

Eliciting Informative Feedback in Peer Review: Importance of Problem-Specific Scaffolding

Ilya M. Goldin^{1,2}, Kevin D. Ashley^{1,2,3}

¹ Intelligent Systems Program

² Learning Research & Development Center

³ School of Law

University of Pittsburgh

Pittsburgh, PA

{goldin,ashley}@pitt.edu

Abstract. In a controlled experiment using Comrade, a computer-supported peer review system, student reviewers offered feedback to student authors on their written analyses of a problem scenario. In each condition, reviewers received a different type of rating prompt: domain-related writing composition prompts or problem/issue specific prompts. We found that the reviewers were sensitive to the type of rating prompts they saw and that their ratings of authors' work were less discriminating with respect to writing composition than to problem-specific issues. In other words, when students gave each other feedback regarding domain-relevant writing criteria, their ratings correlated to a much greater extent, suggesting that such ratings are redundant.

Keywords: computer-supported peer review, ill-defined problem-solving

1 Introduction

Computer-supported peer review deserves the attention of ITS researchers as an instructional activity that seems to bring many benefits to both students and educators. [1] For instance, students benefit in that receiving feedback from multiple peers' on the first draft of an assignment can lead them to improve the quality of their second drafts even more than receiving feedback from an expert. [2] Student authors receive an extra channel of feedback in addition to and distinct from assessment by the instructor or self-assessment [3], and, in playing both roles of author and reviewer, students may learn from engaging in an authentic activity in the many professional domains that institutionalize peer review. One advantage to educators is that when students give each other feedback, they free the educator to focus on other tasks (such as providing struggling students with individual attention).

When augmented with AI techniques, computer-supported peer review may provide ITS research with methods for addressing ill-defined problems even in writing-intensive courses. Ill-defined problem-solving presents a test case for Intelligent Tutoring System technology. [4] ITS have been used for problem-solving when a student's answer or solution procedure can be compared against a gold standard, in domains such as geometry and physics. [5, 6] In contrast, ill-defined

problems may have no correct answer, or multiple defensible answers, or no way to define *a priori* what constitutes an acceptable response. So long as an ITS cannot assess a student's answer, it cannot update its representation of what the student does and does not know, the so-called student model. This precludes it from tutoring the student through guidance, feedback and selection of new problems. Writing-intensive courses often focus on problems that are ill-defined. These problems are usually distinguished by a goal that can be perceived only through analysis and refinement, and by allowing multiple acceptable solution paths. Solvers may frame ill-defined problems differently according to their knowledge, beliefs, and attitudes, thereby yielding different representations for the problem in terms of relevant facts and applicable operators. [7] Analyses of ill-defined problems are often in free-form text since they require arguments and justifications for one solution over others. They exist in many domains, and they are central to domains such as law, where practitioners must map statutes and precedents to the facts of new cases, as the law students had to in the peer-reviewed exercise in the experiment described below.

Free-form student answers to ill-defined problems may be difficult for a computer to interpret, but not for a student's peers. In our computer-supported peer-review system, called Comrade, AI techniques are used to aggregate the feedback that peer reviewers give each other into a student model that estimates attainment of learning objectives. Comrade asks reviewers to provide written feedback as well as numeric ratings of peer work. In this paper, we examine an important aspect of the feasibility of Comrade's design, namely the extent to which student peer reviewers are sensitive to different types of rating prompts. As students evaluate each other's written work, these prompts serve as a scaffold, focusing the reviewers on different aspects of the work. In addition, as Comrade compiles a student model based on the students' feedback, it needs to know whether the reviewers' ratings provide useful information.

A variety of peer review systems has been developed in support of teaching in many domains and according to different instructional strategies and demands. [2, 8-13] Some systems, including SWoRD [2], CPR [8], and Comrade, allow the instructor to specify the rubric according to which reviewers evaluate the peer author's work. The designers of SWoRD purposefully focused prompts on three criteria (insight, logic, and style) that could be applied to writing in any domain. For example, a domain-independent rating point from a SWoRD rubric on the logic of the argument was "All but one argument strongly supported or one relatively minor logical flaw in the argument."¹ It is also possible for a rubric to be highly specific to the assignment. In one deployment of CPR, the rubric contained the question "Does the summary state that the study subject was the great tit (*Parus major*) or the Wytham population of birds? AND does the summary further state that the sample size was 1,104 (egg clutches, 654 female moms, or 863 identified clutches?" [15]

Given the variety of possible strategies that can be employed in creating prompts for peer review, and given the fact that prompts influence the experience of both reviewers and authors, it is important to determine whether some kinds of prompts are more valuable than others. For example, it is desirable to avoid prompts that yield redundant information. It is also possible that some prompts can scaffold reviewers in acquiring domain knowledge better than others. In the work described here, we

¹ For an example of a full SWoRD rubric, see the appendix to [14].

compared the effects of two types of rating prompts: prompts that focus on domain-relevant aspects of writing composition versus prompts that focus on issues directly pertaining to the problem and to the substantive issues under analysis. We considered that when an instructor gives a student a rubric to assess another's paper, interacting with this rubric can cause the student to focus on those issues that are made prominent in the rubric. For example, if the rubric looks at domain-independent issues of writing composition, that communicates to the reviewer that the instructor sees various discourse features as important. We then articulated two kinds of prompts which may be particularly useful to the reviewer. Prompts that focus on writing in the domain can communicate to the reviewer the importance of various domain-specific discourse features, while prompts that relate to the subject of the essay under review can emphasize critical elements of the assignment.

In section 2 of this paper, we describe the Comrade system and the two kinds of prompts (domain writing versus problem/issue specific) it delivered to students in a controlled experiment involving peer-review of a take-home midterm examination essay. In section 3, we present empirical evidence that student peer reviewers are sensitive to the two types of rating prompts and that their ratings of each other's work are less discriminating with respect to writing composition than problem-specific issues. As discussed in section 4, this is pedagogically important. When students give each other feedback regarding domain-relevant writing criteria, their ratings correlated to a much greater extent, suggesting that the ratings are redundant. We discuss the significance of these results for the design, implementation, and evaluation of intelligent tutoring systems that employ peer review as a mechanism for teaching skills of ill-defined problem-solving.

2 Methods

Hypotheses. Our first hypothesis is that peer reviewers are sensitive to the difference between prompts that focus them on writing in the domain (from now, "domain-writing prompts") vs. prompts that focus them on details of the assignment (from now, "problem-specific prompts"). This hypothesis is operationally defined in our study as a between-subjects manipulation with two conditions: domain-writing prompts and problem-specific prompts. To test this hypothesis, we first introduce a definition:

Consider that peer review can be seen as a directed graph. Let every student be viewed as a node. When the student acts as a reviewer, there are outbound edges from this student to the peer authors whose work she is reviewing. When the student acts as an author, there are inbound edges to this student from the other students reviewing her work. Thus, the ratings received by a student are that student's inbound ratings, and the ratings given by a student are that student's outbound ratings.

If peer reviewers are not sensitive to variations in rating prompts, then peer authors' inbound ratings according to different prompts will be highly correlated; if reviewers are sensitive, peer authors' inbound ratings will not be highly correlated.

Our second hypothesis is that when a rubric supports a reviewer in evaluating another student's work, the rubric may act as a scaffold in focusing the reviewer on

key domain concepts, thus making it more likely that the reviewer will understand these concepts. We compared student understanding of key domain concepts before and after reviewing in terms of performance on an objective test, as described below.

Participants. All 58 participants were second or third year students at a major US law school, enrolled in a course on Intellectual Property law. Students were required to take a midterm examination and to participate in the subsequent peer-review exercise. For purposes of ensuring comparability of conditions and interpreting results, the participants' Law School Admission Test (LSAT) scores, and instructor-assigned scores on the midterm were collected (48 of 58 students opted to allow their LSAT scores to be used). The participants were randomly assigned to one of the two conditions in a manner balanced with respect to their LSAT scores. For simplicity of analysis, an author could only receive reviews from reviewers in the same condition.

Participants were asked to perform a good-faith job of reviewing. The syllabus indicated, "a lack of good-faith participation in the peer-reviewing process as evidenced by a failure to provide thoughtful and constructive peer reviews may result in a lower grade on the mid-term."

Apparatus. As noted, we hypothesize that different kinds of ratings prompts focus reviewers on different aspects of the author's work. In this paper, we only examine reviewer responses to rating prompts, although reviewers also gave written evaluations of the same dimensions of peer work that they rated numerically. We collected ratings according to Likert scales (7 points, grounded at 1,3,5,7). Each condition received a different set of rating prompts, either domain-writing (Table 1), or problem-specific (Table 2).

Table 1: Domain-writing rating prompts. Reviewers rated peer work on four criteria pertaining to legal writing.

Issue Identification ("issue")	1 - fails to identify any relevant IP issues; raises only irrelevant issues 3 - identifies few relevant IP issues, and does not explain them clearly; raises irrelevant issues 5 - identifies and explains most (but not all) relevant IP issues; does not raise irrelevant issues 7 - identifies and clearly explains all relevant IP issues; does not raise irrelevant issues
Argument Development ("argument")	1 - fails to develop any strong arguments for any important IP issues 3 - develops few strong, non-conclusory arguments, and neglects counterarguments 5 - for most IP issues, applies principles, doctrines, and precedents; considers counterarguments 7 - for all IP issues, applies principles, doctrines, and precedents; considers counterarguments
Justified Overall Conclusion ("conclusion")	1 - does not assess strengths and weaknesses of parties' legal positions; fails to propose or justify an overall conclusion 3 - neglects important strengths and weaknesses of parties' legal position; proposes but does not justify an overall conclusion 5 - assesses some strengths and weaknesses of the parties' legal positions; proposes an overall conclusion

	7 - assesses strengths and weaknesses of parties' legal positions in detail; recommends and justifies an overall conclusion
Writing	1 - lacks a message and structure, with overwhelming grammatical problems
Quality	3 - makes some topical observations but most arguments are unsound
("writing")	5 - makes mostly clear, sound arguments, but organization can be difficult to follow
	7 - makes insightful, clear arguments in a well-organized manner

Table 2: Problem-specific rating prompts. Reviewers rated peer work on five problem-specific writing criteria (the claims), which all used the same scale.

Claims:	Smith v. Barry for breach of the nondisclosure/noncompetition agreement ("nda") Smith v. Barry and VG for trade-secret misappropriation ("tsm") Jack v. Smith for misappropriating Jack's idea for the I-phone-based instrument-controller interface ("idea1") Barry v. Smith for misappropriating Barry's idea for the design of a Jimi-Hydrox-related look with flames for winning ("idea2") Estate of Jimi Hydrox v. Smith for violating right-of-publicity ("rop")
Rating scale:	1 - does not identify this claim 3 - identifies claim, but neglects arguments pro/con and supporting facts; some irrelevant facts or arguments 5 - analyzes claim, some arguments pro/con and supporting facts; cites some relevant legal standards, statutes, or precedents 7 - analyzes claim, all arguments pro/con and supporting facts; cites relevant legal standards, statutes, or precedents

The researchers conducted the study via Comrade, a web-based application for peer review. For purposes of this study, Comrade was configured to conduct peer review in a manner that approximates the formal procedures of academic publication. In the tradition of the SWoRD and CPR systems, peer review in the classroom usually involves the following sequence of activities:

1. Students write essays.
2. Essays are distributed to a group of N student peers for review.
3. The peer reviewers submit their feedback to the essay authors.
4. The authors give "back reviews" to the peer reviewers.
5. The authors write new drafts of their essays.

Steps 2-5 can be repeated for multiple drafts of the same essay. In SWoRD, reviewers generate feedback (step 3) according to instructor-specified criteria, and authors evaluate the feedback they receive (step 4). Papers are chosen using an algorithm that ensures that the reviewing workload is distributed fairly, and that all authors receive a fair number of reviews. Conventionally, all students act as both authors and reviewers. As authors, they may write in response to the same domain problem or different problems. As reviewers, they may formulate their feedback in different formats, including written comments and numeric ratings.

For this study, we followed phases 1 through 4, and omitted phase 5. In addition, students took a pretest (described below) between phases 1 and 2, and a posttest

between phases 2 and 3. After authors gave back-reviews in phase 4 and before back-reviews were delivered to reviewers, all students were invited to fill out a survey.

Procedure. As stated, the participants' activities in our study proceeded in five phases, with a pretest and posttest before and after the reviewing.

Just prior to the peer-review exercise, participants completed writing a mid-term, open-book, take-home examination. It comprised one essay-type question, and students were limited to writing no more than four double-spaced or 1.5-spaced typed pages. Students had 3 days to answer the exam question.

As is typical of law school essay exams, the question presented a fairly complex (2-page, 1.5-spaced) factual scenario and asked students "to provide advice concerning [a particular party's] rights and liabilities given the above developments." The instructor designed the facts of the problem to raise issues involving many of the legal claims and concepts (e.g., trade secret law, shop rights to inventions, right of publicity, passing off) that were discussed in the first part of the course. Each claim involves different legal interests and requirements and presents a different framework for viewing the problem. Students were expected to analyze the facts, identify the claims and issues raised, make arguments pro and con resolution of the issue in terms of the concepts, rules, and cases discussed in class, and make recommendations accordingly. Since the instructor was careful to include factual weaknesses as well as strengths for each claim, the problem was ill-defined; strong arguments could be made for and against each party's claims.

Based roughly on the legal claims, concepts, and issues addressed in the exam question, the instructor also designed a multiple choice test in two equivalent forms (A and B), each with 15 questions. The test was intended to assess whether student reviewers learned from the peer-reviewing experience. The questions addressed roughly the same legal claims and concepts as the exam, but not in the same way as the exam, involving completely different facts, and in a multiple choice format rather than in essay form. After preparing the tests, the instructor invited several particularly strong students who had taken the same course in prior years to take the test. The instructor then revised the test based on these students' answers to multiple choice questions and other feedback.

On Day 1, students uploaded their anonymized midterm exam answers to Comrade from wherever they had an Internet connection. From Day 3 to 7, students logged in to review the papers of the other students. Each student received four papers to review, and each review was predicted to take about 2 hours. Before a student began reviewing, and again before he received his reviews from other students, each student completed a multiple choice test as the pretest and posttest. To control for differences between the test forms, half of the students in each condition received form A as the pretest and form B as the posttest; the other half received them in the opposite order. On Day 8, students logged in to receive reviews from their classmates. On Day 10, students provided the reviewers with back-reviews explaining whether the feedback was helpful. Students also took a brief survey on their peer review experience.

3 Results

Sensitivity to Prompts. We computed every peer author’s mean inbound peer rating for each rating prompt across the reviewers. For example, for a student in the domain-writing condition, we took four means across reviewers, namely for the prompts "issue", "argument", "writing", and "conclusion" (see Table 1). We examined the distribution of mean inbound peer ratings for each rating prompt to determine the extent to which these different rating prompts yielded non-redundant information from reviewers. Mean inbound peer ratings ranged from a low of 1.86 (problem-specific condition, “idea2” prompt) to a high of 5.54 (domain-writing condition, “writing” prompt) on a 7-point Likert scale (Table 3), showing that peer reviewers do respond to different prompts with different answers.

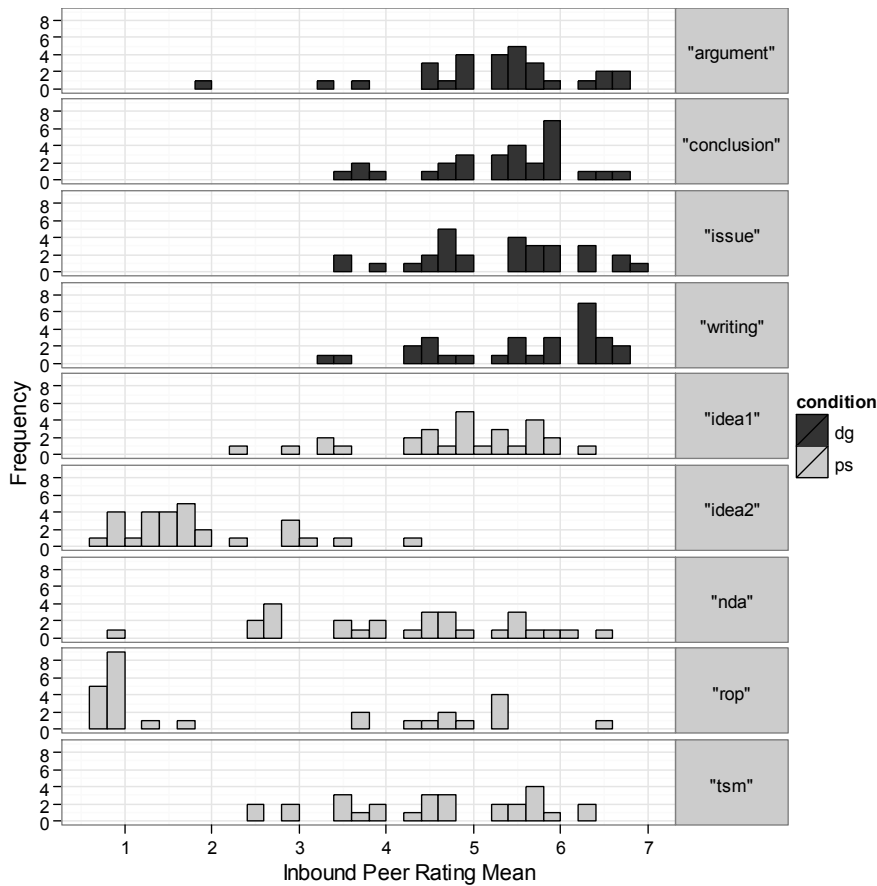


Figure 1: Frequency of inbound ratings per dimension. *dg* = domain-writing dimensions, *ps* = problem-specific.

In particular, they distinguish less among various aspects of writing composition than they do among problem-specific issues. This becomes apparent by visualizing

the frequency of the inbound ratings, as in Figure 1. All the ratings in response to domain-writing prompts tend to the right on the X axis, while the problem-specific ratings have no consistent distribution. Within each condition, we computed pairwise correlations of the mean inbound peer ratings for the prompts in that condition. The mean pairwise correlation among domain-writing ratings is 0.68, while the mean pairwise correlation among ratings in response to problem-specific prompts is 0.15.

Table 3: Mean inbound peer ratings for each rating dimension

<i>Dimension (Domain-Writing)</i>	<i>argument</i>	<i>conclusion</i>	<i>issue</i>	<i>writing</i>	
<i>Mean (SD)</i>	5.23 (1.02)	5.34 (0.838)	5.33 (0.93)	5.54 (0.98)	
<i>Dimension (Problem-Specific)</i>	<i>ideal</i>	<i>idea2</i>	<i>nda</i>	<i>rop</i>	<i>tsm</i>
<i>Mean (SD)</i>	4.83 (0.995)	1.86 (0.878)	4.26 (1.34)	2.65 (2.02)	4.59 (1.12)

Learning Outcomes. We measured student comprehension of some aspects of domain knowledge before and after students gave feedback to each other. Neither the ratings that authors received with respect to domain writing skills nor those related to problem-specific issues were predictive of their performance on pre-test or post-test. For each kind of rating prompt, we used a linear regression to model pre-/post-test performance as a function of mean inbound peer ratings. In each case, the linear models predicted less than 1% of the variance in test performance.

4 Discussion

Our aim is to understand how peer review can bring value to the classroom, and to emphasize those elements of peer review that benefit learners the most. We have presented evidence that student peer reviewers are sensitive to the difference between two types of rating prompts, domain-writing and problem-specific, and that their ratings of each other's work are less discriminating with respect to the former than to the latter. This is likely to be pedagogically important. Since peer reviewers' ratings of different aspects of domain-specific writing composition are highly correlated, they are likely to communicate redundant information to authors, and soliciting these ratings is not an effective use of the reviewers' time. On the other hand, if problem-specific support to reviewers leads to ratings that do not correlate with each other, such ratings are not redundant, and more likely to be informative. In particular, the problem-specific support relates to legal claims, each of which provides a different framework for analyzing the ill-defined problem. The different problem-specific reviews may thus lead authors to frame the problem in different ways, and the exercise of reading and making sense of the somewhat divergent problem-specific suggestions is likely to be pedagogically fruitful. [16, 17] One direction for future research is to examine whether peer authors respond differently to feedback on

writing versus on problem-specific aspects by looking at back-reviews (the authors' responses to reviewers) and subsequent drafts, and by surveying students. If problem-specific support is indeed valuable, this suggests that intelligent and adaptive support for peer review may also benefit students.

Our study complements the research of Wooley [18], which showed that students' subsequent writing improves when they give ratings and written comments, and not only numeric ratings. Subsequent writing quality was operationally defined as expert-assigned scores of student essays. In both conditions of our experiment, reviewers gave feedback as ratings and as written comments. Having thus controlled for the effect identified by Wooley, we found that giving feedback did not contribute to student understanding of key domain concepts, as measured by an objective test. It is possible that our objective test was not sensitive to what students were learning from the act of giving feedback, and it remains for future work to examine in detail the written comments that students give each other, and to look for signs that the prompts seen by students have an effect in the reviewers' and authors' subsequent writing. Author understanding of domain-independent feedback is a critical mediating factor in what feedback authors actually implement in a subsequent draft. [14] Our study suggests that feedback regarding problem-specific aspects may be more useful to authors than feedback on writing composition; making feedback more useful may lead to greater implementation as well.

Although we found that giving feedback in response to either kind of prompt did not contribute to student understanding of key domain concepts, many others have found that prompts can indeed support learning. Renkl and colleagues found positive learning outcomes for students who received metacognitive scaffolding through prompts rather than the support of "cognitive" task-oriented prompts. [19, 20] We explore metacognitive support for peer review in another study. In related work, King describes several discourse patterns that can benefit learning outcomes in settings such as problem solving and peer tutoring. [21] Another way to encourage learning in peer reviewers could be to ensure that all students review low-quality work. [22]

These results have significance for the design, implementation, and evaluation of intelligent tutoring systems that employ peer review as a mechanism for teaching skills of ill-defined problem-solving. For instance, we expect the problem-specific ratings to be especially useful for Comrade; as it compiles a student model based on the students' feedback, it needs to know whether the reviewers' ratings provide useful information. We plan to investigate its impact on reviewers and authors in future work.

References

1. Topping, K.: Peer assessment between students in colleges and universities. *Review of Educational Research*. 68, 249-76 (1998).
2. Cho, K., Schunn, C.D.: Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers and Education*. 48, (2007).
3. Sluijsmans, D.: The use of self-, peer- and co-assessment in higher education: a review of literature. Educational Technology Expertise Centre Open University of the Netherlands, Heerlen (1998).

4. Goldin, I.M., Ashley, K.D., Pinkus, R.L.: Teaching Case Analysis through Framing: Prospects for an ITS in an Ill-defined Domain. Workshop on Intelligent Tutoring Systems for Ill-Defined Domains, 8th International Conference on Intelligent Tutoring Systems. , Jhongli, Taiwan (2006).
5. VanLehn, K., Lynch, C., Schulze, K., Shapiro, J., Shelby, R., Taylor, L., Treacy, D., Weinstein, A., Wintersgill, M.: The Andes Physics Tutoring System: Lessons Learned. *International Journal of Artificial Intelligence and Education*. 15, (2005).
6. Koedinger, K., Anderson, J., Hadley, W., Mark, M.: Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education*. 8, 30-43 (1997).
7. Voss, J.: Toulmin's Model and the Solving of Ill-Structured Problems. *Arguing on the Toulmin Model: New Essays in Argument Analysis and Evaluation*. Springer (2006).
8. Russell, A.: 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 (2004).
9. Gehringer, E.: Strategies and mechanisms for electronic peer review. 30th Annual Frontiers in Education Conference. pp. F1B/2-F1B/7 vol.1 (2000).
10. Zhi-Feng Liu, E., Lin, S., Chiu, C., Yuan, S.: Web-based peer review: the learner as both adapter and reviewer. *IEEE Transactions on Education*. 44, 246-251 (2001).
11. Masters, J., Madhyastha, T., Shakouri, A.: ExplaNet: A collaborative learning tool and hybrid recommender system for student-authored explanations. *Journal of Interactive Learning Research*. 19, 51-74 (2008).
12. Hsiao, I., Brusilovsky, P.: Modeling peer review in example annotation. 16th International Conference on Computers in Education. pp. 357-362 , Taipei, Taiwan (2008).
13. Gouli, E., Gogoulou, A., Grigoriadou, M.: Supporting self-, peer-, and collaborative-assessment in e-learning: the case of the PEEr and Collaborative ASSESSment Environment (PECASSE). *Journal of Interactive Learning Research*. 19, 615 (2008).
14. Nelson, M., Schunn, C.D.: The nature of feedback: how different types of peer feedback affect writing performance (2008).
15. Walvoord, M.E., Hoefnagels, M.H., Gaffin, D.D., Chumchal, M.M., Long, D.A.: An analysis of Calibrated Peer Review (CPR) in a science lecture classroom. *Journal of College Science Teaching*. 37, 66-73 (2008).
16. Pinkus, R., Gloeckner, C., Fortunato, A.: Professional knowledge and applied ethics: a cognitive science approach. under review.
17. McNamara, D., Kintsch, E., Songer, N., Kintsch, W.: Are good texts always better? interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*. 14, 1 - 43 (1996).
18. Wooley, R., Was, C.A., Schunn, C.D., Dalton, D.W.: The effects of feedback elaboration on the giver of feedback. B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. pp. 2375-2380 Cognitive Science Society, Washington, DC (2008).
19. Hübner, S., Nückles, M., Renkl, A.: Prompting cognitive and metacognitive processing in writing-to-learn enhances learning outcomes. 28th Annual Conference of the Cognitive Science Society. (2006).
20. Nückles, M., Hübner, S., Renkl, A.: Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction*. 19, 259-271 (2009).
21. King, A.: ASK to THINK-TEL WHY: A model of transactive peer tutoring for scaffolding higher level complex learning. *Educational Psychologist*. 32, 221-235 (1997).
22. Cho, K., Cho, Y.H.: Learning from ill-structured cases. 29th Annual Cognitive Science Society Conference. p. 1722 Cognitive Science Society, Austin, TX (2007).