Probing the Landscape: Toward a Systematic Taxonomy of Online Peer Assessment Systems in Education

Dymytro Babik
Edward F. Gehringer
Jennifer Kidd
Old Dominion University, jkidd@odu.edu
Ferry Pramudianto
David Tinapple

Follow this and additional works at: https://digitalcommons.odu.edu/teachinglearning_fac_pubs
Part of the Computer Sciences Commons, and the Education Commons

Repository Citation
Babik, Dymytro; Gehringer, Edward F.; Kidd, Jennifer; Pramudianto, Ferry; and Tinapple, David, "Probing the Landscape: Toward a Systematic Taxonomy of Online Peer Assessment Systems in Education" (2016). Teaching & Learning Faculty Publications. 22. https://digitalcommons.odu.edu/teachinglearning_fac_pubs/22

Original Publication Citation

This Conference Paper is brought to you for free and open access by the Teaching & Learning at ODU Digital Commons. It has been accepted for inclusion in Teaching & Learning Faculty Publications by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.
ABSTRACT
We present the research framework for a taxonomy of online educational peer-assessment systems. This framework enables researchers in technology-supported peer assessment to understand the current landscape of technologies supporting student peer review and assessment, specifically, its affordances and constraints. The framework helps identify the major themes in existing and potential research and formulate an agenda for future studies. It also informs educators and system design practitioners about use cases and design options.

Keywords
Peer assessment, peer review, system design, rubric, scale

1. INTRODUCTION
In the twenty years that the web has been widely used in education, dozens, if not hundreds, of online peer assessment systems have appeared. They have been conceived by educators in many disciplines, such as English, Computer Science, and Design, to name a few. Topping [29] highlighted computer-aided peer assessment as an important pedagogical approach to developing higher level competencies. Surprisingly, most of these systems have been designed “from the ground up” — until now, there is little evidence that designers and developers of one system have consulted other systems to see what existing techniques are appropriate to their experience, and what can be done better. Several authors have conducted reviews of existing peer assessment approaches [4, 5, 8, 11, 19, 25, 28]. To the best of our knowledge, however, no one has proposed a systematic research framework for exploring and generalizing affordances and constraints of educational technology-enabled peer assessment systems.

Our Peerlogic project1 is pursuing two primary goals: (1) to systematically explore the domain of technology-enabled peer assessment systems, and (2) to develop an arsenal of web services for a wide range of applications in such systems. We have examined a number of these systems, including such better known ones as Calibrated Peer Review [20], CritViz [26], CrowdGrader [6], Expertiza [9], Mobius SLIP [2], Peerceptiv [5] and peerScholar [14]. We adopt the term “online peer assessment system” to describe the broad range of computer applications purposefully designed and developed to support student peer review and assessment. Specifically, we define an online peer-assessment system as a web-based application that facilitates peer assessment process workflow, such as collecting submission artifacts, allocating reviewers to critique and/or evaluate designated artifacts submitted by peers, setting deadlines, and guiding reviewers on the format of the qualitative and quantitative feedback. This term covers a class of systems described in the literature as “computer (technology, IT, CIT, ICT, network, internet, web, cloud)-aided (assisted, based, enabled, mediated, supported)” peer assessment (review, evaluation) systems (in any combination). Online peer-assessment systems are a subset of a general class of social computing systems that involve peer review (including social networking and social-media applications, such as wikis, blogs, and discussion forums), but are distinguished by having specific workflow constraints and being directed at specific educational goals.

The purpose of this paper is to set up a framework for the systematic review and analysis of the current state of online peer assessment systems. We contrast our study with the earlier surveys by Luxton-Reilly [19] and Søndergaard and Mulder [25], which considered the facilities of individual systems one by one and then contrasted them. Our approach is to discuss functionalities of systems, and then describe how individual systems realize those functionalities. Thus, in a sense, it is a dual of the earlier papers. Alternatively, one might say it applies the jigsaw technique [34] to them. Because of space limitations, this paper only begins to apply the taxonomy, which we will elaborate and extend in a future paper.

We use our framework to examine affordances and limitations of the systems that have been developed since 2005 and how they
address pedagogical, philosophical, and technological decisions. We also exploit the framework to develop a research agenda to guide future studies. In this paper, we will begin to address these important research questions: What is the current state of the online peer assessment in education? How is technology transforming and advancing student peer review?

We address this study to several audiences such as peer assessment researchers, practitioners, system designers and educational technologists. Researchers in learning analytics can learn what peer-assessment data can be extracted and mined. Software designers can learn what has been designed and implemented in the past. Instructors applying peer review in their classes can find what systems and functionality would best meet their needs. Instructors may turn to ed-tech specialists and instructional designers to answer these questions; thus, the latter also constitute an audience for this work. Conversely, marketers of these systems may identify the unique features of their systems so they can inform their constituencies.

2. FRAMEWORK AND METHODOLOGY

2.1 Framework

We applied a grounded theory approach to construct our framework. First, we identified all possible use cases, occurring in the online peer assessment. For this, we used an informal focus group, where faculty using peer assessment in their pedagogy described various situations and scenarios. In addition, academic papers on peer assessment were reviewed and relevant practices were brought to the discussion. Through this discussion of practices, the peer assessment process use cases were identified and categorized. Next, these use case categories were formalized as objectives of the peer assessment process. Thus, we obtained a classification of system-independent peer assessment objectives and respective use cases that support these objectives (Table 1).

Table 1. Primary objectives for online peer assessment systems

<table>
<thead>
<tr>
<th>Objective</th>
<th>Descriptive Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Eliciting evaluation</td>
<td>How do student reviewers input evaluation data (quantitative and qualitative, structured and semi-structured)? What input controls are used to elicit responses?</td>
</tr>
<tr>
<td>II. Assessing achievement and generating learning analytics</td>
<td>How are peer assessment results computed and presented to instructors and to students? What assessment metrics can be used?</td>
</tr>
<tr>
<td>III. Structuring automated peer assessment workflow</td>
<td>What is the process of online peer review? What variations of this process exist?</td>
</tr>
<tr>
<td>IV. Reducing or controlling for evaluation biases</td>
<td>How assessment subjectivity can be reduced or controlled for? What metrics of assessment inaccuracy can be used?</td>
</tr>
<tr>
<td>V. Changing social atmosphere of the learning community</td>
<td>How online peer assessment can be conducted to achieve higher-level learning and other benefits?</td>
</tr>
</tbody>
</table>

These objectives and use cases are system-independent because they are not determined by the system in which they are realized but rather by the user needs independent of any system. In this paper, for illustration purposes, we focus only on objective I (Table 2).

Next, we examined a sample set of online peer assessment systems to identify how these use cases are implemented as functionality (features). In this study, we focus on functionality relevant specifically to the student peer-to-peer interactions in the review and assessment process and ignore complementary functionality that is germane to any learning, knowledge management or communication systems (such as learning-object content management). A given use case may be implemented in various systems as different ensembles of features, with varying design options. Therefore, functionality and design options are system-dependent. For each functionality, specific design options were identified and categorized.

Visually, our framework can be represented as hierarchically organized layers, where the top layer comprises objectives, which determine use cases, supported by functionality, implemented as specific design options (Figure 1).

![Figure 1. Research framework for a taxonomy of online peer assessment systems.](image)

2.2 Data Collection and Analysis

Data collection was conducted through iterative paper presentations, system demonstrations, and discussions, documented as written notes and video recordings (including screencasts) shared online. Over three years, the authors have reviewed and experimented with multiple available systems, designed and implemented their own systems, systematically reviewed literature, and collaborated with other creators and users of systems in research and practice.

Identified, categorized and formalized themes, patterns, use cases, and design choices led to the construction of the framework. Then we used this framework to design questionnaires for surveys and structured interviews to collect additional data on each identified system. Collected data was synthesized in a spreadsheet, with formally defined “cases” and “variables”. Our current sample includes 40 systems described in the literature and found on the web. For the purpose of this paper, we illustrate our analysis with a subsample of selected systems (Figure 2). Finally, the multi-case method will be used to complete our taxonomy and to answer our research questions in the full study.

3. SAMPLE ANALYSIS

To demonstrate the application of our research framework for the analysis of the online peer assessment systems in education, in this paper, we focus on Objective I, “Eliciting evaluation”. We analyze the input mechanisms and controls that students use to conduct peer assessment. In general, the review process involves...
two tasks: (a) providing quantitative evaluations based on some criterion or criteria and using some scale, and (b) providing qualitative critiques or comments to peers’ artifacts. Therefore, this objective is manifested in two distinct use cases: (I) “Eliciting quantitative peer evaluation” and (II) “Eliciting qualitative peer evaluation, critiquing and commenting”. Use case I is supported by two functionalities: rubrics and scales used for quantitative assessment; use case II is also supported by two functionalities: critique artifact media types and contextualization of critiques (Table 2). We present below the taxonomy of specific design choices available for these functionalities and illustrate them with examples of specific systems.

Table 2. The application of the research framework for the analysis of objective I

<table>
<thead>
<tr>
<th>System-independent</th>
<th>System-dependent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective</td>
<td>Use case</td>
</tr>
<tr>
<td>I. Evaluation</td>
<td>Use case I:</td>
</tr>
<tr>
<td>elicitation</td>
<td>Eliciting</td>
</tr>
<tr>
<td></td>
<td>quantitative</td>
</tr>
<tr>
<td></td>
<td>peer evaluation</td>
</tr>
<tr>
<td></td>
<td>Rubrics</td>
</tr>
<tr>
<td></td>
<td>Specific /</td>
</tr>
<tr>
<td></td>
<td>analytic</td>
</tr>
<tr>
<td></td>
<td>Scales</td>
</tr>
<tr>
<td></td>
<td>Ranking</td>
</tr>
<tr>
<td>Use case II:</td>
<td>Critique</td>
</tr>
<tr>
<td></td>
<td>artifact media</td>
</tr>
<tr>
<td></td>
<td>types</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Contextuali</td>
</tr>
<tr>
<td></td>
<td>Contextuali</td>
</tr>
<tr>
<td></td>
<td>Contextuali</td>
</tr>
</tbody>
</table>

3.1 Eliciting Quantitative Peer Evaluation

3.1.1 Rubrics

Rubrics are used at all levels of education to evaluate a wide variety of products. A rubric is an assessment tool that communicates expectations for an assignment submission. A well-designed rubric must consist of three essential components: evaluation criteria, quality level definitions, and a scoring strategy [20]. Evaluation criteria are the factors deemed to be important on which the goodness of the submission will be judged. Quality-level definitions specify achievement levels (e.g., “meets standards”, “needs improvement”) and help assessors understand what evidences those levels. The scoring strategy translates reviewer judgments into usable, often numeric, representations.

Rubrics can be categorized as holistic or specific/analytical [13, 15]. In a holistic rubric, a submission is judged as a whole, with a single value or category representing its overall quality. In contrast, a specific/analytic rubric requires evaluations on several distinct criteria.

In the context of peer review, we found that the term “rubric” has been used more loosely to describe a multitude of evaluative processes and structures. Some systems offer wide flexibility in design of rubrics that may or may not contain all three elements, while other systems are more restrictive. This leaves to the instructor assessment decisions, such as the type of rubric, the number of criteria, the number of achievement levels, the point value for each level, whether to use definitions, numeric scales, or both to delineate achievement levels. For example, in Canvas and Expertiza, a rubric can vary from a series of open-ended questions with no established quality levels or quantitative scores to an elaborate rubric with multiple criteria, detailed definitions, and a complex scoring strategy. In CritViz, a rubric is a set of questions that reviewers have to consider when evaluating peers’ submissions. Mobius SLIP supports creation of a qualitative rubric complete with the essential components but elicits holistic quantitative evaluation (Figure 2).

Typically, online peer review systems, e.g., Expertiza, Calibrated Peer Review, Peerceptiv, and Canvas, support specific/analytical rubrics because they generate more detailed feedback that helps students understand their performance on each of these criteria. Specific rubrics provide a more granular picture of artifacts’ strengths and weaknesses and more guidance to students as they complete subsequent revisions or assignments. Some systems, such as Mobius SLIP and CritViz favor holistic evaluations (even if some specific rubrics are provided); noticeably, these systems also rely in ranking (rather than rating) evaluations. Holistic rubrics make more sense for overall ranking, as it may be tedious for evaluators to rank multiple products on each of several criteria.

Limited choices in rubric design reduce the instructor’s control over pedagogical implications of using different rubric types, but free them to focus on other aspects of instruction. Instructors new to assessment may appreciate not having to make too many of these decisions. Some systems fall in the middle, dictating some parameters, but allowing flexibility with others. For example, Peerceptiv allows instructors to determine the number of criteria, but requires each criterion to have a 7-point scale, unaccompanied by elaborated definitions. If rubric design is a critical factor in the institution’s use of peer review process, instructors must carefully vet and select the system which best fits their assignment and assessment requirements.

In the context of peer review, rubrics are also associated with higher student achievement [18] and higher reliability of peer evaluations [12, 30]. Several studies suggested students need to engage with the rubrics in order for them to be effective [20]. Providing rubrics when an assignment is first given and asking students to complete self- and peer reviews were shown to be effective ways to facilitate this engagement.
Figure 2. Screenshot of selected online peer assessment systems
While rubrics are typically viewed as an assessment tool, many researchers suggested that they have a second, often overlooked, instructional purpose. When used formatively, rubrics can illuminate strengths and weaknesses and suggest a direction for future improvements. Rubrics help students understand what to change in their work and help educators see where future instruction should be directed. Interestingly, studies of student perceptions of rubrics suggested that students value these formative purposes. Students observed that rubrics clarify the objectives for their work, help them plan their approach, check their work, and reflect on feedback from others. They also report producing better submissions, earning higher grades, and feeling less anxious about assignments when they are provided with a rubric [20].

Empirical studies support students’ impressions, providing evidence that rubrics support teaching and learning and contribute to higher achievement [13, 20].

Online peer review systems offer a variety of means for supporting the formative use of rubrics. Some allow different rubrics to be used for different rounds of peer review within a single assignment; others offer calibration to show students how peer evaluations compare to the instructor assessment on a selected sample assignment. Many systems allow student achievement scores to be calculated in different ways depending on whether peer review is used for formative or summative purposes. These features, while important to this discussion, are beyond the purview of this paper, and will be discussed in a future publication.

3.1.2 Scales

In general, quantitative evaluations may be conducted using either ranking or rating [9]. Rating refers to the comparison of different items using a common absolute, or cardinal, scale (either numeric or categorical). Ranking, sometimes also called forced-distribution rating, means comparing different items directly one to another on a relative, or ordinal, scale [22]. Both ranking and rating have their strengths and weaknesses, and there is still little consensus as to which has a greater predictive validity [1, 16, 17].

Generally, ranking and rating are expected to correlate, but some studies have demonstrated that ordinal (ranking-based) evaluations contain significantly less noise than cardinal (rating-based) evaluations [23, 32]. A cardinal scale in the context of peer evaluations is also susceptible to score inflation, whereas an ordinal scale is immune to this problem [9]. When a cardinal scale is used, an evaluator may “smokescreen” his preferences by giving all evaluated artifacts the same rating, and may severely inflate scores by giving all artifacts the same high ratings (similarly, he can severely degrade scores by giving all artifacts the same low ratings). Thus, a cardinal scale is very vulnerable to social or personal biases (e.g., “never give the highest rating”) and idiosyncratic shocks (e.g., mood or inconsistency in evaluation style). When an ordinal scale is used, an evaluator must construct an explicit total ordering of artifacts (based on their perceived quality) over others [24]. This makes the evaluation more robust. Psychological evidence suggests that evaluators are better at making comparative judgments than absolute ones [26, 31].

The ordinal scale also has its drawbacks. It forces evaluators to discriminate between artifacts that may be perceived to have very similar quality as much as between the artifacts whose qualities may be far apart. Some ordinal scales may implicitly emphasize items earlier in the list and lead to their higher ranking. Evaluating on ordinal scales places higher cognitive load on the evaluators because it requires them to compare multiple items against each other. Thus, rubrics that use ordinal scales tend to contain fewer criteria, and consequently, they may not draw evaluators’ attention to as many salient features of the artifact under review. Scores from rating-based systems are usually determined by calculating a weighted average of scores given to various criteria, which means they depend on multiple, independent decisions by each evaluator, rather than a single decision of how to rank this submission relative to others.

Most online peer assessment systems are rating-based, e.g., Calibrated Peer Review, Peerceptiv, PeerGrade. Typically, a rating scale is presented as a drop-down menu or validated text box. Ranking-based systems have been also gaining prominence thanks to the strengths of the ordinal evaluation approach. In CritViz, for example, students have to “drag and drop” submission artifacts to position them in the ranking order according to the reviewers perception of their quality. Yet other systems attempt to take advantage of combining both evaluation scales in a single control. For example, in Mobius SLIP, the SLIP Slider control (figure 2) allows recording ratings on the 0-100 scale as well as ranking, which then can be used separately to generate analytics and grading data. Naturally, for such controls to function, they should exclude the possibility of assigning the same rating to any two artifacts, but they allow placing two artifacts close to each other to indicate approximately the same level of quality. Another example of a system that supports both ranking and rating scales is PeerScholar [14], where the instructor can configure an assignment to have either a rating scale or a ranking scale. Inasmuch as long rubrics also seem to elicit more textual feedback, systems that use ranking may provide the author with less feedback on the quality of the submission and guidance how to improve it [35].

3.2 Eliciting Qualitative Peer Evaluation

3.2.1 Critique Artifact Media Types

Critiques, as the verbal component of reviews, can be provided in different formats. The most obvious and typical design choice is to prompt the reviewer to post a plain-text comment in a text box. Most systems provide a web form combining rubric questions and text boxes to fill out. Plain-text feedback is the most basic and arguably the fastest way to provide feedback. Textual critiques can be enhanced by allowing rich-text format (varying font faces and sizes, bullet points, alignment, hyperlinks, etc.) using the WYSIWIG editors. Including a hyperlink in the text feedback further enhances the options by referencing an externally hosted copy of the submission artifact (which can be edited and/or annotated) or by referencing externally hosted multimedia critique artifacts, such as voice and video recordings, screencasts and HTML documents. Only a few systems (e.g., Canvas) allow internal hosting of multimedia critique artifacts, but arguments have been made that this type of critiques substantially improves the provider’s efficiency and the recipient’s experience.

The next step up in providing rich critiques is inline file annotation. Several systems take advantage of the third party APIs allowing inline file annotations of submission artifacts
uploaded as files. For instance, Mobius SLIP and Canvas utilize a document viewer called Crocodoc, which renders various file formats as an HTML document and allows reviewers to select portions of the document and annotate them in place. Annotation includes highlighting, commenting, adding text and primitive graphic elements. This feature is similar to adding comments in a Microsoft Word file or a Google doc. Crocodoc supports both non-anonymous and anonymous file annotation. While the Crocodoc API is used by a number of systems, after its acquisition by Box in 2013, it is expected to be replaced by a new API with similar, and possibly, more advanced inline file annotation functionality. Web annotation is another possible implementation of inline annotation in the web-based online peer assessment systems [33] but no systems in our illustrative sample rely on it; therefore, this option needs to be explored further. To the best of our knowledge, no existing online peer review systems offer its “native,” custom-built inline file annotation functionality.

Since text critiques may not offer the desired expressiveness and clarity that other media may provide, users have been requesting to allow reviewers to attach multimedia files containing critique artifacts (e.g., images, voice or video recordings) as an alternative to inserting URLs to such externally hosted files in the plain- or rich-text comments. Such an option, for example, would allow reviewers, who are more comfortable using traditional media (e.g., pen and paper), to write their critiques offline, then scan them into pdf or image files, and then attach them to the original submission artifacts. For another example, some reviewers may also be more productive when providing their critiques as voice or screencast recordings made directly in the system. In our sample only Canvas offers such options, but since they are available in other social learning applications, such as VoiceThread (voicethread.com), it is reasonable to expect increasing availability of such functionality in online peer assessment in the near future.

3.2.2 Contextualization of Critiques

A number of factors influence how well the author of the submission artifact is able to understand and relate to a reviewer’s feedback: spatial relationship of the critique artifacts with the submission artifact, placing critiques in the specific context of the submission artifact, and the granularity of comments. For example, directly annotating an issue in a fragment of the submission artifact, rather than trying to explain in the overall, “detached”, critique where the issue is located and how to fix it, simplifies communication between the reviewer and the author. Moreover, annotation is more suitable for providing specific fine-grained comments, while filling out a text box is more appropriate for more global comments.

We define this aspect of eliciting qualitative evaluation as the contextualization of critiques. Naturally, the system interface design determines how much critiques can be contextualized in relation to submission artifacts. Moreover, the interface implementation of other functionalities, such as rubrics, scales and critique artifact media types closely interplays with the implementation of critique contextualization. Contextualization of critiques, thus, has two options: (a) “detached”, non-contextualized (“single comment per submission”); (b) contextualized (“multiple comments in various fragments of the submission”). While the former is typically available in all systems in our sample, the latter is implemented as either an entry space (textbox) associated with a specific criterion/question in the rubric (e.g., Expertiza, CritViz) or as inline file annotation with Crocodoc (e.g., Mobius SLIP, Canvas). Further exploration of this functionality and design options for its implementation will be provided in the full study.

4. CONCLUSION

We have presented our initial attempt at formulating the research framework for a taxonomy of educational online peer assessment systems. This framework enables researchers of technology-supported peer assessment to understand the current landscape of technologies supporting student peer review and assessment, specifically, its affordances and constraints. Importantly, this framework helps identify the major research questions in existing and potential research and formulate agenda for the future studies. It also informs educators and system design practitioners about use cases and design options in this particular branch of educational technology.

Using a grounded theory approach, we identified several primary objectives for online peer assessment systems and combined them in the research framework. To illustrate the application of this framework in this research-in-progress paper, we presented a sample analysis of how use cases supporting the objective of eliciting quantitative and qualitative peer evaluations are implemented in several different systems. In the future, full study, we intend to apply the multi-case method to conduct a complete analysis of the objectives based on a large sample of online peer assessment systems.

5. REFERENCES


