well as a comment box for students to use in their review. He discusses the marking process at length with the students in a lecture session, giving students the opportunity to practice answering a question themselves and then marking answers of contrasting quality to that question. Not engaging in the review process results in students losing marks allocated to tutorial participation; students are also advised that Aropä participation is taken into account when considering particular circumstances regarding final grades (e.g. marginal fails).

Abdoul Aziz Fall in the Faculty of Medicine at Cyberjaya University College of Medical Sciences in Malaysia have used Aropä extensively since 2014, conducting 190 successful peer-review assignments in 34 topics, with class sizes totalling 3,780 student reviewers. Abdoul's use of the system covers medical areas including Motivation and Emotion, Personality Disorders, and Social Psychology and Religion. Students are in groups, and there is one submission per group: the submissions are the slides from the groups' presentations that all students attend. All students belonging to groups that have submitted their slides are invited to review the group submissions, and most students do so – typically with a turn-around period of less than one day (and sometimes overnight). Students are rewarded with marks for participation.

Todd Whittaker teaches Computing Science at Franklin University, and has used Aropä for three classes per year since the Fall Semester of 2014. These are small classes (typically less than 20 students). There are five assignments for each class: these are draft documents for consecutive stages of a team project; in some cases, the submission is a team effort, in others, submissions are done by individuals. Student then submit revised documents for summative assessment. Todd uses Aropä's 'review marking' feature for each assignment: every student who reviewed the submitted drafts is randomly allocated four reviews to mark, using a rubric that considers the tone of the review, the extent of constructive feedback, and the accuracy of the review. The marks allocated to the drafts by peers form a small part of the final assessment mark, as do the marks allocated to review quality. Not participating in any step of the review process will result in zero marks for that step.

Peter Bier teaches the annual 'Introduction to Engineering Computation and Software Development' class to all first year engineering students at the University of Auckland. Since 2010, he has used Aropä for classes ranging in size from 575 to 876. Students are required to submit a set of Matlab files – as many as 16 files in one assignment. Students review work submitted by three of their peers, with a rubric that covers both style and functionality. With such large class sizes, the standard Aropä feature that allows extensions to be given for late submissions is unwieldly (since each student needs to be explicitly given an extension); Peter gets around this by creating a duplicate 'extension' version of the assignment with later deadlines. While students receive 2% of their final grade for completing their reviews, the feedback they receive is for formative purposes only.

Subjects

The universality of peer-review activities is clearly demonstrated by the range of subjects for which Aropä is used. Topping (1998, p251), in his review of 109 peer-review articles found: "peer assessment is potentially applicable to virtually all areas". Originally developed for peer-review of programming code in Computing Science assignments, Aropä is now used in most subjects offered in Higher Education institutions (see Figure 2). Each Aropä class and assignment has a subject associated with it, usually designated by the instructor, or, if not, by ourselves (based on the content of the submissions and/or our knowledge of the subject area of the instructors). The 476 classes in our data set fall into 37 subject areas (including an 'other' category used for classes that do not fit any of these subjects – one class only).

[Figure 2 goes here]

We group the subjects into six 'disciplines': Arts (**Arts**), Medical, Veterinary & Life Sciences (**MVLS**), Science & Engineering (**SciEng**), Social Science (**SocS**), Academic Skills (**Acad**), and Other (**Other**). The first four of these disciplines classify subjects within the four academic colleges of the authors' university, and 'Academic Skills' relates to generic transferrable subjects (e.g. Research Skills). The number of Aropä classes in each discipline is shown in Figure 3.

[Figure 3 goes here]

We consider associations between the peer-review assignment configurations and the disciplines. We focus on the academic disciplines: Arts, MLVS, SciEng and SocS. Removing 14 academic skills classes (22 assignments) and 1 'other' class (2 assignments) leaves 461 classes and 1,109 assignments for analysis.

Student effort: number of submissions each student must review. Does the effort expected of students vary according to discipline, where effort is represented by the number of reviews students are required to complete? Figure 4 shows the overall distribution, and the distribution with respect to discipline.

[Figure 4 goes here]

An independent samples Kruskal-Wallis test ($n=1,019$, $df=3$, $K=152.06$, $p<0.001$) reveals significant pairwise differences: MVLS < SocS ($p<0.001$); MVLS < SciEng ($p<0.001$); Arts < SocS ($p<0.001$); Arts < SciEng ($p<0.001$); SocS < SciEng ($p=0.019$). These results suggest MVLS and Arts classes tend to require a lower reviewing load than SciEng and SocS classes, with SocS classes also being less demanding than SciEng classes.

Student effort: length of rubric. Does the effort expected of students vary according to discipline, where effort is represented by the length of the reviewing rubric? (Figure 5).

[Figure 5 goes here]

An independent samples Kruskal-Wallis test (n=1,019, df=3, K= 134.43, $p<0.001$) reveals significant pairwise differences: MVLS< SocS ($p<0.001$); MVLS < SciEng ($p<0.001$); MVLS < Arts ($p<0.001$); SocS < SciEng ($p<0.001$); SocS < Arts ($p<0.001$). We conclude MVLS assignments have the shortest rubrics, closely followed by SocS. Arts and SciEng use the longest rubrics.

We calculated the correlation co-efficient between effort as represented by reviewing load and effort as represented by length of rubric, to see whether assignments with a high reviewing load tend to have shorter rubrics. The correlation is positive for all assignments (0.36), indicating that this is not the case overall. However, it is negative for Arts (n=90, r= -0.08) and SocS (n=306, r= -0.16), positive for SciEng (n=201, r= 0.26) and strongly positive for MVLS (n=422, r= 0.49).

Rubric style. Does the nature of the reviewing rubrics differ according to discipline? Are some subjects more likely to use rubrics that provide quantitative (radio button) marks rather than qualitative (comment boxes) feedback? Our metric in this case is, for

each rubric, the proportion of rubric items that are sets of radio buttons. A high proportion suggests a rubric focusing on quantitative, summative data; a low proportion suggests that formative feedback is emphasised (Figure 6).

[Figure 6 goes here]

An independent samples Kruskal-Wallis test (n=1,019, df=3, K= 61.63, $p<0.001$) reveals significant pairwise differences: SocS < SciEng ($p=0.005$); SocS < MVLS ($p<0.001$); Arts < MVLS ($p<0.001$); SciEng < MVLS ($p=0.017$). MVLS assignments tend to have the highest proportion of radio button sets in their rubrics; SocS assignments have the lowest.

Routine activity. Can we identify disciplines where instructors tend to use more than one peer-review assignment in a class? That is, subjects where the peer-review activities are considered a routine activity during the class (rather than simply being a one-off event)? (Figure 7).

[Figure 7 goes here]

An independent samples Kruskal-Wallis test (n=461, df=3, K= 3.25, p=0.354) reveals no significant differences between discipline with respect to the number of peer-review assignments in a class.

Number of days for reviewing. Do different subject categories give different time periods for reviews to be completed? (Figure 8).

[Figure 8 goes here]

An independent samples Kruskal-Wallis test (n=1,017, df=3, K= 109.55, p<0.001) reveals significant pairwise differences: Arts < SciEng (p<0.001); MVLS < SciEng (P<0.001); SocS <SciEng (p=0.017); Arts < SocS (p<0.001); MVLS < SocS (p<0.001). SciEng assignments have the longest period for reviewing, closely followed by SocS. There is no significant difference between Arts and MVLS.

Restricting access to feedback. Aropä allows instructors to prevent students from reading feedback on their own work unless they have completed all or some of their reviews. Does this practice vary between the subject categories? Since this feature was

only implemented in March 2011, only assignments conducted since then are included in this analysis (956, Table 6).

All authors can see the feedback on their submission	782
Only authors who have competed at least one review can see feedback on their submission	65
Only authors who have competed all their reviews can see feedback on their submission	109

Table 6. Choices with respect to restricting students' access to their feedback, over all assignments.

We consider the percentage assignments (within each discipline) for which all students are able to see their feedback (even if they have not themselves engaged in the review process), and the percentage of assignments for which feedback is restricted (Table 7).

	Arts	**MVLS**	**SciEng**	**SocS**
All authors can see the feedback on their submission	81.48	82.73	84.62	78.98
Authors must have engaged in the review process before they can see feedback on their submission	18.52	17.27	15.38	21.02

Table 7. The percentage of assignments for which students are restricted in their access to feedback, according to discipline.

While the majority of instructors are happy for students to take advantage of the efforts of their peers (even if they have not themselves put any effort into reviewing), SocS instructors tend to be stricter regarding restricting access to feedback, with SciEng instructors being more lenient.

Incentives

Aropä cannot collect information about aspects of the activity that are external to the system and which are communicated to students by the instructor. In particular, we have no way of knowing whether students have been given any incentive to complete their reviews. The feature that prevents students from seeing their reviews unless they have completed some or all of their own reviews was implemented in March 2011. Since then 20% of successful assignments have used this feature.

A survey of 41 active Aropä instructors revealed that 10 did not offer any incentive to students with respect to engagement in the peer-review activity, but that others used a variety of methods to encourage students to complete reviews.

Giving marks for participation was the most popular category (12) – students are given marks if they complete their reviews. This is clearly a simple method, since assessment is simply computational: did the student complete their reviews or not? The second most popular category (7) was giving marks for review quality – students are given marks based on the quality of the feedback that they produce. This is often done using the second level 'review marking' feature in Aropä, with tutors or instructors marking the reviews according to a specified rubric.

Some instructors use the grades provided by the peer-reviewers as summative marks (after excluding outliers), contributing to the students' final grade. This is rare, but appears to be successful, since this method has been repeated in subsequent years.

In some cases, students are encouraged to take part because of an implicit incentive – usually because the instructor makes explicit links between this activity and other summative activities. For example, doing reviews gives students access to the rubric used by markers in summative assessment (which they would not get otherwise), or students are reminded that engaging in the peer-review activity will help their performance in later summative exercises. One instructor invites only those students who have completed reviews to an additional workshop where the marking rubric is discussed in detail, thus giving additional guidance to these students for later assessments. Some instructors simply make engaging the peer-review activity mandatory for the award of credit; others simply tell students it is mandatory (with no suggestion that it could ever be optional).

Discussion

Most of the literature that reports on specific peer-review activities simply present the values of relevant configuration parameters; however, some authors include discussion about their configuration choices, and in a few cases, explicit comparisons are made.

Anonymity (P5): In Aropä, reviewers are always anonymous; in only 4% of the assignments were authors identifiable. Mulder et al (2014b) suggest that the fact students felt feedback was "reliable and valid" may have been due to the anonymity of both authors and reviewers. Lu & Bol (2007) performed an experimental comparison of anonymous vs identifiable peer-review, and found that students in the former condition produced more critical feedback, and their own writing skills improved more, when compared with the latter condition. Rotsaerti et al (2017) used a transitional process from anonymous to non-anonymous peer-review in the context of providing comments on student oral presentations, so as to encourage a constructive dialogue in face-to-face discussion sessions. They discovered that by following a staged transition approach,

later non-anonymous feedback was of comparable quality to the earlier anonymous feedback.

Rubric design (P4): The majority of Aropä rubrics comprised an equal balance between comments and radio buttons. Nicol et al (2014) make a rubric-based distinction between 'peer-review' (the rubric consists of only open ended questions, and is designed for feedback) and 'peer-assessment' (the rubric requires students to mark or rate their peers' work quantifiably). They present data that indicates that students are divided as to whether it is a good idea for peers to assign marks (mostly because they believe peers don't have the expertise to do so), and argue that many of the problems that arise from implementing peer-review (fairness, biases, review reliability) can be avoided if students 'review' rather than 'mark' their peers' work. Both Walker (2015) and Peters et al (2017) highlight the benefits of peer-review being used for qualitative formative feedback. While not commenting explicitly on the nature of the rubric, Mulder et al (2014b) recommend the use of 'structured review forms' to help reduce variation in review quality – suggesting that a rubric consisting of only one comment box would not be sufficient. 35 (3.4%) of the rubrics for the 1,043 Aropä primary assignments comprised only one comment box. Aropä allows for the quantifiable marks (given by radio buttons in the rubric) to be suppressed when feedback is presented to the author, ensuring students focus on formative comments rather than summative marks (P11): 27% of the assignments made use of this feature.

Tutor reviewers (P12): Only 9% of the Aropä assignments included tutors as reviewers of the submissions (in addition to peer reviewers). Topping (2009) comments that students react differently to feedback from 'adults' and peers, and that it is a good idea for instructors to check some of the peer-reviews against their own assessments. Pearce et al (2009) describe a case study where one of the reviews received was always

written by an instructor, thus guaranteeing that "at least one review by a reviewer with some experience" (p19). Mulder et al (2014b) recommends adding one tutor reviewer to each submission, so as to address the common complaint from students about review quality variation. Harland et al (2017) included tutor-marking, stating explicitly that students did not know which comments came from peers and which from tutors – although students claimed they would be able to tell the difference and would "likely pay more attention to teacher comments because they were subject experts" (p806). The fact that reviewers are always anonymous in Aropä means tutors would need to explicitly identify themselves in the text of their reviews if knowledge of their status is required.

Review period (P3): Most Aropä assignments allowed a week for reviews to be written (although this period ranges from less than 24 hours to 59 days). Topping (2009, p25) suggests that "a peer assessor with less skill at assessment but more time in which to do it can produce an assessment of equal reliability and validity to that of a teacher." He does not indicate how much this 'more time' might be, and we could find no other comments in the literature about the most appropriate review period.

Number of reviews (P2): The typical review allocation in Aropä is two submissions per student. Topping (2009) highlights that one of the most important aspects of peer-review is that feedback is 'plentiful'. Harland et al (2017) point out that multiple reviews mean there is a chance conflicting opinions might be expressed; in their context (the review of an ecology research proposal grant application), such conflicts occur in real-world scenarios and so they are not concerned about this. Nicol et al (2014) suggest that receiving more reviews means students are more likely to get the feedback they need, rather than the feedback the teacher has chosen and has had time to create. Their students reported that comparing peers' submissions against each other

was useful, and that doing so informed their own work – suggesting that students should review at least two peer submissions. Pearce et al (2009) are more specific – they suggest "instructors should consider a minimum of two to three reviewers per assignment, and ideally more (e.g. four to six)" (p19). In a rare example of a study of the effect of parameter choice, they present student satisfaction data from three consecutive identical instantiations of a peer-review activity over three years, where students' perception of the extent to which the reviews helped in improving their work increased when the number of reviews each student received had increased from 1, through 2, to 3 – although this is not a remarkable result in the context of an assignment where peer-reviews provide formative feedback intended to be used to improve the original submission.

Allocating reviews according to topic (P8): Where allocations are made according to the submission topic, between-tags were more popular (2%) than within-tags (0.5%). Harland et al (2017) comment that they allocated submissions to students who themselves had written on different topics (the Aropä 'between-tags' option), and found no problem in an ecology class with students not having sufficient "subject knowledge and …exposure to different ecological methods and paradigms" (p805) to be able to provide useful feedback. Nicol et al (2014) note that by allocating students to review submissions on the same topic, there is a greater chance students will reflect on their own work by comparison. Pearce et al (2009) suggest that using between-tags allocation of topics reduces the possibility of plagiarism – although this relies on even distribution of topics amongst the class.

Allocating reviews according to student ability: Topping (1998) suggests peers might be matched according to complementary ability (that is, "more expert assess those less expert"), yet also describes a case study (Topping 2009) where review teams

are groups of three students with "roughly similar ability in writing." The pre-2009 prototype version of Aropä included an option for 'streaming' students, but with little demand for this feature, it was discontinued. No requests for it to be reinstated have been submitted – indeed, if students are aware of these groups, this function can be effected through appropriate use of the 'within-tags' feature.

Self-review (P10): Although asking students to self-review their own work as well as their peers is easy to specify in Aropä, only 4% of the assignments used this feature. Nicol et al (2014) suggest that asking students to review their own submission using the same rubric as used for the peer-reviews can encourage students to "rethink" their own assignment as part of the peer-review activity; they also indicate that in other studies they have conducted, self-reflection naturally occurs as a result of the peer-review process, even if no explicit self-review stage is required. Topping (1998) describes conflicting results from the literature when comparing the reliability of self-review and peer-review.

Marking reviews (P15): Patton (2012) advocates summative assessment of the quality of student reviews. Pearce et al (2009) discuss arguments for and against making the reviews summatively assessed by tutors, suggesting this will not necessarily motivate students to improve the quality of their reviews, and noting the additional load this places on tutors (who need to read original submissions as well as multiple reviews). Of the 76 review-marking assignments in Aropä, 16 used tutors for marking the reviews, while 60 entailed reviews being marked by students.

Rebuttals (P15): Harland et al. (2017) introduced a new rebuttal phase into an existing peer review exercise: students were required to provide a "clear rationale for accepting or rejecting all comments." The quality of the revised work was found to be an improvement over the previous instantiation of the activity. Nicol et al (2014)

reinforce the view that rebuttals ("students' responses to [feedback]") are an important aspect of 'productive' learning. 17 of the 60 Aropä peer review-marking assignments required students to provide a response to their own reviewers' comments (only 1.6% of all the peer-review activities).

No useful discussion on the remaining Aropä parameters were found in the literature, suggesting they are not interesting enough to discuss (P1 (submission type and number), P9 (manual adjustments to automatic allocations)), they are flexible features offered by Aropä but not typically used elsewhere (P6 (restrictions on reviewers), P13 (weighting marks in the rubric)), or that no studies reflect on their effect on learning outcomes (P14 (restrictions on viewing feedback), P7 (group vs individual submissions)). No existing literature reports explicitly on comparisons between disciplines; when more than one subject area is discussed (e.g. Mulder et al (2014b), Huston et al (2015), Reinholz (2016)), these are three or four individual case studies from which generalisations cannot be made.

Conclusion

This paper demonstrates the range of peer-review activities that are conducted in practice, and the many decisions instructors need to make in preparing a peer-review assignment. These are not speculative or theoretical assessment ideas: this is what instructors are actually doing in their classes – over many institutions and subject areas.

Some instructors have published papers reporting their use for Aropä (in most cases unbeknownst to us in advance) in Engineering (Patterson, 2009), Music (November (2011), Health (November & Day, 2012), Scientific writing (Finlay et al, 2012), Physiotherapy (MacDonald, 2013), Anatomy (Welsh, 2014), and over several disciplines (Huston et al, 2015). The assignments analysed in this paper were not conducted primarily for the purpose of academic publication, but, we assume, because

instructors are aware of the evidence base demonstrating that peer-review is a worthwhile learning activity.

Our analysis of the Aropä data demonstrates that practical peer-review activities do not follow one simple model, and different instructors do very different things – different ways of allocating submissions to reviewers, different means of restricting access to reviews, different timings, different anonymity parameters, different length rubrics etc. Each of these parameters is a particular choice made by an instructor when preparing a peer-review assignment. While it is useful to know the configuration norms, parameter choice at extremes does not necessarily result in unsuccessful assignments. One assignment has a rubric containing 59 items (35 radio button sets and 24 comment boxes), where students were required to review four of their peers' submissions; in one class, students are required to submit 16 separate files for each assignment. Several assignments give students less than one day to prepare their reviews.

There are some variations between academic disciples. Science and Engineering instructors tend to be more demanding of their students (but also give longer time for the review period); Social Science students are less likely to be allowed to read their reviews if they have not written any; the rubrics associated with MVLS assignments tend to more quantitative than those in Social Science; Arts instructors prefer long rubrics.

While we might like to think that students believe that what we ask them to do is essential for their learning, without incentives to complete their reviews many students will not do so. Students often complain that they think that marking is the instructors' job, or that an assignment activity should be complete at the time of submission – reviewing is seen as additional unnecessary work. Interestingly, students also tend to welcome the feedback from their peers. They therefore sometimes do not appreciate the

necessary balance of the system: peer-feedback will only be created if students provide it.

This data analysis comes with the caveat that only peer-review assignments as represented in Aropä have been considered; there will be many other peer-review activities conducted – on paper or using other online systems (e.g. PRAZE (Mulder & Pearce, 2007), PeerGrade (https://intercom.help/peergrade/)). However, the amount of Aropä data (1,103 successful assignments, 116 instructors, 36 subject areas, 24 institutions, 10 countries) is sufficient to provide a comprehensive overview.

Aropä is well-established as a reliable and robust peer-review system, but is more than simply an academic service. The ability to collect and analyse extensive real-world practical data provides an opportunity to gain a comprehensive overview of peer-review activities. We believe that it is too early to look at trends, but expect that in a few years' time we will be able to report on whether there are changes in the way peer-review activities are designed over time. Our collection and analysis of real-world data is useful for supporting and encouraging instructors (both novices and experts) in preparing, conducting, and reflecting on successful peer-review assignments in their classes.

Acknowledgements. We are grateful for the support of the School of Computing Science at the University of Glasgow and the Faculty of Science at the University of Auckland for providing hosting services. Many instructors over the years have been willing to try out the various features of Aropä, and we are grateful to those named in this paper for providing case studies. Ethical approval for this study was given by the

Ethics Committee of the College of Science and Engineering, University of Glasgow (ref: 300160176).

References

Bhalerao, A. & Ward, A. (2001) Towards electronically assisted peer assessment: a case study, *ALT-J*, 9:1, 26-37, DOI:10.1080/09687760108656773

Brill, J.M. (2016) Investigating peer review as a systemic pedagogy for developing the design knowledge, skills, and dispositions of novice instructional design students. *Educational Technology Research and Development*, 64: 681-705.

Cheng, K-H., Liang, J-C. and Tsai, C-C. (2015) Examining the role of feedback messages in undergraduate students' writing performance during an online peer assessment activity. *Internet and Higher Education*, 25: 78-84.

Çevik, Y-D (2015) Assessor or assessee? Investigating the differential effects of online peer assessment roles in the development of students' problem-solving skills, *Computers in Human Behavior*, 52, 250-258, DOI:10.1016/j.chb.2015.05.056

Cho, Y-H & Cho, K. (2011) Peer reviewers learn from giving comments, *Instructional Science* 39, 629–643, DOI:10.1007/s11251-010-9146-1

Falchikov, N. and Goldfinch, J. (2000) Student peer assessment in Higher Education: a meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70:3, 287-322.

Finlay, C., Tierney, A., McVey, M., and Gray, J. (2012) Online peer review of scientific writing – Aropä. In *5th Annual University of Glasgow Learning and Teaching Conference*. Available at: http://eprints.gla.ac.uk/91940/.

Harland, T., Wald, N. & Randhawa. H. (2017) Student peer review: enhancing formative feedback with a rebuttal, *Assessment & Evaluation in Higher Education*, 42:5, 801-811, DOI:10.1080/02602938.2016.1194368

Huston, E., McVey, M., McCallum, S., Deeley, S., Hau, L. and Finlay, C. (2015) Diverse use of the web-based peer assessment tool, Aropä. *Scottish QAA Enhancement Themes Conference*. Available at: http://www.enhancementthemes.ac.uk/pages/searchresultdetails/docs/case-studies/diverse-use-of-the-web-based-peer-assessment-tool-aropa

Li, L., Steckelberg, Lui, X. and A.L. (2010) Assessor or assessee: how student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, 41:3: 525-536.

Lu, R. and Bol, L. (2007) A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback, *Journal of Interactive Online Learning*, 6:2, 100-112, ISSN: 1541-4914

MacDonald, D. (2013) Online peer assessment in undergraduate physiotherapy education. *Scottish QAA Enhancement Themes Conference*. Available at: http://www.enhancementthemes.ac.uk/pages/docdetail/docs/paper/online-peer-assessment-in-undergraduate-physiotherapy-education.

Mulder, R.A. & Pearce, J.M. (2007). PRAZE: Innovating teaching through online peer review. In *Proceedings of ASCILITE - Australian Society for Computers in Learning in Tertiary Education Annual Conference 2007*, 727-736.

Mulder, R., Baik, C., Naylor. R. & Pearce, J. (2014a) How does student peer review influence perceptions, engagement and academic outcomes? A case study,

Assessment & Evaluation in Higher Education, 39:6, 657-677, DOI:10.1080/02602938.2013.860421

Mulder, R.A., Pearce, J.M. and Baik, C. (2014b) Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education*, 15:2, 157-171.

Nicol, D. (2010) The foundation for graduate attributes: developing self-regulation through self and peer-assessment. *Quality Assurance Agency for Higher Education report*. http://www.reap.ac.uk/reap/public/Papers/DN_The%20foundation%20for%20Graduate%20Attributes.pdf

Nicol, D., Thomson, A. and Breslin, C. (2014) Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education*, 39:1, 102-122.

November, N. (2011) Literacy loops and online groups: promoting writing skills in large undergraduate music classes, *Journal of Music History Pedagogy*, 2:1, 5-23.

November, N. and Day, K. (2012) Using undergraduates' digital literacy skills to improve their discipline-specific writing: a dialogue. *International Journal for the Scholarship of Teaching & Learning*, 6:2, article 5.

Nulty, D. (2011) Peer and self-assessment in the first year of university, Assessment & Evaluation in Higher Education, 36:5, 493-507, DOI:10.1080/02602930903540983

Patterson, D. (2009) Two peer review systems for summative assessment in a Green

Processing engineering course. *Conference on Engineering Our Future: Are We up to the Challenge?* Engineers Australia, 2281-2295.

Patton, C. (2012) 'Some kind of weird, evil experiment': student perceptions of peer assessment, *Assessment & Evaluation in Higher Education*, 37:6, 719-731, DOI:10.1080/02602938.2011.563281

Pearce, J., Mulder, R. & Baik, C. (2009) *Involving Students in Peer Review: Case Studies and Practical Strategies for University Teaching*, Centre for the Study of Higher Education, University of Melbourne.

Peters, O., Körndle, H. & Narciss, S. (2017) Effects of a formative assessment script on how vocational students generate formative feedback to a peer's or their own performance. *European Journal of Psychology of Education*, DOI:10.1007/s10212-017-0344-y

Purchase, H. and Hamer. J. (2017) Perspectives on peer-review: eight years of Aropä. *Assessment & Evaluation in Higher Education*, 1-15.

Reinholz, D. (2016) The assessment cycle: a model for learning through peer assessment, *Assessment & Evaluation in Higher Education*, 41:2, 301-315, DOI:10.1080/02602938.2015.1008982

Rotsaert, T, Panadero, E. & Schellens, T. (2017) Anonymity as an instructional scaffold in peer assessment: its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education*, DOI:10.1007/s10212-017-0339-8

Simpson, G. and Clifton, J. (2016) Assessing postgraduate student perceptions and measures of learning in a peer review feedback process. *Assessment & Evaluation in Higher Education*, 41(4): 501-514.

Topping, K. (1998) Peer assessment between students in colleges and universities, *Review of Educational Research*, 68:3, 249-276

Topping, K. (2005) Trends in peer learning, *Educational Psychology*, 25:6, 631-645, DOI:10.1080/01443410500345172

Topping, K. (2009) *Peer Assessment, Theory Into Practice*, 48:1, 20-27, DOI:10.1080/00405840802577569

Walker, M. (2015) The quality of written peer feedback on undergraduates' draft answers to an assignment, and the use made of the feedback, *Assessment & Evaluation in Higher Education*, 40:2, 232-247, DOI:10.1080/02602938.2014.898737

Weaver, K.F., Morales, V., Nelson, M., Weaver, P.F., Toledo, A. and Godde, K. (2016) *CBE—Life Sciences Education*, 15, ar15.

Welsh, M. (2014) The impact of a new interactive anatomy curriculum for nursing students. *Higher Education Academy STEM Conference*. 2014.

Figure captions

Figure 1. A typical Aropä rubric, including radio buttons and comment boxes. This rubric is used by Gordon Curry at the University of Glasgow for a Geological History peer-review assignment.

Figure 2: Subjects which have used Aropä for at least five classes. The other subjects with fewer than five classes are: Chemistry, Music, Geography, Mathematics, Modern Languages, Economics, Film Studies, Education, Geology, Health Science, Public Policy.

Figure 3. The number of classes in each discipline.

Figure 4: Number of submissions each student was asked to review. Overall distribution and distribution within each discipline. The rectangles show the upper and lower quartiles, the horizontal lines show the median, and the black circles show the mean.

Figure 5: Number of items in the rubric. Overall distribution and distribution within each discipline.

Figure 6. Proportion of radio button items in the rubric: 0 represents rubrics that contain only comment boxes, 1 represents radio buttons solely containing sets of radio buttons. Overall distribution and distribution within each discipline.

Figure 7: Number of assignments per class. Overall distribution and distribution within each discipline.

Figure 8: Number of days given for the reviewing period (excluding outliers of 71 and 118 days; there are two data points at 59). Overall distribution and distribution within each discipline.