A systematic review of tools that support peer assessment

Andrew Luxton-Reilly*

Computer Science Department, The University of Auckland, Private Bag 92019, Auckland, New Zealand

(Received 6 July 2009; final version received 20 September 2009)

Peer assessment is a powerful educational technique that provides significant benefits to both staff and students. Traditionally, peer assessment has been conducted using pen-and-paper in small classes. More recently, online tools have been developed to enable peer assessment to be applied in large class. In this article, the tools that support peer assessment are reviewed and analysed, revealing the common features and significant differences. Future directions for research on peer assessment tools are suggested.

Keywords: peer assessment; peer-assisted learning; peer review; social learning

1. Introduction

The massification of tertiary education has impacted on quality and quantity of interactions between instructors and students (Ballantyne, Hughes, & Mylonas, 2002). Although the opportunities for instructors to provide detailed feedback on students' work have decreased, a high degree of individualized feedback for students can be maintained by engaging them in tasks that promote learning by interacting with each other. Hamer et al. (2008) report growing interest in the use of contributing student pedagogies among computer science educators. They define a contributing student pedagogy (CSP) as:

A pedagogy that encourages students to contribute to the learning of others and to value the contributions of others (p. 195).

Contributing student pedagogies characteristically involve the use of new (web-based) technologies. They encompass a wide range of activities, including that of peer assessment which has been defined as:

…an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status (Topping, 1998).
Peer assessment is not a new assessment strategy. Peer assessment has been used in many institutions for more than 50 years (Sluijsmans, Brand-Gruwel & van Merrinboer, 2002), in a wide range of higher education contexts such as academic writing, science, engineering, business and medicine (Falchikov, 1995, Freeman & McKenzie, 2002). Peer review has been used as a learning process to improve the quality of computer programs for at least 30 years (Anderson & Shneiderman, 1977).

In a review of the peer assessment literature, Topping (1998) concludes that peer assessment has been used in a wide variety of contexts and that it can result in gains in the cognitive, social, affective, transferable skill and systemic domains. The majority of the studies reviewed showed an acceptably high level of validity and reliability. A subsequent review of peer assessment by Dochy, Segers and Sluijsmans (1999) showed that peer assessment can be valuable as a formative assessment method, and that students find the process sufficiently fair and accurate.

Ballantyne et al. (2002) report significant benefits of peer assessment, but at the cost of significant administrative overheads. Online tools can alleviate this overhead by providing appropriate administrative support for the management of the peer review process. The use of these tools enable peer assessment to be used in contexts such as large classes where it would be infeasible without such support. Furthermore, a number of features such as anonymous online discussion or automated weighting of reviews cannot be provided by traditional pen-and-paper or face-to-face peer reviews.

The experience of using an online tool for reviewing assignments is qualitatively different to both face-to-face reviewing and using pen-and-paper to mark. An early study by Price and Petre (1997) of instructors using electronic marking reported numerous benefits over paper-based marking, including: improved legibility, easy reuse of comments, faster turn-around time, lower administrative overheads and fewer administrative errors.

Plimmer and Apperley (2007) note that the act of marking paper-based assignments often involves scanning the assignments, physically reordering the scripts and making annotations on the scripts as a reminder of the critical points, summarising grades and as feedback to students. The location of annotations on paper provides an easily identifiable reference point which is clumsy to replicate with an online system. The authors advocate the use of digital ink as a means to retain the traditional advantages of pen-and-paper marking while using electronic systems to relieve the administrative burden imposed by paper.

Murphy and Wolff (2005) compared ‘Minute Papers’ created electronically with those created using pen-and-paper. They found that the response rate using pen-and-paper was higher, but the length of the student responses was much higher in the electronic version.

McLuckie and Topping (2004) note that although both face-to-face and online peer assessment activities involve many similar skills, there are important differences. Face-to-face activities involve socio-affective elements which are difficult to develop in online interactions. Other skills, such as interactive process management, that are essential for online environments are less critical for face-to-face interaction. Figl, Bauer, and Mangler (2006) noted qualitative differences between reviews conducted between teams using traditional pen-and-paper, an online tool, and face-to-face. Students reported that communication was easier face-to-face, and that it was significantly easier to give hints and helpful feedback using pen-and-paper compared to filling in an online form.
The significant qualitative differences observed when different mediums are used to conduct peer review highlight the importance of reviewing the research on tools that support peer assessment. This is of particular interest to computer science educators because the majority of the tools described in this review have been developed by computer science instructors for use in computer science classrooms.

Webster and Watson (2002) claim that the paucity of review articles published in the information systems field impedes research progress. Although the literature on peer review has previously been reviewed, the literature on the tools that support peer assessment has not. In this article we review the currently available tools, compare and contrast the features provided by the tools and analyse these features with respect to the findings from the literature. The following research questions are addressed.

1. What are the common features and important differences between online tools?
2. How does the implementation of features relate to the main findings reported by reviews of the peer assessment literature?
3. What directions for future research are indicated?

2. Method
A systematic literature review is a process that seeks to aggregate empirical data using a formal protocol. Kitchenham describes the process as:

'a means of evaluating and interpreting all available research relevant to a particular research question of topic area or phenomenon of interest' (Kitchenham, 2004).

The procedures for practising evidence-based literature reviews have been well documented in the medical domain (Sackett, Richardson, Rosenberg, & Haynes, 1997). More recently, the steps for conducting evidence-based reviews in software engineering have been identified and documented (Kitchenham, 2004). Brereton, Kitchenham, Budgen, Turner, and Khalil (2007) report that systematic literature reviews help researchers rigorously and systematically aggregate outcomes from relevant empirical research.

2.1. Data sources and study selection
Primary studies were identified by searching the IEEE Xplore, ACM Digital Library, Google Scholar, Citeseer, ScienceDirect and SpringerLink electronic databases. The Journal of Computer Assisted Learning, Computer Science Education and Computers and Education were also searched. The title, abstract and keywords were searched for the phrases ('peer assessment' OR 'peer review' OR 'peer evaluation'). As not all of the databases supported boolean phrases in the same way, the search was adapted as required to obtain equivalent results.

The title and abstracts of the search results were assessed for relevance. Studies that mentioned the use of peer assessment in large classes were scanned to determine if technology was used to assist the assessment process. Studies that mentioned the use of software to support peer assessment were scanned to determine the nature of the software and how it was used.

To be included in this review, the software must have been designed specifically for the purpose of supporting peer assessment activities. This excluded a number of
studies that discussed the use of standard communication software such as email (Downing & Brown, 1997), forums (Mann, 2005) and wikis (Lutteroth & Luxton-Reilly, 2008; Xiao & Lucking, 2008) for peer assessment. Tools such as TeCTra (Raban & Litchfield, 2007) and SPARK (Freeman & McKenzie, 2002) that are designed to be used in the context of group projects to support the assessment of an individual’s contribution within a team are explicitly excluded from this study. Although these kinds of peer reviews have elements of commonality with the peer review of artefacts, the review process is qualitatively different. In the review of teammates, students are typically assessing in a competitive way (because a higher grade for a teammate normally results in a lower personal score), and they are often required to evaluate personal qualities and impressions built up over time, rather than assessing a distinctive artefact at a specific time.

Studies that describe a proposal for a tool that has not yet been built, such as RRAS (Trivedi, Kar & Patterson-McNeil, 2003), or that describe a prototype such as PeerPigeon (Millard, Sinclair, & Newman, 2008), which has not been used in a classroom at the time of publication, are excluded from this review.

In summary, studies that describe a software tool designed for peer review in an educational setting and used in at least one course are included in this review. The following software tools are considered outside the scope of this review:

- tools used for the peer review of an individual contribution to team;
- standard technologies designed for another purpose and used for peer review (e.g., word processor change tracking, email, forums and wikis);
- tools designed for peer review that have not been implemented and used in the classroom; and
- conference management tools and other software designed to manage peer review in a professional rather than educational setting.

The reference lists of all primary research reports were searched for other candidate reports.

3. Results

The results of the review are organised into three subsections based on the kind of tools that were identified during the review process. The first subsection summarises generic tools that have been designed to be flexible and support peer assessment in a variety of different disciplines and contexts. The second subsection summarises tools that have been designed to support peer assessment in a specific domain such as the review of a specific kind of artefact such as written reports or computer programs. The final subsection summarises tools that have been purpose-built for a specific course, or which require manual modification to the software to adapt it for use in other contexts.

Each subsection contains a table that summarises the relevant tools. It lists the name of the tool and the year of the first published report about it. Rubric designs are described as ‘flexible’ if the administrator has the ability to modify the rubric for a given assessment, and ‘fixed’ if the rubric cannot be modified. The rubric criteria are coded as ‘b’ if the tool supports boolean criteria (e.g., check boxes), ‘d’ if the tool supports discrete choices (such as a drop-down list or a forced choice between a finite number of specified criteria) ‘n’ if the tool supports numeric scales (e.g., rating a
solution on a 1–10 scale), and ‘t’ if the tool supports open ended textual comments (e.g., suggestions to improve the solution). If the quality of the reviews is assessed, then the source of the feedback is noted (i.e., either a student or an instructor evaluates the quality of the reviews). The opportunity for dialogue to occur between the reviewers and the authors is coded. The way in which the tool allows workflow to be specified by the instructor is listed. Finally, a summary of the kinds of evaluation performed with the tool is included.

3.1. Generic systems

A number of the systems reported in the literature are designed to be highly configurable and support peer review activities in a wide range of disciplines and contexts. Although some systems have only been used in a limited context at the time of publication, the design of those systems indicates that they could be used in a variety of disciplines and context. This section describes these very flexible systems.

Table 1 summarises the generic peer assessment tools.

3.1.1. PeerGrader

The PeerGrader (PG) system reported by Gehringer (2000) allows students to submit an arbitrary number of web pages for review, allowing students to include multimedia resources. Reviewers and authors are able to communicate anonymously via a shared web page. After the initial feedback phase, authors are given an opportunity to revise their work. At the end of the revision period, the reviewers are required to allocate a grade. When the reviews are completed, the students are required to grade the reviews on the basis of how helpful and careful the review was.

An initial evaluation of PeerGrader in a standard data structures and algorithms course exposed some problems with review allocations. Because assignments were allocated to reviewers on the basis of student enrollments, students who didn’t submit assignments or reviews (because of dropping the course, or simply choosing not to

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Rubric design</th>
<th>Rubric criteria</th>
<th>Discuss</th>
<th>Backward feedback</th>
<th>Flexible workflow</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PeerGrader</td>
<td>2000</td>
<td>Flexible</td>
<td>b,d,n,t</td>
<td>Shared page</td>
<td>Student</td>
<td>No</td>
<td>Student survey</td>
</tr>
<tr>
<td>Web-SPA</td>
<td>2001</td>
<td>Flexible</td>
<td>d,n,t</td>
<td>Public comments</td>
<td>None</td>
<td>Fixed</td>
<td>Validity, performance improvement</td>
</tr>
<tr>
<td>OPAS</td>
<td>2004</td>
<td>Flexible</td>
<td>b, d, n, t</td>
<td>Debrief</td>
<td>None</td>
<td>Script</td>
<td>Student survey</td>
</tr>
<tr>
<td>CeLS</td>
<td>2005</td>
<td>Flexible</td>
<td>b,d,n,t</td>
<td>Peers instructor</td>
<td>Unknown</td>
<td>Script</td>
<td>Validity</td>
</tr>
<tr>
<td>PRAISE</td>
<td>2005</td>
<td>Flexible</td>
<td>b,t</td>
<td>None</td>
<td>None</td>
<td>Fixed</td>
<td>Student survey, usage statistics</td>
</tr>
<tr>
<td>Aropä</td>
<td>2007</td>
<td>Flexible</td>
<td>b,d,n,t</td>
<td>None</td>
<td>Student</td>
<td>Limited</td>
<td>Student survey, staff interview</td>
</tr>
</tbody>
</table>
participate) caused other students to receive too few assignments to review, or to receive too few reviews on the assignments they submitted. As reviewing can only begin after an assignment is submitted, assignments that were submitted late left little time for the reviewers to complete their reviews. Gehringer notes that dynamically allocating assignments to reviewers may go some way towards alleviating these problems.

3.1.2. Web-SPA

Web-SPA (Sung, Chang, Chiou, & Hou, 2005) is designed to guide students through self and peer assessment activities. Instructors have some flexibility to configure the type of activity by configuring parameters such as setting a group or individual assignment, and defining the method of scoring used by the rubric (discrete scale, percentage or no scoring). An instructor can define criteria which are scored according to the method chosen in the initial configuration.

The Web-SPA system uses a fixed workflow to progressively engage students in the peer assessment activity. Initially, the students assess themselves. Having completed an evaluation, they compare their own evaluation with others in their group. The groups select the best and worst examples. The system will randomly present each individual with exemplars of the best and worst cases chosen by other groups to review. Once the reviews have been conducted, the system presents the best and worst examples from the entire class. The act of re-reviewing exemplars is designed to help students identify what is good and bad in a given assignment.

The authors conducted a study with 76 high school students in a computer and information science course. The study found considerable consistency between instructor and peer marks. It also found that the quality of work improved after the peer review activities.

3.1.3. Online peer assessment system

The Online Peer Assessment System (OPAS) (Trahasch, 2004) has been designed to support a wide range of peer assessment activities with flexible submission and marking criteria. Collaboration scripts are used to formalise the structure and workflow of the peer assessment process. An artefact submitted for review can be a single document or a zip file. Submissions can come from individual authors, groups, or the instructor. Reviews can be assigned randomly, manually, or using a combination of random and manual. The system supports the allocation of reviews within groups. The review rubrics are flexible and contain criteria that can be assessed using radio buttons, list boxes, numeric scales or with open-ended feedback. Multiple review cycles are supported. An overview of the rankings and criteria is displayed to students at the completion of the review and the best example of each is displayed. A forum supports discussion after the completion of the reviews.

The system was evaluated with a class of 76 students enrolled in an algorithms and data structures course in computer science. A student satisfaction survey was completed in which students were generally positive.

3.1.4. Collaborative e-learning structures

Collaborative e-Learning Structures (CeLS) (Ronen, Kohen-Vacs, & Raz-Fogel, 2006) is a system designed to support collaborative learning activities, including peer
review with flexible work processes using collaboration scripts. An instructor can create new activities or use the structure of an existing activity. The assessment activities can include all the standard elements of a web form, but may additionally include activities that involve ranking or sorting a set of artefacts.

A prototype of the CeLS system was piloted in 2003–2004 in Israel by nine universities, five schools and four in-service teacher courses. In total, 1600 students used CeLS in 48 different courses, although the nature of the collaborative activities was not reported.

Kali and Ronen (2005) report on the use of CeLS for peer review in three successive semesters of an undergraduate educational philosophy course. Students were asked to use the system to evaluate a group presentation on a scale of 1–7 and write feedback in text fields for three grading criteria. After an initial evaluation of the system, a fourth criterion was introduced to allow students to write their own opinion, which was not considered to be a grading criterion. This was intended to explicitly distinguish between objective and subjective viewpoints. A third design iteration introduced the idea of evaluating students as reviewers. Instead of assigning grades according to the results of the peer review, the reviews themselves were evaluated by an instructor and 15% of the students’ grades were calculated based on the quality of the reviews.

3.1.5. PRAISE

PRAISE (de Raadt, Toleman, & Watson, 2005) supports the peer review of documents according to a rubric defined by an instructor. The rubric consists of objective binary criteria, and a holistic open-ended comment. The system waits until a specified number of reviews have been received (e.g., 4 or 5), and thereafter immediately allocates an assignment to review when a student submits. Assignment reviews can be flagged for moderation by the author if they feel that the review is unfair.

PRAISE has been used in at least five different courses, across the subjects of Computing, Accounting and Nursing. Student surveys, usage statistics, time management and moderation required have all been analysed. Student attitudes and practices of novice programmers were found to differ from those of non-programmers (de Raadt, Lai, & Watson, 2007).

3.1.6. Aropā

Aropā (Hamer, Kell, & Spence, 2007) is a generic web-based system that supports the administration and management of peer assessment activities in a variety of contexts. The authors submit files directly to the Aropā system. The reviewers download the files for off-line viewing. The reviews are conducted online by filling in a web form, which is customised by an instructor for the review activity. After students receive their reviews, they may be required to provide feedback on the quality of the reviews (according to a rubric defined by the instructor).

The allocation of authors to reviewers can be automatic or manual. If automatic, the instructor can define a subset of authors, a subset of reviews and the number of reviews to allocate to each reviewer. This system of allocation can accommodate a wide range of peer assessment activities, including intra- or inter-group reviews.
The authors report that Aropā has been used in over 20 disciplines with a diverse range of classes, ranging in size from 12 to 850. It has been used for formative feedback on drafts, critical reflection after an assignment and for summative assessment. Each of these varieties of peer assessment differs in the timing, style of the rubric and degree of compulsion and awarding of marks.

3.2. **Domain-specific systems**

Many of the systems are designed to support peer review activities in a specific domain, such as reading and writing essays, or reviewing Java programming code. Systems that are designed for use in a specific domain are described in this section, and summarised in Table 2.

3.2.1. **Calibrated peer review**

Calibrated Peer Review (CPR) (Chapman & Fiore, 2000) has been designed to help students develop writing skills through peer assessment. Instructors use CPR to create assignments that include specifications, guiding questions to focus the development of a good solution, and examples of solutions with corresponding reviews. Students write and submit short essays on the specified topic. The review process requires students to engage in a training phase where they must evaluate the quality of three sample essays. Their reviews are compared to the samples provided by the instructor and feedback is given to the students about their review performance. Students are not permitted to participate in real reviews until they can perform adequately on the samples. Although widely used, few evaluation studies have been published. A recent report suggests that the use of CPR did not improve writing skills or scientific understanding (Walvoord, Hoefnagels, Gaffin, Chumchal, & Long, 2008).

3.2.2. **C.A.P.**

The C.A.P. system (Davies, 2000) (originally, computerized assessment including plagiarism, and later computerized assessment by peers) is designed for peer assessment of written documents such as research reports and essays. It has evolved substantially from its initial implementation, and continues to be actively studied and improved.

C.A.P. includes a predefined list of comments. Each student can configure the list by adding their own comments. In addition, comments can be assigned a rating to specify how important they are to the reviewer. The review process requires students to summatively assess essays by allocating a numeric value for four fixed criteria (such as ‘Readability’), and to provide formative feedback by choosing comments from the configurable list. The students also provide a holistic comment in an open-text area.

After an initial review period, students are given an opportunity to see the comments that other reviewers selected, and may choose to modify their own review as a result (Davies, 2009). Once the review stage is fully complete, the marks are used to calculate a compensated average peer mark from the ratings submitted by the reviewers. The choice of comments and the ratings are evaluated and used to generate an automated mark for the quality of reviewing (Davies, 2004).
Table 2. Domain-specific peer assessment tools.

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Domain</th>
<th>Rubric design</th>
<th>Rubric criteria</th>
<th>Discuss</th>
<th>Backward feedback</th>
<th>Flexible workflow</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>CPR</td>
<td>1998</td>
<td>Essays</td>
<td>Flexible</td>
<td>b,n</td>
<td>None</td>
<td>Auto</td>
<td>None</td>
<td>Validity, student survey, writing performance</td>
</tr>
<tr>
<td>C.A.P.</td>
<td>2000</td>
<td>Essays</td>
<td>Fixed</td>
<td>n,d,t</td>
<td>Private author/reviewer</td>
<td>Auto</td>
<td>None</td>
<td>Validation, student surveys, higher-order skills, comment frequency, use of review features, compare with self-assessment</td>
</tr>
<tr>
<td>Prakotomat</td>
<td>2000</td>
<td>Programs</td>
<td>Fixed</td>
<td>d,t</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Student survey, usage correlation</td>
</tr>
<tr>
<td>Sitthiworachart</td>
<td>2003</td>
<td>Programs</td>
<td>Fixed</td>
<td>d,n,t</td>
<td>Reviewers</td>
<td>Student</td>
<td>None</td>
<td>Student survey, validity</td>
</tr>
<tr>
<td>SWoRD</td>
<td>2007</td>
<td>Essays</td>
<td>Fixed</td>
<td>n,t</td>
<td>None</td>
<td>Student</td>
<td>Limited</td>
<td>Validity</td>
</tr>
<tr>
<td>PeerWise</td>
<td>2008</td>
<td>MCQ</td>
<td>Fixed</td>
<td>n,t</td>
<td>Public feedback</td>
<td>Student</td>
<td>None</td>
<td>Usage, effect on exam performance, quality of questions, validity</td>
</tr>
<tr>
<td>peerScholar</td>
<td>2008</td>
<td>Essays</td>
<td>Fixed</td>
<td>n,t</td>
<td>None</td>
<td>Student</td>
<td>None</td>
<td>Validity</td>
</tr>
</tbody>
</table>
The students are given the opportunity to anonymously discuss their marks with the reviewer. The reviewer may choose to modify the marks on the basis of the discussion (Davies, 2003).

The C.A.P. system has been used in a number of studies of peer assessment, particularly around the assessment of the review quality. Studies show that marks for reviewing are positively correlated with both essay marks and marks in a multiple choice test (Davies, 2004). The upper two quartiles of students are more critical with comments than the lower two quartiles (Davies, 2006).

3.2.3. Praktomat

The Praktomat system (Zeller, 2000) is designed to provide feedback on programming code to students. The authors submit a program to Praktomat which automatically runs regression tests to evaluate the correctness of the code. The authors who have submitted code to the system can request a program to anonymously review. The reviews use a fixed rubric that focus on a number of specific style considerations that the instructor uses for final grading purposes. The code is displayed in a text area that can be edited to allow annotations to be entered directly into the code. The review feedback is purely formative and plays no part in the final grade, nor is it required. The students can review as many programs as they wish.

Students reported that they found the system useful, both in terms of automated testing, reviewing programs and having programs reviewed by others. The grades obtained by students for program readability increased both with the number of sent reviews and the number of received reviews, although no formal statistical analysis was performed.

3.2.4. Sitthiworachart

The system developed by Sitthiworachart and Joy (2004) was based on the OASYS system. It was designed for the peer review of programming assignments. A fixed rubric is used to assess program style and correctness using Likert-scale ratings. An asynchronous communication tool is provided to allow reviewers to anonymously discuss the assignments they are reviewing throughout the process. An evaluation study showed that the peer ratings correlate significantly with instructor ratings, and that students are better able to make accurate objective judgements than subjective ones.

3.2.5. Scaffolded writing and rewriting in the discipline

Scaffolded writing and rewriting in the discipline (SWoRD) (Cho & Schunn, 2007) is a tool designed specifically to support writing practice. At the time of publication, it had been used in 20 courses in four different universities between 2002 and 2004.

An instructor using SWoRD defines a pool of topics, from which students select those they want to write about and those they want to review. SWoRD balances the allocation of topics, so some students may have a reduced set of choices. Students submit drafts and a self-assessed estimate of grade.

The review structure is fixed, and uses pseudonyms to ensure that the identity of the authors remains confidential. Reviewers evaluate the writing according to
three dimensions: flow; logic; and insight. For each dimension, reviewers rate the work on a scale of 1–7 and provide a written comment about the quality of the writing.

SWoRD assumes that the average grade given by a group of student reviewers is an accurate assessment. It calculates the accuracy of an individual reviewer using three different metrics – systematic difference, consistency and spread. These three metrics are calculated for each of the three dimensions of flow, logic and insight, giving nine measures of accuracy. The nine measures of accuracy obtained for each reviewer are normalised and combined to calculate a weighted average grade for a submitted piece of writing. All the drafts are published with their pseudonyms, ratings and associated comments. The authors revise the drafts and submit final papers, along with feedback about the usefulness of the reviews they received. The review cycle is repeated with the revised papers.

An evaluation of SWoRD was conducted with 28 students in a research methods course. A controlled experiment comparing a single expert reviewer, single peer reviewer and multiple peer reviewers showed the greatest improvement between draft and final paper occurred when the author received multiple peer reviews, and the least improvement occurred with a single expert reviewer.

3.2.6. PeerWise

PeerWise (Denny, Luxton-Reilly & Hamer, 2008a) supports the development of an online multiple-choice question (MCQ) database by students. The MCQs submitted using PeerWise become available for other students to use for revision purposes. When students answer a question, they are required to review the question and enter a holistic rating (0–5) of the quality. They are also encouraged to write a holistic comment in a text area. The author of a question has the right to reply to any given comment, although there is no facility for a continuing discussion. Although the comments are visible to all users, the individual ratings are averaged and only the aggregate rating is displayed. All interaction between students is anonymous.

Numerous studies have evaluated aspects of PeerWise, including the usage (Denny, Luxton-Reilly, & Hamer, 2008b), effect on exam performance (Denny, Hamer, Luxton-Reilly, & Purchase, 2008) and the quality of the questions (Denny, Luxton-Reilly, & Simon, 2009). A study of the validity of the peer assessed ratings (Denny et al. 2009) found that the correlations between ratings of students and the ratings of instructors who taught the course were good (0.5 and 0.58). The authors conclude that students are reasonably effective at determining the quality of the multiple choice questions created by their peers.

3.2.7. peerScholar

The peerScholar (Paré & Joordens, 2008) system was designed to improve the writing and critical thinking skills of students in a large undergraduate Psychology class. In the first phase, students are required to write two abstracts and two essays. The second phase requires students to anonymously assess five abstracts and five essays by assigning a numeric grade (1–10) and writing a positive constructive comment for each piece of work. Finally, in the third phase, students receive the marks and comments as feedback. An accountability feature allows students to submit a mark (1–3) for each of the reviews they received.
A study was conducted to compare expert marks with the marks generated through the peer review process. The authors found that the correlation between expert and peer marks was good, and that it improved when the accountability feature was applied to students.

3.3. Context-specific systems

Some of the systems reported in the literature have been written for use in a specific course and must be rewritten if they are to accommodate other contexts. Although most of these systems have the potential to be developed further in the future, at the time of publication they were bound to the specific context in which they were developed. Table 3 summarises the tools in this category.

3.3.1. Peers

The Peers (Ngu & Shepherd, 1995) system was implemented in Ingres, a commercial database management system. The students were able to anonymously suggest assessment criteria and alter weightings on existing criteria before the submission of assignments. The assignments were allocated to students who were able to anonymously review them and provide marks for the criteria that were cooperatively developed. A short evaluation study found a good correlation between instructor and student marks. However, the student survey that was conducted found that all the students preferred to have instructor assessment in addition to the peer evaluation, suggesting that students did not trust the outcomes of peer assessment.

3.3.2. NetPeas

NetPeas (Lin, Liu, & Yuan, 2001), initially known as Web-based Peer Review or WPR (Liu, Lin, Chiu, & Yuan, 2001), requires students to submit documents in HTML format. Initially, the system only supported a single holistic rating and an open-ended comment, but was later revised to support numerous specific criteria involving both a rating (1–10 Likert scale) and an open-ended comment for each criterion. The system supports the modification of assignments by students which allows drafts to be revised after an initial review period.

Evaluation studies have looked at correlations between review ability and examination scores, different thinking styles, specific and holistic feedback and student attitude. The authors conclude that being a successful author, or a successful reviewer alone may not be sufficient for success in a peer review environment.

3.3.3. OASYS

OASYS (Bhalerao & Ward, 2001) is designed to support self-assessment and provide timely formative feedback to students in large classes without increasing academic workload. It is a hybrid system used to assess students using a combination of multiple choice questions and free-response questions. The system automatically marks the MCQ questions and uses peer review to provide summative feedback to students about their answers to the free-response questions.

Although this system has been designed and used in the context of a programming course, the authors note that it could easily be adapted for more
Table 3. Context-specific peer assessment tools.

<table>
<thead>
<tr>
<th>Name</th>
<th>Year</th>
<th>Context</th>
<th>Rubric design</th>
<th>Rubric criteria</th>
<th>Discuss</th>
<th>Backward feedback</th>
<th>Flexible workflow</th>
<th>Evaluation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peers</td>
<td>1995</td>
<td>Computer Science</td>
<td>Flexible</td>
<td>n</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Student survey, validity</td>
</tr>
<tr>
<td>NetPeas</td>
<td>1999</td>
<td>Computer Science, Science</td>
<td>Fixed</td>
<td>n,t</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Student survey, rubric comparison, thinking styles</td>
</tr>
<tr>
<td>OASYS</td>
<td>2001</td>
<td>Computer Science</td>
<td>Fixed</td>
<td>d,t</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Student survey, admin costs</td>
</tr>
<tr>
<td>Wolfe</td>
<td>2004</td>
<td>Computer Science, Mathematics, Marketing, Psychology</td>
<td>Fixed</td>
<td>n,t</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Usage</td>
</tr>
<tr>
<td>PEARS</td>
<td>2005</td>
<td>Computer Science</td>
<td>Fixed</td>
<td>n,t</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>Rubric comparison</td>
</tr>
</tbody>
</table>
widespread use in other disciplines. An evaluation which compared the time taken to mark paper tests with the time required to mark using the OASYS system was performed. Students using the system received feedback more rapidly with less staff time required than paper-based tests.

3.3.4. Wolfe

Wolfe (2004) developed a system in which students posted their assignments on their own web site and submitted the URL to the peer review system. The reviewers remained anonymous, but they knew who they were reviewing. The reviewers were presented with the list of all the assignments submitted and were expected to submit a score (1–10) and a holistic comment about each assignment. The students were required to submit a minimum number of reviews, but no maximum was set. The web site listed the number of reviews that had already been submitted for each assignment and students were asked to ensure the numbers were roughly even, but the request was not enforced.

The system was used in computer science, mathematics, marketing and psychology courses, but required manual recoding to adapt it to each new context. Wolfe notes that roughly 70% of the reviews were superficial. He reports on the use of the system in a small software engineering course (34 students). The students were required to submit a minimum of 10 reviews, but could conduct additional reviews if desired. The majority of students received more than the minimum 10 reviews, and the majority of those reviews were submitted by students ranked in the top third of the class.

3.3.5. PEARS

PEARS (Chalk & Adeboye, 2005) is designed to support the learning of programming skills. Students submit Java files directly to the system, conduct peer reviews, respond to feedback and may resubmit reviewed work. In the published study, students used two different rubrics to review Java code. The first rubric contained 16 specific binary criteria (yes/no, and not applicable), while the second rubric used a text area to submit open-ended holistic feedback about the strengths and weaknesses of the reviewed work and a single overall score out of 10.

The authors report that over two-thirds of the students prefer to write reviews using holistic feedback, that they preferred receiving holistic feedback, and that the holistic feedback written by students had a significant positive correlation with the marks allocated by a tutor.

4. Discussion

In this section, the common elements of the systems are discussed and unique approaches are identified.

4.1. Anonymity

Ballantyne et al. (2002) suggest that students should remain anonymous to alleviate student concerns over bias and unfair marking. The majority of systems use a double-blind peer review process, ensuring that students remain anonymous throughout the entire process. Bhalerao and Ward (2001) report that anonymity is
a statutory requirement in their institution. Developers of peer review software would be well advised to consider their own institutional regulations regarding the privacy of student grades.

In some cases, student presentations are being assessed (Kali & Ronen, 2005), or students are working in teams on different projects, in which case students performing a review would be aware of the identity of the person they were reviewing. In such cases, there is no need to ensure a double-blind review occurs. Flexible systems such as OPAS and Aropa may be configured to have different levels of anonymity for a given activity (e.g., double-blind, single blind, pseudonym, or open reviewing).

Notably, the system developed by Wolfe (2004) ensured the anonymity of the reviews, but the identity of the authors was known to the reviewers.

### 4.2. Allocation and distribution

A variety of methods are employed to distribute artefacts produced by an author to a reviewer (or most commonly, multiple reviewers). The simplest approach is simply to allocate the reviews randomly. A spreadsheet specifying the allocation of assignments from author to reviewer is accommodated by Aropa, PeerGrader and OPAS. Although some systems (such as Aropa) support the allocation of assignments by groups (to allow inter-or intra-group reviews), many do not.

PRAISE system waits until a minimum number of submissions are received before it begins to allocate assignments to reviewers. After the threshold has been reached, an author that submits an assignment is immediately allocated assignments to review. The major benefit of this approach is a reduction in time between submission and review. However, no analysis of the consequences of this strategy has yet been conducted. It is possible that better students (who complete the assignment and submit early) will end reviewing each other while weaker students who submit later will be allocated weaker assignments to review. Further investigation may be warranted to explore the implications of this allocation strategy.

The use of exemplars can help students to identify what is good or bad in a given assignment. These exemplars can act as a ‘yard-stick’ by which students can measure their own performance and that of others. To ensure that students see a diversity of assignments, OASYS uses the marks for an MCQ test in the distribution algorithm to ensure that each reviewer receives one script from authors in each of the good, intermediate and poor MCQ categories. Web-SPR uses multiple review cycles to ensure that students are exposed to examples of the best and worst assignments. SWoRD makes all the drafts, reviews and ratings publicly available for students to peruse, providing students with the opportunity to compare the best and worst submissions. At the completion of the review phase, OPAS displays a summary of the rankings for each criteria assessed and the top ranked assignment for each criteria is available for students to view. Although Aropa does not systematically provide students with the best and worst reviews, during the allocation phase, it has been seeded with a sample solution provided by the instructor to ensure all students see a good solution.

#### 4.2.1. Unrestricted reviewing

The PeerWise system has no system of allocation. Instead, students can choose to answer as many MCQ questions as they wish. Each time a question is answered a
review is required. Students tend to choose the questions with the highest rating, therefore the better questions are reviewed more frequently. Poor questions are infrequently reviewed.

Wolfe (2004) allowed students to choose who they reviewed (and the identities of the authors were known). The number of reviews that each artefact had received was displayed and reviewers were asked to ensure that they were approximately even, but this requirement was not enforced by the system.

Since reviewing is optional in the Praktomat system, the process of review allocation uses a non-random strategy to encourage students to participate and contribute high-quality reviews. Praktomat uses a set of rules to determine which artefacts are reviewed next. The artefact that has had the minimum number of reviews is selected. Programs whose authors have composed a greater number of reviews are selected by preference. Praktomat tries to allocate reviews mutually, so a pair of authors review each others’ programs.

4.3. Marking criteria

A variety of different approaches to designing marking criteria are apparent. Students are rarely invited to participate in the design of the marking criteria, although numerous authors report that criteria are discussed with students before the review process. Some systems use very specific criteria whereas others use a more holistic general rating.

Systems that are designed to be used in a wide variety of conditions (i.e., those classified as ‘generic’ systems) support instructor-designed marking forms. These forms are typically constructed from the components that make up standard web forms, and support check boxes, discrete lists, numeric scales and open responses in text areas. CeLS has a very flexible design that can accommodate a range of assessment activities including selection, assigning a numeric value and free-text comments, but also more complex assessments such as ranking and sorting.

Systems that are designed to operate in a more restricted domain frequently use a fixed structure for the assessment process and may provide few options for the configuration of the marking schema.

Falchikov and Goldfinch (2000) conducted a meta-analysis that investigated the validity of peer assigned marks by comparing peer marks with teacher marks. They recommend that it is better to use an overall global mark rather than expecting students to rate many individual dimensions. However, Miller (2003) found that more specific, detailed rubrics provided better differentiation of performance at the cost of qualitative feedback. Rubrics that provided more opportunities to comment elicited a greater number of qualitative responses and a larger number of comments.

An evaluation study comparing holistic with specific feedback using the PEARs system found that the majority of students preferred both writing and receiving the holistic feedback (Chalk & Adeboye, 2005). They also found that there was no correlation between the students’ scores and the tutors’ scores when using the rubric with specific criteria, but a significant positive correlation was found between students and tutors when the holistic rubric was used.

PRAISE uses objective binary criteria to ensure consistency between reviewers. A holistic comment is also supported.

Kali and Ronen (2005) report that an explicit distinction between objective and subjective criteria improves the quality of the review. Students like having the option
to express their personal, subjective opinion (which does not contribute to the
ggrading process), and distinguishing their subjective view from the objective grading
criteria improves the correlation between student and instructor marks.

CAP requires students to use numeric scales to summatively assess an essay, but
they are also expected to provide formative feedback by selecting comments from a
defined list. The importance of each comment in the list is weighted by the reviewer,
allowing the CAP system to automatically compare the comments applied by
different reviewers in an attempt to estimate the effectiveness of a given reviewer.

Open-ended feedback requires students to write prose that states their opinion in
a critical, yet constructive way. It is certainly possible that the formative feedback
provided by this approach is more useful to students than that obtained through
check boxes or simple numeric scales. However, further research is required to
identify the conditions under which specific feedback is more valuable than holistic
feedback for both the reviewers and the authors who receive the review.

4.4. Calculating the peer mark

Many of the systems use a simple mean value, although a variety of other methods of
calculating the peer mark are employed.

The peerScholar has a fixed workflow design in which each artefact is reviewed by
five different authors. An average of the middle three values was used to calculate the
final mark. This reduces the impact of a single rogue reviewer from the calculation.

Aropä uses an iterative weighting algorithm (Hamer, Ma & Kwong, 2005) to
calculate the grade. This algorithm is designed to eliminate the effects of rogue
reviewers. The more that a reviewer deviates from the weighted average, the less their
review contributes to the average in the next iteration. When the weighted averages
have settled, the algorithm halts and the values are assigned as grades.

CPR requires students to go through a training stage where the grades assigned
by students are compared with the expected grades. The students receive feedback on
their grading performance and must be able to accurately apply the criteria before
they are permitted to begin reviewing work submitted by their peers. The degree to
which a reviewer agrees with the 'ideal' review set by the instructor determines a
'reviewer competency index' which is later used to weight the reviews when a
weighted average is calculated.

SWoRD calculates a weighted grade based on three accuracy measures:
 system differences, consistency and spread. The system assumes that the average
of all the reviewers of a given artefact is an accurate measure. The 'systematic' metric
determines the degree to which a given reviewer is overly generous or overly harsh (a
variation of a t-test between the reviewer and the average marks across all the
reviews). The 'consistency' metric determines the correlation between the reviewer
marks and the average marks (i.e can the reviewer distinguish between good and
poor papers). Finally, the 'spread' metric determines the degree to which the reviewer
allocates marks too narrowly or too widely. These metrics are combined to form an
accuracy measure which is factored into the weighting for reviewer marks.

CAP initially used a median value (Davies, 2000) to eliminate the effect of 'off the
wall' reviewers, but was subsequently modified to calculate a compensated peer mark
(Davies, 2004). The compensated peer mark is a weighted mark that takes into
account whether a given reviewer typically overestimates the grade, or under-
estimates the grade (compared to the average given by peers). Although the overall
effects of the compensation are minor, students feel more comfortable knowing that they will not be disadvantaged by a ‘tough’ marker.

4.5. Quality of reviews

The quality of the reviews created by students is of significant concern to both instructors and students. A number of systems offer the opportunity to provide feedback to the reviewer about the quality of their reviews. However, there are few studies that have investigated the quality of the reviews, the value of the feedback to the students, or investigated how the rubric format, or quality assurance methods have affected the quality of the feedback.

4.5.1. Validity of reviews

One aspect of quality is the ability of peers to mark fairly and consistently. The metric most commonly used to determine if students can mark effectively is the correlation with the marks assigned by an instructor.

Falchikov and Goldfinch (2000) conducted a meta-analysis comparing peer marks with teacher assigned marks. They found a mean correlation of 0.69 between teacher and peer marks over all the studies they considered. Paré and Joordens (2008) found a small but significant difference between expert and peer marks in psychology courses using the peerScholar system. The correlation between the expert and peer marks was low, but increased after they introduced the facility for students to grade the reviews they received. They conclude that the averaged peer marks are similar to the averaged expert marks in terms of level and ranking of assignments. Sitthiworachart and Joy (2008) conducted a study that compared tutors’ and peers’ marks for a number of detailed marking criteria for assignments in a first-year programming course. They found high correlations between tutors’ and students’ marks for objective criteria, but lower correlations between tutor and student marks for subjective criteria.

Wolfe (2004) reports that sufficiently large numbers of reviews result in reliable averages, although this was an anecdotal observation by the author rather than the result of a formal study. It is worth noting that the system used by Wolfe resulted in a larger number of reviews being contributed by the better students than the poorer students.

4.5.2. Quality of formative feedback

There are few studies that have investigated the nature of formative feedback provided by students in holistic comments, and compared the value of those comments with those provided by instructors. A study using SWoRD revealed that formative feedback from multiple peer reviews was more useful for improving a draft than feedback from a single expert.

4.5.3. Backwards feedback

The term backwards feedback is used to describe the feedback that an author provides to a reviewer about the quality of the review. This feedback can be formative, in the form of a comment, or can be summative in the form of a numeric value.
Ballantyne et al. (2002) suggest that teachers award marks for the feedback provided by peers to boost student engagement and commitment to the task. The system created by Wolfe (2004) did not contain any assessment of review quality, and he estimates that ~70% of the reviews were superficial. Many of the more recently developed tools require students to assess the quality of the reviews they have received, either summatively or with formative feedback.

The ‘tit for tat’ approach used in Praktomat allocates reviews on a paired basis where possible, so a reviewer knows that they are reviewing the work of the person that will be reviewing them in turn. This encourages students to produce high quality reviews in the hope that the recipient will be doing the same. Although this is a feasible strategy for formative assessment, it is not appropriate for summative assessment where it would be likely to encourage grade inflation.

Kali and Ronen (2005) decided not to grade assignments on the basis of the peer assessments, but instead to grade the quality of the reviews. They report that grading students on the quality of the reviews rather than the peer assessed marks for their assignments reduced tensions and produced higher correlations between the marks assigned by students and instructors. This grading was performed by instructors.

PEARS allows authors to respond to their reviewers, giving feedback on the usefulness of the reviews they received. However, this feedback is purely formative and is not used in assessment criteria.

SWoRD requires authors to provide feedback to the reviewers about the quality and usefulness of the review. This feedback is purely formative and plays no part in the final grade.

Aropa can be configured to require students to formally review a number of reviews using an instructor-defined rubric. The instructor can specify that students assess the reviews they have received, or the reviews can be considered to be artefacts in their own right and allocated anonymously and randomly to be reviewed by a student that has no vested interest.

4.6. Dialogue

The systems considered here vary substantially when it comes to supporting discussion within the peer assessment framework. PeerGrader allows authors and reviewers to access and contribute to a shared web page where discussions can occur. The instructor can configure the system to make the comments posted to the shared web page visible to the other students allocated to review the same author’s work. This allows either private discussions between authors and reviewers, or a group discussion between the author and all the reviewers of their work. Web-SPA uses a similar approach in which students can post short messages to a public page.

OPAS includes a discussion forum which is available for students to post after the completion of the review. This encourages reflection on the criteria and quality of the work produced. The highest degree of discussion is provided by the Sitthiworachart system, which provides reviewers with the capacity to communicate with both the author and all the other reviewers assigned to review the given assignment. A chat system allows them to communicate in real time, or leave messages for each other if they are not available.
4.7. Workflow

SWoRD is designed for students to progressively improve essay drafts using formative peer feedback. It uses a fixed process for a given cycle, but the instructor can define the number of review cycles that occurs before the final submission.

PeerGrader allows the author to revise their work at any time through the reviewing process. When the author submits a revised version, an email message is sent to all the reviewers. The older version is archived and a new discussion page is created for the revised version. The collaboration scripts used by OPAS support multiple review cycles where students can progressively improve drafts on the basis of feedback.

Miao and Koper (2007) show how collaboration scripts can be used to describe the structure of interactions that occur in the process of peer review. Using a script to describe the peer assessment process, a tool can automatically generate documents adhering to the IMS Learning Design specification (IMS LD) and IMS Question and Test Interoperability specification (IMS QTI) that can be viewed using an appropriate player. However, the authoring tools used to create the scripts are complex and require a significant degree of technical expertise.

CeLS is extremely flexible and allows instructors to create a wide range of peer assessment activities with varying workflow. The authors report the flexibility resulted in a large number of variants of basic structures which could be confusing. The authors suggest that further work is required to categorize the structures to ensure that the variety of options is not overwhelming.

There appears to be a significant trade-off between flexibility and ease-of-use. Systems that have more flexible workflow have used collaboration scripts or domain specific languages to express the complex processes, but this flexibility makes them too difficult to use for a non-technical person.

5. Conclusion

This review makes a significant contribution by summarising the available tools that support online peer assessment. These tools have been classified as generic, domain specific and context specific. The major features have been compared and discussed. Although a variety of different tools have been reported in the literature, few of them have been thoroughly evaluated. There is a clear need for more usability studies and further evaluation studies that investigate the differences between the approaches taken.

Aropañ, SWoRD and C.A.P. have the most sophisticated processes for identifying ‘good’ reviewers and weighting student assigned grades accordingly. A variety of different algorithms are applied to weight the peer marks in an attempt to establish a more accurate measure of the ‘true’ quality of an assignment. Comparative studies that investigate the benefits of these different approaches are required.

Since the peer assessment process uses the output from one student as the input to another student, online tools need to provide a mechanism to deal with late or missing submissions. Many of the systems support both manual and automatic allocation of reviews, but PRAISE is the only system that dynamically allocates reviews during the submission process. Some systems, such as PeerWise and that of Wolfe, do not limit the number of reviews that a student can perform. In such systems, students with higher grades tend to contribute more than weaker students, resulting in a greater
amount of higher quality feedback being produced. This approach looks promising, and future tools should support unlimited reviewing where possible, although further research is required to investigate this approach more carefully.

All of the systems considered in this study are web-based and use standard web forms for the entry of the review. Only one of the systems (Praktomat) supports direct annotation on the product of review, something that has always been possible on paper-based reviews. None of the tools currently support the use of digital ink to provide annotations during the peer review process.

Although some tools supported instructor designed marking criteria, others specified a fixed schedule. The marking criteria varied between binary criteria and a holistic overall rating and open-ended text. There is no clear indication of the impact of each approach. Future work is required to evaluate the effectiveness of different forms of rubrics for both the reviewer and the recipient of the review. Although numerous studies have considered the correlation between instructor-assigned grades and student-assigned grades, no studies have thoroughly investigated the quality of the formative feedback (comments) provided by students.

Many of the tools support some form of feedback between reviewer and author, but few support full discussion. The impact of discussion at different stages of the peer assessment process has not been investigated. The support of discussion between reviewers and between reviewers and authors warrants further study.

Instructors in computer science have the expertise to develop online tools that support peer assessment, and the opportunity to evaluate those tools in the classroom. The majority of online tools described in this article (13 of 18) have been used in computer science courses, but most are unavailable for use outside the context in which they were developed, and none of them have been widely adopted. It is likely that peer assessment tools in the immediate future will continue to be developed by computer science educators for use in their own classrooms, informed by reports of the current tools. However, it would contribute significantly to the computer science community if future peer assessment tools were designed for use in multiple institutions.

References


