Redesigning Educational Peer Review Interactions Using Computer Tools: An Introduction

Ilya M. Goldin, Kevin D. Ashley & Christian D. Schunn

Carnegie Mellon University - University of Pittsburgh, PA | United States

Abstract: Peer review is a family of instructional techniques. Historically, these have been employed in writing and many other educational domains. Modern computer technologies facilitate the use of peer review, which is especially relevant to educational settings where it is not practical to administer peer review manually. The use of computer support for peer review has shed light on many important scientific questions, some of which we summarize. These findings set the context for the papers in this special issue, which demonstrate how computer support for peer review enables research on peer review itself and on its pedagogical significance.

Keywords: peer review, technology-enhanced learning

journal of
WRITING RESEARCH

Goldin, I. M., Ashley K.D., & Schunn, C.D. (2012). Redesigning Educational Peer Review Interactions Using Computer Tools: An Introduction. *Journal of Writing Research, 4*(2), 111-119. http://dx.doi.org/10.17239/jowr-2012.04.02.1

Contact: Ilya M. Goldin, Carnegie Mellon University, Human-Computer Interaction Institute, 5000 Forbes Avenue, Pittsburgh, PA 15213, PA | USA – goldin@cmu.edu.

1. What is Peer Review in the Educational Context?

We consider as peers those who have a similar academic role in a particular context (e.g., peers in a class, researchers in a field), although they may have different skill levels and prior experience. Review is any process (face-to-face or document-mediated) by which evaluation and feedback is applied to an academic object. And thus, peer review is a review process completed by academic peers. Broadly construed, peer review has a wide range of traditional uses in all academic domains at every level of expertise. Peers can review any academic product in both formative and summative ways, and often there are both formative and summative elements at once. Peer review is especially popular for writing exercises for both instructional and pragmatic reasons, and thus an important topic for the Journal of Writing Research.

Given the diversity of contexts in which peer review can happen, it should not be surprising that there are many variations on peer review, even just within peer review of writing. To name just a few variations: the number of rounds of peer review, the number of reviewers, the instructions given to reviewers, the tools used to support peer review, the relative topic expertise of reviewers relative to authors, the formality or structure used to guide review, the time given to peer review, the diversity of knowledge in the reviewers for a given paper, the anonymity of writer or reviewer, and the use of synchronous or asynchronous communication between reviewer and author. Additionally, there are distinctions among peer review and similar (or similar-sounding) techniques, including peer tutoring (Bruffee, 1984), peer response, peer editing, peer evaluation, and peer criticism or critique, all of which may take place in writing-oriented courses (Armstrong & Paulson, 2008).

One taxonomy of peer review (Gielen et al., 2010) delineates five clusters of issues:

- decisions concerning the use of peer assessment
- link between peer assessment and other elements in the learning environment
- interaction between peers
- composition of assessment groups
- management of the assessment procedure

Each issue in these clusters could further be a binary opposition, a multidimensional concept, or a continuum, making for a potentially infinite set of peer review configurations. Adding complexity, what may not be obvious is that these clusters interconnect. For instance, decisions on the use of peer assessment may bear on the composition of assessment groups, e.g., because peer groups consisting only of novices may lead to poor peer-learning outcomes. Presumably these variations change the nature of what feedback is provided and what impact that feedback has. Thus, in any given setting, a peer review process ought to be tailored to the desired outcome.

Peer review research has a long and interdisciplinary history. For instance, issues relating to peer review were noted in studies of writing instruction by Kenneth Bruffee, who investigated the potential for students to learn from each other rather than from the instructor (Bruffee, 1984). Bruffee traces the development peer tutoring to studies of collaborative learning at a British medical school in the 1960s. The writing instructor Peter Elbow explored various aspects of feedback on writing, such as what feedback is valuable at what point for a work in progress, and how students can benefit from sharing their work with peers even if they receive no feedback (Elbow, 2000). The potential for peer review is captured in the meta-analysis of Hillocks, which found that the environmental mode of instruction, "characterized by peer-group activity… [that] involves highly structured problem-solving tasks" was more effective than other modes, including individual teacher-student conferences (Hillocks Jr, 1986, p. 199). While these early works exemplify some of the activities that might be termed peer review, papers in this Special Issue illustrate how modern computer tools allow researchers to study peer review methodically and empirically.

A critical distinction in possible goals of peer review of writing is textual improvements to the document versus improvement in the skills of the participants of the peer review exchange. In professional settings, the primary goal is the document itself: the grant, conference paper, journal article are all being evaluated *per se*, and the purpose of feedback is to improve the document. In typical educational settings, the primary goal is the skills set of the participants, with only a secondary goal of improving the document being evaluated. This critical distinction likely implies that structures for effective professional peer review may not be optimal for effective educational peer review. We focus on the case of educational peer review.

2. Computer-Supported Peer Review in Education

At many different points in time, technological change has had a large effect on writing (e.g., the invention of paper, the printing press, and the typewriter). Computers are now a ubiquitous component of writing. Similarly, technology has had a large influence on peer review of writing at professional levels. Today, professionals often review writing through the lens of general computer tools applied for purposes of peer review. For example, documents arrive via the Internet, they may be viewed via general tools such as Adobe Acrobat or Microsoft Word, and then feedback returns to the author via the Internet once again.

Peer review of writing in education can also make use of such general computer tools. Students can simply email around documents for peer commentary. Feedback can be transmitted as changes to the original document (e.g., PDF mark-up), or via synchronous or asynchronous side channels (e.g., email or video conferencing). Using general tools, the computers bring to peer review of writing the same advantages and disadvantages that they bring to many other kinds of tasks: formal externalization of (review) objects that face-to-face review would leave implicit, connecting of individuals across time and space, and reducing the communication bandwidth such that more miscommunications might occur (Hinds & Kiesler, 2002).

The advantage of using generally available tools is that students can continue such peer reviewing practices in many other academic and professional contexts. But special-purpose tools can bring potential advantages both for research and practice. For example, specialized tools allow researchers to track the interactions of peers in greater depth, to manipulate precise components of the peer interaction, to add supports for reviewers in the peer review process, and to provide instructors with insights into what students are learning (or not) from the peer interactions.

Early forms of computer-supported peer commentary on writing can be found in the 1990s (Neuwirth, Chandhok, Charney, Wojahn, & Kim, 1994; Scardamalia & Bereiter, 1994). As far back as 2000, Peer Grader, which was a specialized application for peer review in education, was used to review student-written research papers, to support students in compiling bibliographies relevant to class lectures, to annotate lecture notes, to make up original problems, to review other students' designs, and to do weekly reviews in independent-study courses (Gehringer, 2000). Further, specialized tools enable peer review in instructional settings where its manual implementation would be practically impossible, e.g., courses with hundreds of students. This is achieved through automation of key processes, such as collection of student assignments, distribution of these to peers for review, collection of reviews with regard to a rubric, delivery of this structured feedback to peer authors, blinding reviewers and authors for anonymous communication, assignment of reviewers to authors, and back-evaluation of reviews from authors to reviewers.

Some milestone systems in the history of computer-supported peer review are PREP Editor (Neuwirth et al., 1994), CSILE (Scardamalia & Bereiter, 1994), Praktomat (Zeller, 2000), Peer Grader / Expertiza (Gehringer, 2000), SWoRD (Cho & Schunn, 2007), and Aropa (Hamer, Kell, & Spence, 2007).

Some key research findings are as follows. In general, preventing authors from knowing reviewer identity increases the number of critical reviewer comments and improves writing performance on a transfer task (Lu & Bol, 2007). Domain-specific educational practices may be exploited to integrate peer review for maximal pedagogical usefulness; for example, the choice of peer review rubrics and commenting prompts shapes what is done and what is learned (Wooley, Was, Schunn, & Dalton, 2008). Further, peer assessment, self-assessment, and collaborative assessment may be combined within a single system to enrich the space of instructional activities (Gouli, 2006).

Reviews, whether numeric or textual, may be evaluated automatically with machine learning techniques, which can serve as a basis for formative or summative assessment (Cho, 2008; Ramachandran & Gehringer, 2010; Xiong, Litman, & Schunn, 2010). Criteria-based self-assessment (Li & Kay, 2005) can be used to generate a “scrutable” student model (Weber & Brusilovsky, 2001), i.e., one that a student can examine and modify. Rather than assigning students to review works selected randomly or letting students choose works to review, peer works may be assigned or recommended to individual reviewers based on characteristics of the reviewer, author, and the work itself (Crespo García, Pardo, & Delgado Kloos, 2006; Masters, Madhyastha, & Shakouri, 2008), but the literature on effective group composition is not definitive, cf. (Hsiao & Brusilovsky, 2008; Webb, Nemer, & Zuniga, 2002). Students who peer review papers that score low in terms of peer assessment may produce better second drafts than students who peer review high-scoring papers (Cho, Schunn, & Kwon, 2007).

The finding that summative peer assessment is very similar to summative assessment by an instructor has been noted multiple times. Combining the opinions of multiple reviewers for each essay provides a more reliable estimate of the quality of the essay than a single reviewer's opinion; for example, if the correlation of reviewer and instructor scores is 0.6, an effective reliability of combined reviewer scores of 0.9 requires about 6 reviewers (Cho & Schunn, 2007). By calibrating reviewers before reviewing begins, it is possible to ensure a minimum reviewer accuracy (Russell, 2004). Reviewers may be evaluated via the numeric ratings they produce, e.g., in terms of metrics such as systematic difference, consistency, and spread (Cho & Schunn, 2007; Goldin, 2012). Taking reviewer differences into account may help in computing summative assessments of the works of authors (Hamer, Ma, & Kwong, 2005; Lauw, Lim, & Wang, 2007) as well as the works of reviewers (Goldin, 2012), and these evaluations may be computed at the same time as the quality of the peer author works under review. Evaluations of reviewers may also be used to grade reviewer effort (Gehringer, 2000), and communicated to the reviewers to help them monitor and improve their performance, privately or to the whole class as public praise of good performance (Gehringer, Gummadi, Kadanjoth, & Andrés, 2010).

As with any assessment technique, validity and reliability are key issues of interest with regard to peer review. If validity of summative peer assessment is defined as the convergence of peer assessment to instructor assessment, the instructor may have a different view of the validity of an exercise than the individual student. This is because the instructor's impression of validity is a kind of average that takes into account all the papers in the class, while an individual student author's impression depends on whether the peer ratings received by that author deviate from the instructor's grade (Cho, Schunn, & Wilson, 2006). The general tension between validity and reliability has been noted in peer assessment: peer review may demonstrate a "convergence of different raters on a 'single truth'", or it may "uncover the presence of multiple perspectives about the performance being assessed, which do not necessarily have to agree" (Miller, 2003).

3. How Papers in this Special Issue Contribute to the Field

The papers in this Special Issue represent multiple ways of using computer tools to redesign and analyze peer review interactions. The technology described in Crinon's paper is the simplest (Crinon, 2012). Crinon used distance learning with email to conduct peer review across four geographically separated classrooms of 4th and 5th graders. Over the course of months, the students engaged in activities aimed at writing and revising an episode from a novel. In the process, students engaged in a two-week peer review exchange conducted exclusively via email. Students from two classes served as reviewers; students from the other two received reviews and provided back reviews. Crinon then compared the authors' episode-rewrites with those of the reviewers, and related all of them to the written exchanges between reviewer and author. Although the computer support was simple email, the Internet medium enabled the researcher to assemble enough participants for a meaningful experiment, ensure that each author received multiple reviews, balance the numbers of high- and low-achieving students across groups, and conduct the peer-reviewing exchange over a geographic distance and an extended time period without overly disrupting classroom routines. The potential is apparent for linking dispersed classrooms in peer-review activities and thus assembling a more diverse body of authors and readers or for adding a pedagogically meaningful collaborative activity to distance learning. In addition, the Internet medium produced a written electronic record of the entire exchange between author and reviewer. More automated peer review tools might have made the research effort easier, but perhaps also complicated the manipulation; many peer review tools force equal assignment of papers for review to all contributing authors.

While Crinon *manually* analyzed, related, and compared the texts of the chapter rewrites and the reviews, the electronic medium lends itself to more automatic analyses, which can then be applied to even large-scale studies which could investigate situational moderators (e.g., under which circumstances do the observed effects hold true?). The papers of Xiong, et al. and of Leijen and Leontjeva investigate automated analysis of written peer comments (Leijen & Leontjeva, 2012; Xiong, Litman, & Schunn, 2012). Xiong, et al. applied computerized techniques (Natural Language Processing and Machine Learning) to analyze peer reviewers' feedback according to two types of features previously shown to help authors improve their texts, namely whether the feedback localizes a problem in the author's text and offers a concrete solution. Techniques like these can help writing researchers to discover and investigate other features of reviewers' advice and authors' texts, modifications, and backreviews (comments from authors back to reviewers) that may relate to student learning to write better. In addition, Xiong et al. illustrate how the techniques can be applied not only after the fact as part of research activities but also immediately as reviewers prepare and submit their reviews. A system could highlight parts of the reviews that include the desirable features or flag their absence in time for reviewers to improve their reviews before authors receive them. Leijen and Leontjeva also used machine learning techniques in analyzing electronic peer-review records. They investigated linguistic and other review features of peer feedback and how these may influence second language learners of English to accept or reject revision advice in their academic writing. The results suggest that directive comments and multiple peer comments on the same topic influence authors' revisions.

Similarly, Goldin and Ashley provide an example of how computer-supported peer review enables research in alternative ways to collect and analyze peer feedback (Goldin & Ashley, 2012). A strength of peer review is that assessments of each paper may be elicited from multiple reviewers, e.g., to increase reliability of the aggregate assessments or to encourage diversity of reviewer perspectives. Fortunately, computer-supported peer review makes it relatively easy to solicit and manage multiple peer reviews of each paper. Goldin and Ashley conducted an experiment comparing two types of rubrics for guiding peers reviewing: one that focuses generally on domain-relevant aspects of writing and another that focuses specifically on conceptual aspects of the problem scenario students are asked to analyze. Statistical analyses of the peer reviewers' rubric-guided feedback scores can answer such questions as whether a rubric is valid, reliable, redundant, and helpful. Such information is relevant not only to writing researchers but also provides instructors with feedback on a rubric's pedagogical effectiveness. Computer support facilitated both the administration of the peer review exercise in a class of sixty students using two different rubrics, and the subsequent data analysis.

In sum, the four papers reported in this special issue represent a range of contributions of peer review computer support in conducting writing research on, and improving the pedagogical effectiveness of, the peer review process. Each of these papers only addresses a very specific issue, but in doing so takes a step towards key questions in peer review research, such as these:

- Which outcomes are achievable via peer review, and which are not? What are the necessary and sufficient conditions to implement each outcome via peer review, and what changes do these outcomes require of peer review?
- How do we assess the outcomes and connect them to the peer review process? What aspects of peer review processes do we instrument, and how should we analyze the resulting data such that we can understand how we are achieving the outcomes of interest, or why we are failing to achieve them?
- How do individual differences among learners, including cognitive and metacognitive differences, interact with peer review effectiveness? What can peer review reveal about cognitive and metacognitive differences among participants?

While much work remains, we hope that the papers in this Special Issue help researchers and practitioners of peer review in pursuing these questions.

References

Armstrong, S. L., & Paulson, E. J. (2008). Whither "Peer Review"? Terminology Matters for the Writing Classroom. *Teaching English in the Two-Year College*, *35*(4), 398–407.

Bruffee, K. (1984). Peer tutoring and the "conversation of mankind." *Writing centers: Theory and administration*, 3–15.

Cho, K. (2008). Machine classification of peer comments in physics. In R. S. J. de Baker, T. Barnes, & J. E. Beck (Eds.), *Proceedings of the 1st International Conference on Educational Data Mining* (pp. 192–196). Presented at the Educational Data Mining, Montreal, Canada.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*, *48*(3), 409–426. doi:10.1016/j.compedu.2005.02.004

Cho, K., Schunn, C. D., & Kwon, K. (2007). Learning writing by reviewing in science. *8th International conference on computer-supported collaborative learning* (pp. 141–143). New

Brunswick, NJ, USA: International Society of the Learning Sciences. Retrieved from http://portal.acm.org/citation.cfm?id=1599600.1599626

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891–901. doi:10.1037/0022-0663.98.4.891

Crespo García, R. M., Pardo, A., & Delgado Kloos, C. (2006). Adaptive Peer Review Based on Student Profiles. In M. Ikeda, K. Ashley, & T.-W. Chan (Eds.), *Intelligent Tutoring Systems*, Lecture Notes in Computer Science (Vol. 4053, pp. 781–783). Springer Berlin / Heidelberg. doi:10.1007/11774303_99

Crinon, J. (2012). The dynamics of writing and peer review at primary school. *Journal of Writing Research*, 4*(2)*.121-154.

Elbow, P. (2000). Everyone can write: Essays toward a hopeful theory of writing and teaching writing. Oxford University Press, USA.

Gehringer, E. F. (2000). Strategies and mechanisms for electronic peer review. *30th Annual Frontiers in Education Conference* (Vol. 1, pp. F1B/2–F1B/7 vol.1). Presented at the 30th Annual Frontiers in Education Conference. doi:10.1109/FIE.2000.897675

Gehringer, E. F., Gummadi, A., Kadanjoth, R., & Andrés, Y. M. (2010). Motivating effective peer review with extra credit and leaderboards. *Proceedings of the 2010 American Society for Engineering Education Annual Conference & Exposition*.

Gielen, S., Dochy, F., & Onghena, P. (2010). An inventory of peer assessment diversity. *Assessment & Evaluation in Higher Education*, *36*(2), 137–155. doi:10.1080/02602930903221444

Goldin, I. M. (2012). Accounting for Peer Reviewer Bias with Bayesian Models. Proceedings of the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems. Chania, Greece.

Goldin, I. M., & Ashley, K. D. (2012). Eliciting formative assessment in peer review. *Journal of Writing Research*, *4*(2), 203-237.

Gouli, E. (2006). Supporting Self- Peer- and Collaborative-Assessment through a Web-based Environment. *World Conference on Educational Multimedia, Hypermedia and Telecommunications* (p. 2192). Presented at the EDMEDIA.

Hamer, J., Kell, C., & Spence, F. (2007). Peer assessment using Aropä. *Proceedings of the 9th Australasian Conference on Computing Education* (Vol. 66, pp. 43–54). Ballarat, Victoria, Australia: Australian Computer Society, Inc.

Hamer, J., Ma, K. T. K., & Kwong, H. H. F. (2005). A method of automatic grade calibration in peer assessment. *Proceedings of the 7th Australasian Conference on Computing Education* (Vol. 42, pp. 67–72). Newcastle, New South Wales, Australia: Australian Computer Society, Inc.

Hillocks Jr, G. (1986). *Research on Written Composition: New Directions for Teaching.* National Council of Teachers of English, 1111 Kenyon Rd., Urbana, IL 61801 (Stock No. 40750, $19.00 member, $24.75 nonmember). Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED265552

Hinds, P. J., & Kiesler, S. (Eds.). (2002). *Distributed Work* (1st ed.). The MIT Press.

Hsiao, I.-H., & Brusilovsky, P. (2008). Modeling peer review in example annotation. *16th International Conference on Computers in Education* (pp. 357–362). Presented at the 16th International Conference on Computers in Education, Taipei, Taiwan.

Lauw, H. W., Lim, E., & Wang, K. (2007). Summarizing review scores of “unequal” reviewers. *Proceedings of the 7th SIAM International Conference on Data Mining* (pp. 539–544).

Leijen, D. A. J., & Leontjeva, A. (2012). Linguistic and review features of peer feedback and their effect on the implementation of changes in academic writing: A corpus based investigation. *Journal of Writing Research*, *4(2),* 178-202.

Li, L., & Kay, J. (2005). Assess: Promoting Learner Reflection in Student Self-Assessment. Workshop on Learner Modelling for Reflection, to Support Learner Control, Metacognition and Improved Communication between Teachers and Learners at 12th International Conference on Artificial Intelligence in Education (pp. 32–41). Amsterdam.

Lu, R., & 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), 100–115.

Masters, J., Madhyastha, T., & Shakouri, A. (2008). ExplaNet: A collaborative learning tool and hybrid recommender system for student-authored explanations. *Journal of Interactive Learning Research*, *19*(1), 51–74.

Miller, P. J. (2003). The effect of scoring criteria specificity on peer and self-assessment. *Assessment & Evaluation in Higher Education*, *28*(4), 383–94. doi:10.1080/0260293032000066218

Neuwirth, C. M., Chandhok, R., Charney, D., Wojahn, P., & Kim, L. (1994). Distributed collaborative writing: a comparison of spoken and written modalities for reviewing and revising documents. *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, CHI '94 (pp. 51–57). New York, NY, USA: ACM. doi:10.1145/191666.191693

Ramachandran, L., & Gehringer, E. F. (2010). Automated Metareviewing. *Proceedings of the Workshop on Computer-Supported Peer Review in Education, ITS 2010*. Pittsburgh, PA.

Russell, A. A. (2004). Calibrated Peer Review: A writing and critical thinking instructional tool. *Invention and Impact: Building Excellence in Undergraduate Science, Technology, Engineering and Mathematics (STEM) Education*. American Association for the Advancement of Science.

Scardamalia, M., & Bereiter, C. (1994). Computer Support for Knowledge-Building Communities. *Journal of the Learning Sciences*, *3*(3), 265–283. doi:10.1207/s15327809jls0303_3

Webb, N. M., Nemer, K. M., & Zuniga, S. (2002). Short circuits or superconductors? Effects of group composition on high-achieving students' science assessment performance. *American Educational Research Journal*, *39*(4), 943.

Weber, G., & Brusilovsky, P. (2001). ELM-ART: An adaptive versatile system for Web-based instruction. *International Journal of Artificial Intelligence in Education*, *12*(4), 351–384.

Wooley, R., Was, C. A., Schunn, C. D., & Dalton, D. W. (2008). The effects of feedback elaboration on the giver of feedback. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *30th Annual Conference of the Cognitive Science Society* (pp. 2375–2380). Presented at the 30th Annual Conference of the Cognitive Science Society, Washington, DC: Cognitive Science Society.

Xiong, W., Litman, D., & Schunn, C. D. (2010). Assessing Reviewers' Performance Based on Mining Problem Localization in Peer-Review Data. In R. S. J. d. Baker, A. Merceron, & P. I. J. Pavlik (Eds.), *3rd International Conference on Educational Data Mining*. Pittsburgh, PA.

Xiong, W., Litman, D., & Schunn, C. D. (2012). Toward improving the quality of peer feedback through Natural Language Processing. *Journal of Writing Research*, *4*(2).155-176.

Zeller, A. (2000). Making students read and review code. *SIGCSE Bull.*, *32*(3), 89–92. doi:10.1145/353519.343090