Available online at www.sciencedirect.com

ScienceDirect

Computers & Education 48 (2007) 409–426

COMPUTERS & EDUCATION

www.elsevier.com/locate/compedu

Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system

Kwangsu Cho *, Christian D. Schunn

Learning Research and Development Center, 3939 O'Hara Street, University of Pittsburgh, Pittsburgh, PA 15260, USA

Accepted 7 February 2005

Abstract

This paper describes how SWoRD (scaffolded writing and rewriting in the discipline), a web-based reciprocal peer review system, supports writing practice, particularly for large content courses in which writing is considered critical but not feasibly included. To help students gain content knowledge as well as writing and reviewing skills, SWoRD supports the whole cycle of writing, reviews, back-reviews, and rewriting by scaffolding the journal publication process as its authentic practice model. In addition, SWoRD includes algorithms that compute individual reviewer's review accuracy, which is in turn used to support the various drawbacks of reciprocal peer reviews (e.g., variation in motivation or ability of reviewers). Finally, this paper describes an empirical evaluation showing that the SWoRD approach is effective in improving writing quality in content classes.

Keywords: Adult learning; Architectures for educational technology system; Distributed learning environments; Improving classroom teaching; Post-secondary education

1. Introduction

Professional and academic success in all disciplines depends, at least in part, upon writing skills. Nevertheless, recent nation-wide surveys in the US reported that most of students across all ages

* Corresponding author. Tel.: +1 412 624 2679.
E-mail address: kwangsu@pitt.edu (K. Cho).

0360-1315/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compedu.2005.02.004

have fundamental difficulties in their writing skills. For example, a study reported that only one percent of students tend to have effective writing skills, while 85% of students are at a basic writing skill level (NAEP, 1998).

Additional findings cast some doubt on formal attention to writing practice in colleges and universities. National Commission on Writing in American Schools and Colleges (2003) found that content classes outside English composition classes are providing *near-total neglect of writing*. This unfortunate situation appears to be caused by instructors' workload in reading and generating feedback on student writing. As a result, students, especially in large content courses, rarely practice writing and rewriting. Instead, students tend to be confined to multiple-choice exams with no writing assignments and not even short essay writing in exams. This lack of writing practice is likely to have a strong negative impact on the development of effective writing skills, and probably on the development of content knowledge.

It seems a natural choice to replace instructor or expert reviews with reciprocal peer reviews to remedy the unfortunately current situation. Peer reviews can help instructors spend more time on other aspects of teaching by reducing the instructors' workload associated with writing activities (Rada, Michailidis, & Wang, 1994). However, there are several different problems in administrating reciprocal peer reviews in content courses. First, reciprocal peer reviews may be fundamentally limited in that student peer reviewers are novices in their disciplines. Thus, their feedback and evaluation could be inaccurate relative to the feedback generated by a subject-matter expert or instructor. Second, students are likely to be inexperienced in constructing helpful reviews because students generally are not trained in how to generate helpful comments to others' writing. Finally, instructors often complain about the work required to administrate reciprocal peer reviews such as selecting reviewers for writers and exchanging writing and reviewing.

Therefore, we have developed a web-based reciprocal peer review system called SWoRD to addresses the above and related issues. Since we observe that learning systems developed for research often fail to survive in real classrooms, SWoRD has been trying to lower its technology threshold for students as well as instructors to use while keeping the system secure and stable. As a result, SWoRD has been used successfully in 20 courses across four US universities in the past two years.

This paper describes how SWoRD (version 3.0) addresses the above issues in its implementation. This paper is organized the following way: (1) the journal publication process as an authentic practice model is first described; (2) reciprocal peer reviews and related problems are discussed; (3) SWoRD's design and implementation are explained; (4) a summary of an empirical evaluation obtained thus far with SWoRD is provided; and (5) ongoing work with SWoRD is briefly introduced. Note that this paper only focuses on the key activities in SWoRD that seem most relevant for supporting student learning. Other various management functions and instructor interface are not discussed.

2. Journal publication process as an authentic model

By applying a cognitive work analysis to the journal publication process and writing practice in college and university courses, we became aware of some key steps of writing practice in small and large classes, and identified the steps that are missed in the courses (see Fig. 1). In this section, we

Journal Pub.	Small Classes	Large Classes	SWoRD
Authors select topics	Teacher gives writing topics	Teacher gives writing topics	Teacher sets a pool of topics for students to choose
Authors write papers	Students write papers	Students rarely write papers.	Authors write and publish papers
Reviewers in the domain generate feedback	Teachers read the papers	Some teachers do not take a careful look at papers even if students submitted the papers	Reviewers review papers
Authors get the feedback	Students get the grade with feedback.	Students get the grade, but not specific feedback on their writing	Authors get feedback and give feedback to reviewers
Authors carefully digest the feedback	Students typically do not read the feedback	Students typically do not read the feedback	Authors read and evaluate the feedback
Authors revise the papers	Students are rarely asked to revise papers	Students are rarely asked to revise their papers	Authors revise papers
			Authors send feedback to reviewers
Authors publish the papers			Authors publish the papers

Fig. 1. Cognitive work analysis of writing practice.

discuss the major problems of writing practice in classes. How SWoRD addresses them is discussed in a later section.

It was found that writing practice in classes is a function of instructor feedback, which plays a major barrier to increasing writing practice. Instructors consider feedback crucial to developing students' writing skills and steering their thoughts in the process of writing and rewriting. However, instructors are simply overwhelmed by the workload of reading student writing and giving feedback on it. As a result, they rarely ask students to write papers in classes. In other words, students practice writing only when instructors can provide them with feedback. Even when they do provide feedback, instructors tend to generate ambiguous or evaluative feedback such as "Good job", "You could be better", etc. (Coupe, 1986) because instructors are often unable to take enough time to generate specific feedback.

When instructors administrate writing practice, students are rarely asked to revise their writing based on feedback. It is simply because rewriting doubles instructors' workload. Even students in small classes are in a similar situation. Thus, students in general are unlikely to have the opportunities to rewrite their papers. Because of this, students are losing critical chances to develop their writing skills by incorporating feedback into their papers. Writing researchers emphasize the practice of revision and rewriting using feedback to advance writing skills (Schriver, 1990).

Finally, writing and feedback are privatized by a student author and an instructor. Other students in a class do not have access to peer writing and feedback given to it. However, writing researchers have argued that publicizing student papers to their peers may have students put more effort into writing because students are made aware of audiences out there (Cohen & Riel, 1989; Jonassen, 1996; Schriver, 1990). However, in classes, there is typically no publication process to the class. The only audience for a student's writing is their instructor.

3. Reciprocal peer reviews and problems

One way to include and improve writing practice in content classes is to implement reciprocal peer reviews in the writing process. In reciprocal peer reviews, individual students take two roles: one of writer and one of reviewer. Studies have showed that reciprocal evaluation has the advantage of reducing teacher workload (Rada et al., 1994). Moreover, peer collaboration is effective in that students working alone are unlikely to detect their own misunderstanding (Markman, 1979) and contradictions in text (Otero & Kintsch, 1992), to consider audience (Wess, 1980), but students working collaboratively are better able to avoid these problems (Cho, Schunn, & Lesgold, 2002).

Nevertheless, reciprocal peer reviews have clear drawbacks. First, undergraduate students are generally novices in their disciplines. Thus, their lack of subject-matter knowledge tends to impair the accuracy of determining the quality of papers that novices review. For example, expert-novice studies showed that novices focus on style issues, not making theoretical commitments (Flower, Hayes, Carey, Schriver, & Stratman, 1986). This issue may also increase the concern of student reviewees about grades assessed by unqualified people like themselves.

Second, there are likely to be many individual differences between peer reviewers such that there are some good reviews and some poor reviews. For example, some student reviewers might be unable to distinguish between good and poor papers, or perhaps some may not put in the effort to properly evaluate the papers. Also, peer reviewers might give authors all the same evaluations.

Finally, peer reviews should not be just critical, but also constructive. In face-to-face collaboration, people try to be kind and so tend to give ambiguous advice, while in networked collaboration people tend to be critical on task (Sproull & Kiesler, 1991). However, critical advice is not always constructive especially in networked collaboration. Also people can take task-oriented critiques personally (Crampton, 2001), causing strong emotional reaction.

In order to obtain the practical benefits of reciprocal peer reviews, the above problems must be addressed in the system. SWoRD differs from other such systems mainly in that it seriously considers these problems through the simulation of the journal publication process as an authentic writing practice model. Therefore, the next section describes how SWoRD addresses these issues in its implementation.

4. SWoRD characteristics

SWoRD is a web-based client–server application, supporting reciprocal peer reviews in writing and reviewing practices. It is especially useful for large-scale content classes where writing and rewriting are hard to administrate, although SWoRD can also serve small-scale or skills classes. It is also an asynchronous system that does not provide users with any synchronous tools such as IRC interfaces for real-time communications between writers and reviewers. This approach is consistent with the findings of some studies showing that the amount of interaction is not correlated with the quality of writing (Rada et al., 1994) and that people prefer to work on writing in an asynchronous mode rather than synchronous mode (Hartman et al., 1995), and even reduce their interactions during writing (Galegher & Kraut, 1996). In this section, we describe the core SWoRD activities first and then elaborate some key features of the system.

4.1. Authentic writing and reviewing process

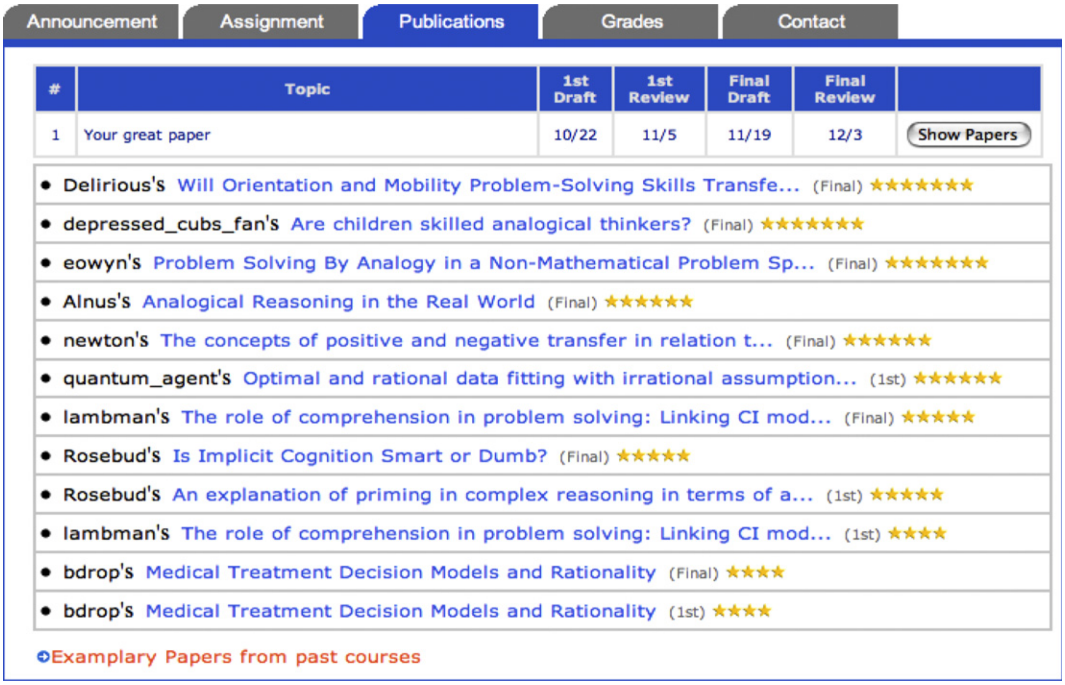
As shown in Fig. 1, SWoRD mainly simulates the journal publication process (with some modifications) as its writing practice model in classes. This section describes what students basically do in SWoRD. In the beginning, instructors register their courses in SWoRD, set a pool of topics for writing and reviewing, specify due dates for writing and reviewing activities, and some policies such as grace periods for late submissions, penalty for late submissions, and number of peer drafts for reviews. Instructors can just follow recommended defaults instead of setting their own policies. For example, SWoRD recommends that each due time and date should be Monday 5 p.m. with 2 days grace periods.

Then, students select topics that they want to write on and those that they want to review on. If there is only one topic, the topic is automatically assigned to students. In the case that multiple topics are available to students, SWoRD controls the number of writers and reviewers in each topic to provide an approximate even balance of reviews to each paper. Therefore, latecomers may not have as many topic choices. Then, based on a clear schedule of due dates for their chosen topics, students do writing and reviewing activities in five phases as follows.

In phase I, student writers submit their first drafts to SWoRD. When submitting drafts, they also provide estimates of their own writing grades for increased metacognition about their writing skills. SWoRD, then, generates sets of drafts based on instructors' policy for how many papers each student should review. Each subset is then distributed to reviewers who chose the topic for review. As shown in Fig. 2,[1] all the papers that were submitted are published to the class on the publication page, where each piece of writing is displayed with its author's pseudonym and its title. Then, any student in the class can access and read any peer writing.

In phase II, reviewers download a set of papers assigned to them, and the reviewers evaluate each paper in the set on the basis of three evaluation dimensions: flow, logic, and insight (defined in a later section). The reviewers submit written comments on how to improve the papers and then rate the quality of the drafts on a seven-point rating scale (1: Disastrous to 7: Excellent). At the end of phase II, when the first draft review deadlines have passed, SWoRD automatically determines the accuracy of each reviewer's numerical ratings using three scales (systematic differences, consistency, and spread) applied to each of the writing dimensions (flow, logic, and insight). The details of these scales will be described in a later section.

[1] Examples were from an undergraduate psychology course for non-majors. All other figures were from similar courses.



#	Topic	1st Draft	1st Review	Final Draft	Final Review	
1	Your great paper	10/22	11/5	11/19	12/3	Show Papers

- Delirious's Will Orientation and Mobility Problem-Solving Skills Transfe... (Final) ★★★★★★★
- depressed_cubs_fan's Are children skilled analogical thinkers? (Final) ★★★★★★★
- eowyn's Problem Solving By Analogy in a Non-Mathematical Problem Sp... (Final) ★★★★★★★
- Alnus's Analogical Reasoning in the Real World (Final) ★★★★★★
- newton's The concepts of positive and negative transfer in relation t... (Final) ★★★★★★
- quantum_agent's Optimal and rational data fitting with irrational assumption... (1st) ★★★★★★
- lambman's The role of comprehension in problem solving: Linking CI mod... (Final) ★★★★★
- Rosebud's Is Implicit Cognition Smart or Dumb? (Final) ★★★★★
- Rosebud's An explanation of priming in complex reasoning in terms of a... (1st) ★★★★★
- lambman's The role of comprehension in problem solving: Linking CI mod... (1st) ★★★★
- bdrop's Medical Treatment Decision Models and Rationality (Final) ★★★★
- bdrop's Medical Treatment Decision Models and Rationality (1st) ★★★★

Examplary Papers from past courses

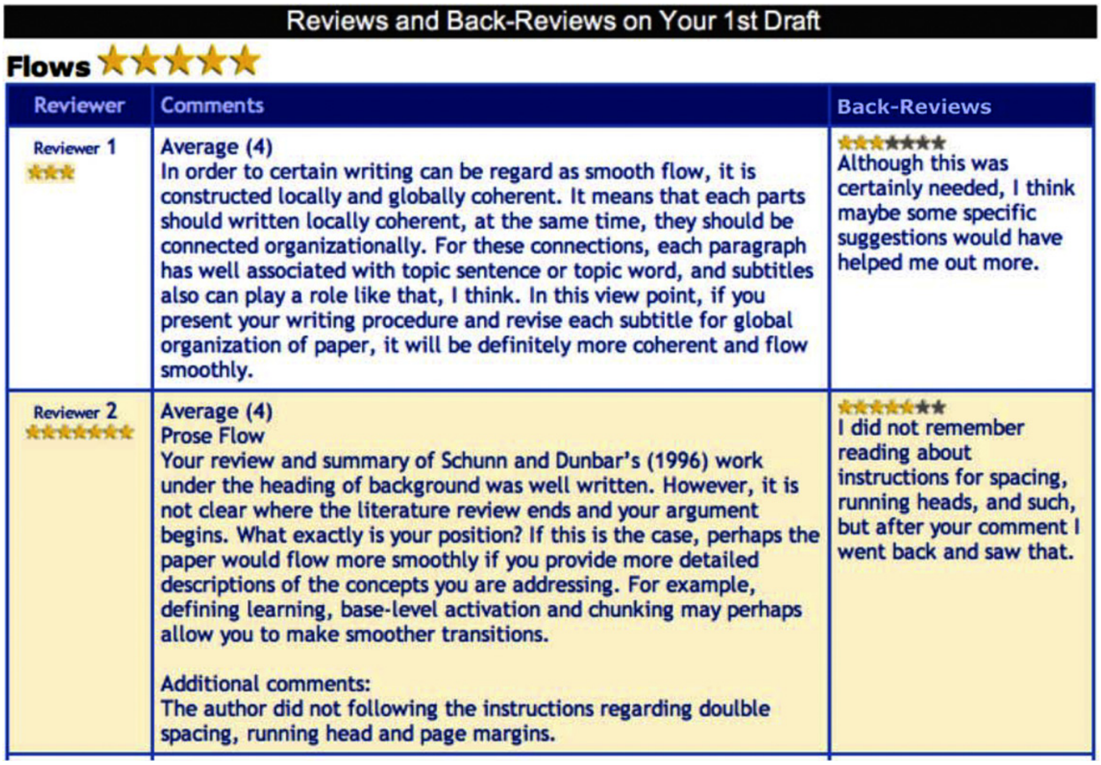
Fig. 2. Publication page interface.

The review accuracy scales have several functions. First, the review scales serve as a weighting function to add validity to the writing scales: reviews from reviewers that are generally less accurate count less towards a given author's grade. Second, the review scales serve as a calibration tool for the reviewers, giving them feedback on how the reviewers might have some misconceptions about what constitutes good writing. Third, the review scales serve as part of an incentive system for the reviewers by providing grades for their numerical rating activities – peer review schemes can suffer from low motivation to take the difficult reviewing task seriously.

Then, SWoRD provides authors with reviewers' feedback and their weighted ratings (see the middle column of Fig. 3) and the reviewers with feedback on their review accuracy. At the same time, the publication page displays drafts in the order of their grades with a number of stars based on their ratings. Any student can access reviewers' comments and grades as well as the papers.

In phase III, the authors rewrite their papers based on comments provided and turn in the final drafts. Throughout the third phase, reviewers reflect upon the feedback that they receive on their reviews. Reviewers are given a table which gives their accuracies on each of the three scales for each of the three dimensions. Clicking on any of the numbers brings up a page which verbally and graphically explains what aspects of their ratings produced the number (see Figs. 4–6 as examples). Also, reviewers are given an interface with which they can unpack the accuracy of their reviews by comparing their own reviews with other reviewers' on same papers (see Fig. 7).

In phase IV, writers back-review their reviewers' feedback in terms of how helpful the written feedback (not the numerical rating) was for revising their paper on a seven-point rating scale (1: Not helpful at all to 7: Very helpful) along with written comments (see the right half of Fig. 3).



Reviews and Back-Reviews on Your 1st Draft

Flows ★★★★★

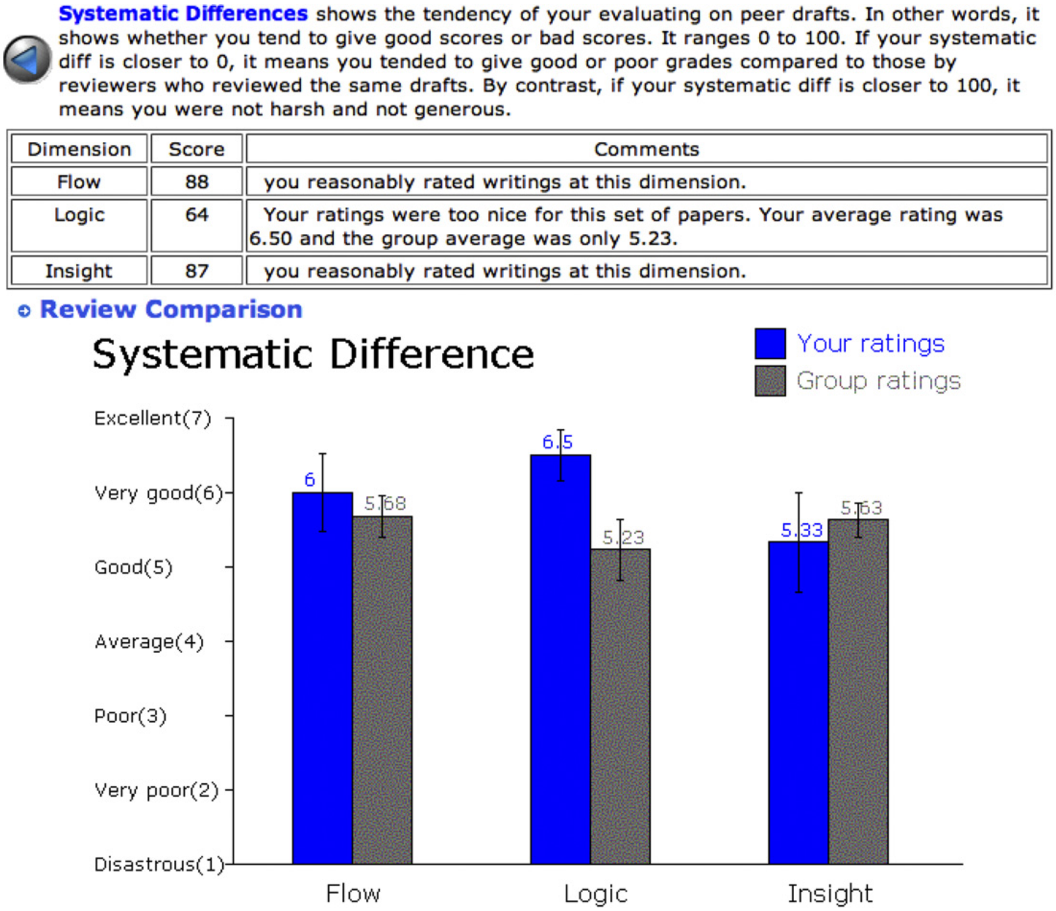
Reviewer	Comments	Back-Reviews
Reviewer 1 ★★★	Average (4) In order to certain writing can be regard as smooth flow, it is constructed locally and globally coherent. It means that each parts should written locally coherent, at the same time, they should be connected organizationally. For these connections, each paragraph has well associated with topic sentence or topic word, and subtitles also can play a role like that, I think. In this view point, if you present your writing procedure and revise each subtitle for global organization of paper, it will be definitely more coherent and flow smoothly.	★★★☆☆☆☆ Although this was certainly needed, I think maybe some specific suggestions would have helped me out more.
Reviewer 2 ★★★★★★★	Average (4) Prose Flow Your review and summary of Schunn and Dunbar's (1996) work under the heading of background was well written. However, it is not clear where the literature review ends and your argument begins. What exactly is your position? If this is the case, perhaps the paper would flow more smoothly if you provide more detailed descriptions of the concepts you are addressing. For example, defining learning, base-level activation and chunking may perhaps allow you to make smoother transitions. Additional comments: The author did not following the instructions regarding doulble spacing, running head and page margins.	★★★★★☆☆ I did not remember reading about instructions for spacing, running heads, and such, but after your comment I went back and saw that.

Fig. 3. A partial view of peer feedback and back-review interface. The stars next to Flows indicate the overall flow quality of the writing. The stars under each reviewer describe each reviewer's reliability that SWoRD computed. For example, the reviewer 1 has three, while the reviewer 2 has seven. Thus, SWoRD considers reviewer 2 more reliable. Finally, the stars under the back-reviews column indicate the helpfulness ratings given from the author. Thus, the author in the example gave 3 points to the reviewer 1's written comments and 5 points to the reviewer 2's comments.

It should be noted that this step is done by authors after their revisions are turned in, which reduces the tit-for-tat strategy between writers and reviewers. In other words, both the time gap and actual revision activity allow the writers to become more objective raters. In an earlier version of the system that did not include this cooling off period, authors often gave much higher helpfulness ratings to reviewers who gave overly generous ratings. After this back-review period, reviewers receive their writers' comments on their feedback.

In phase V, as the final cycle, the same set of the reviewers of the first drafts read the final drafts and generate another round of ratings and written comments. Reviewers can access their comments on the first draft of each paper as well as the back reviews on those comments. At the end of phase V, after reviews are turned in, SWoRD again computes each individual reviewer's review accuracy, provides writers with their reviewers' comments and weighted grades, and provides reviewers with SWoRD feedback on their reviewing accuracy.

Finally, receiving the final round of reviews on their final writing, authors (after another cooling off period) back-review the reviewers' 2nd round of feedback in terms of how helpful the feedback would be in revising their rewriting, again using the seven-point rating scale with written comments.



Systematic Differences shows the tendency of your evaluating on peer drafts. In other words, it shows whether you tend to give good scores or bad scores. It ranges 0 to 100. If your systematic diff is closer to 0, it means you tended to give good or poor grades compared to those by reviewers who reviewed the same drafts. By contrast, if your systematic diff is closer to 100, it means you were not harsh and not generous.

Dimension	Score	Comments
Flow	88	you reasonably rated writings at this dimension.
Logic	64	Your ratings were too nice for this set of papers. Your average rating was 6.50 and the group average was only 5.23.
Insight	87	you reasonably rated writings at this dimension.

Fig. 4. The interface used to explain the systematic difference dimension of reviewing accuracy.

In each step of the above activities, SWoRD sends reminders and confirmation emails on their tasks as well as places the reminders on their announcement page. It is important to post announcements because some students do not receive reminders due to over-quota email accounts or other technical problems. Another important feature is that every deadline has a grace period (instructor-determined but typically two days) during which papers, reviews, and back reviews can be turned in late, but with penalty. After this grace period is over, submissions are released to authors or reviewers, as the case may be. Submissions cannot be accepted into the system beyond this point.

4.2. Anonymous writing and reviewing activities

Students are anonymous both as author and reviewer. That is, reviewers do not know who wrote a given paper, and authors do not know who reviewed their paper. Anonymity is important because writers and reviewers are likely to be less critical when identities are known (Crampton, 2001). However, there are situations in which writers and reviewers need to somehow identify

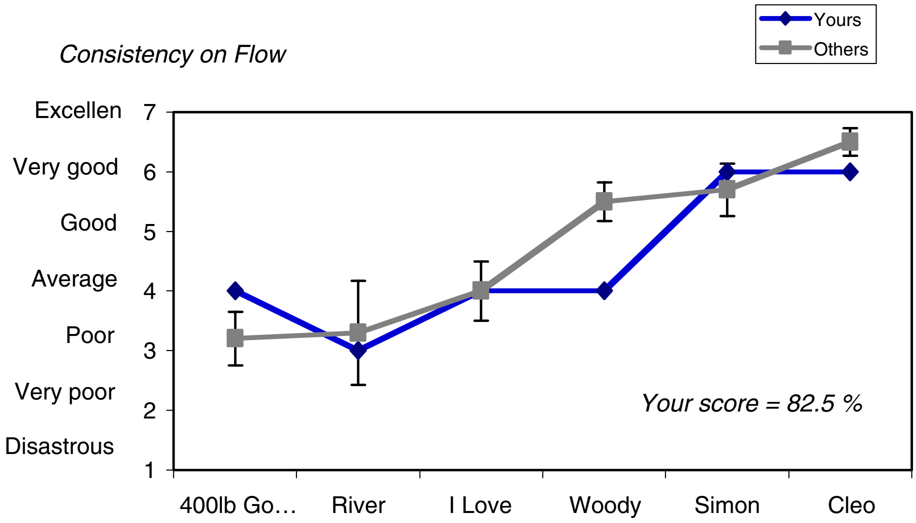
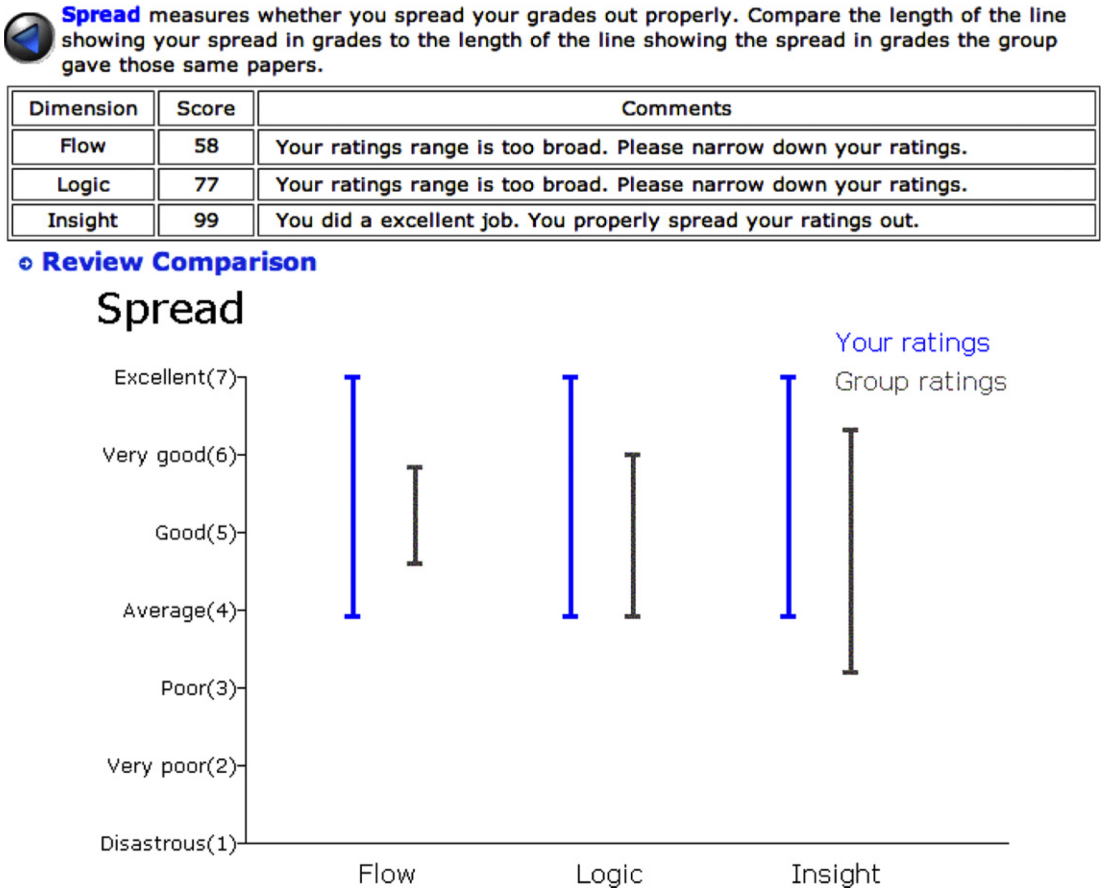


Fig. 5. An example of reviewer support on consistency showing the extent which a reviewer's grades across papers she reviewed were consistent with those by other reviewers who reviewed the same papers. The pattern across the drafts shows that the reviewer's grades are consistent with other reviewers except for the drafts by the author 400lb Gorilla, Woody, and Cleo. If the reviewer clicks on an author's name, then she can read her own reviews and the reviews of others on the given papers.

their reviewers and writers (e.g., reviewer coordinating printouts of papers with online versions). For this coordination function, writers use pseudonyms that they created when opening SWoRD accounts. Student writers submit their writings to SWoRD, which distributes the writings with the authors' pseudonyms to reviewers. When receiving feedback from reviewers, writers are not given any identity information to prevent authors from biasing their evaluations of other students' papers on the basis of what feedback the writers received from those same students as reviewers.

4.3. Individual writing and rewriting

SWoRD emphasizes individual effort in writing and rewriting papers rather than asking students to work on writing with peers collaboratively. Although some research has shown that knowledge constructed during collaboration could be transferred to individuals, other studies have showed that only some members benefited from collaboration (Webb, 1995). Thus, it is still an open question how much each member in a team can develop their writing skills and knowledge in depth from collaboration. Zammuner (1995) showed that students who practice writing benefit most when they work individually and then cooperate, but not when they continuously cooperate. Therefore, it seems reasonable to have students write their own papers and work on them with peer feedback. This approach is supported by studies showing that, for complex tasks like writing, the pattern for successful networked collaboration is for participants to focus on their own work while collaborating on higher order structure (Vera, Kvan, West, & Lai, 1998) and that working on complicated tasks in the presence of others may prevent people from properly focusing on the details of the task (Kiesler & Cummings, 2002). It should be noted that SWoRD



Spread measures whether you spread your grades out properly. Compare the length of the line showing your spread in grades to the length of the line showing the spread in grades the group gave those same papers.

Dimension	Score	Comments
Flow	58	Your ratings range is too broad. Please narrow down your ratings.
Logic	77	Your ratings range is too broad. Please narrow down your ratings.
Insight	99	You did a excellent job. You properly spread your ratings out.

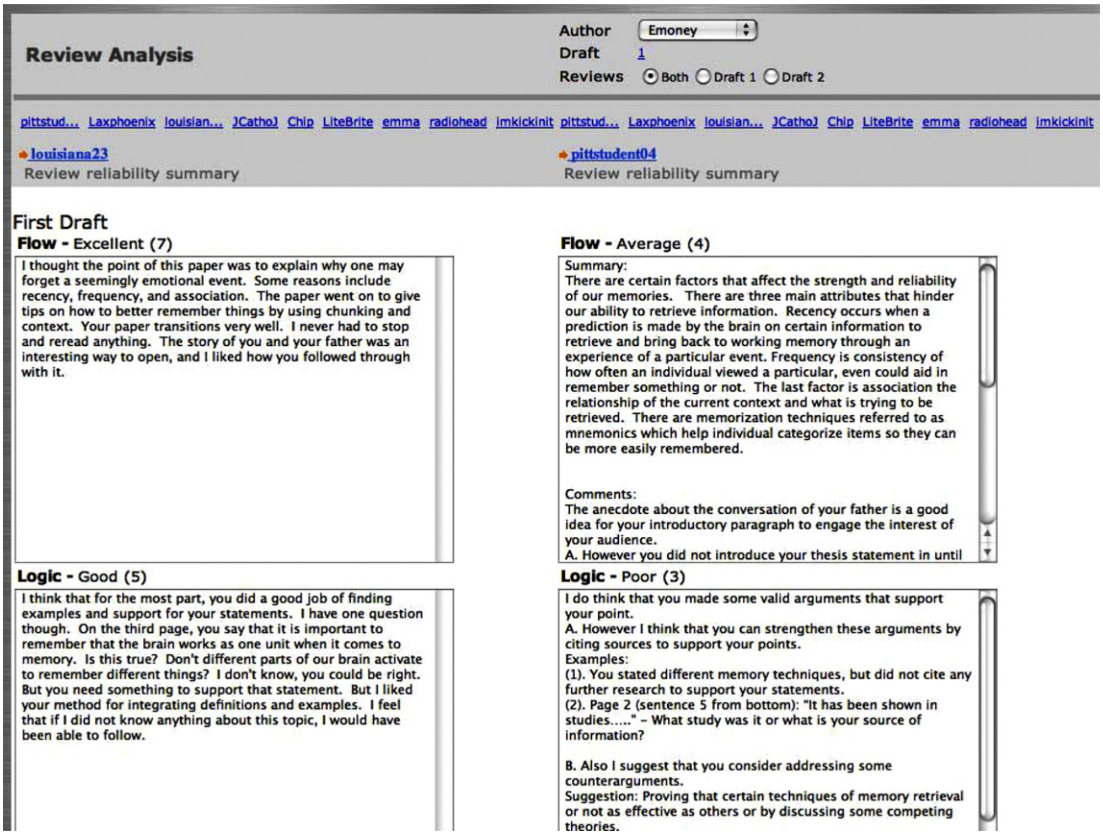
Review Comparison

Fig. 6. The interface used to explain the spread dimension of reviewing accuracy.

requires students to rewrite their drafts using feedback based upon the assumption that students can develop their writing skills and domain knowledge effectively when students revise their drafts using feedback.

4.4. Multiple peer feedback and grading

SWoRD emphasizes the role of multiple peers in generating feedback on peer writing. There are several important benefits of multiple peer reviews in SWoRD. First, writers can improve their audience conception by having multiple peer feedback (Schriver, 1990). In other words, students can come to revise their writing from the readers' points of view rather than from the more common knowledge telling strategy point of view. Second, multiple reviews could reduce blind spots and omissions of any given individual review because more reviews will mean that more errors are caught. Third, multiple reviewers could reduce the negative impact of incorrect feedback. Fourth, multiple reviewers may be in agreement on some specific problems, and this multiplicity of comments on a given problem may be especially persuasive or salient to a student when the student is



Review Analysis

Author: Emoney
Draft: 1
Reviews: ● Both ○ Draft 1 ○ Draft 2

pittstud... Laxphoenix louisian... JCathoJ Chip LiteBrite emma radiohead lmkickinit pittstud... Laxphoenix louisian... JCathoJ Chip LiteBrite emma radiohead lmkickinit

louisiana23
Review reliability summary

pittstudent04
Review reliability summary

First Draft

louisiana23:

Flow - Excellent (7)

I thought the point of this paper was to explain why one may forget a seemingly emotional event. Some reasons include recency, frequency, and association. The paper went on to give tips on how to better remember things by using chunking and context. Your paper transitions very well. I never had to stop and reread anything. The story of you and your father was an interesting way to open, and I liked how you followed through with it.

Logic - Good (5)

I think that for the most part, you did a good job of finding examples and support for your statements. I have one question though. On the third page, you say that it is important to remember that the brain works as one unit when it comes to memory. Is this true? Don't different parts of our brain activate to remember different things? I don't know, you could be right. But you need something to support that statement. But I liked your method for integrating definitions and examples. I feel that if I did not know anything about this topic, I would have been able to follow.

pittstudent04:

Flow - Average (4)

Summary:
There are certain factors that affect the strength and reliability of our memories. There are three main attributes that hinder our ability to retrieve information. Recency occurs when a prediction is made by the brain on certain information to retrieve and bring back to working memory through an experience of a particular event. Frequency is consistency of how often an individual viewed a particular, even could aid in remember something or not. The last factor is association the relationship of the current context and what is trying to be retrieved. There are memorization techniques referred to as mnemonics which help individual categorize items so they can be more easily remembered.

Comments:
The anecdote about the conversation of your father is a good idea for your introductory paragraph to engage the interest of your audience.
A. However you did not introduce your thesis statement in until

Logic - Poor (3)

I do think that you made some valid arguments that support your point.
A. However I think that you can strengthen these arguments by citing sources to support your points.
Examples:
(1). You stated different memory techniques, but did not cite any further research to support your statements.
(2). Page 2 (sentence 5 from bottom): "It has been shown in studies....." – What study was it or what is your source of information?

B. Also I suggest that you consider addressing some counterarguments.
Suggestion: Proving that certain techniques of memory retrieval or not as effective as others or by discussing some competing theories.

Fig. 7. A partial view of the review analysis interface. The interface shows lousiana23's and pittstudent04's reviews side by side on Emoney's first paper. By changing the options given on the top, reviewers can compare different pairs of reviews.

revising his or her paper. Finally, for students to take the feedback seriously, the ratings need to count for actual grades, and the validity and reliability of the grades depends upon there being ratings from multiple reviewers. Our prior work with SWoRD found that peer grading is reliable (correlations of approximately 0.6 with instructor grades) and valid (Cho & Schunn, 2003).

The SWoRD default number of peer reviewers per draft is six. In other words, each student draft receives feedback from six peers. A single reviewer is the most convenient but least reliable. By having more evaluators per task, reliability can be increased and the *true* assessment would be closer, but with cost. By applying the *Spearman–Brown formula* (Rosenthal & Rosnow, 1991), Fig. 8 shows that the number of evaluators is a function of mean reliability. Suppose that an effective reliability across reviewers of 0.90 is acceptable. If a mean reliability of individual reviewers is 0.60, then the number of reviewers should be 6. The lower a mean reliability is, the more evaluators per task are necessary. With a moderate mean reliability between 0.5 and 0.7, 4–9 evaluators are needed. With a lower mean reliability, more than nine evaluators are necessary. Considering task difficulties and time, an acceptable degree of reliability, and educational benefits, we considered six peer reviewers most desirable in SWoRD.

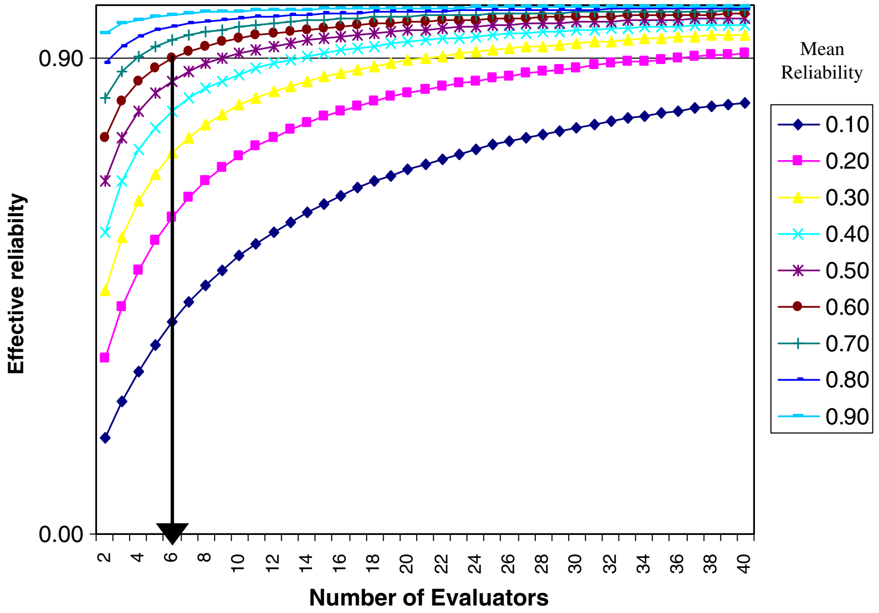


Fig. 8. Effective reliability of multiple reviews as a function of the number of reviewers and the mean reliability of individual reviewers.

4.5. Multidimensional writing evaluation: flow, logic and insight

SWoRD defines reviewing as a constructive process of detecting dissonance between the text and the writer's intention for constructing new knowledge (Flower et al., 1986; McCutchen, 2000). This process is described by the three review dimensions of flow, logic, and insight. The flow dimension, the most basic level, considers the extent to which a paper involves faults or problems in prose flow (Flower et al., 1986). Thus, can the reader understand the main points and the transition between points? The logic dimension examines the extent to which a paper is logically coherent in terms of text structure that organizes various facts and arguments. In particular, reviewers judge how well the main arguments are supported with evidence. The insight dimension refers to the extent to which each paper contributes new knowledge and insight to the reader. In classes, this is operationally defined as new knowledge or insight beyond required class texts and materials. Ackerman (1990) found that writers with more disciplinary knowledge made their arguments more elaborated and specific, and that their writing involved more new information and made it outstanding. For each dimension, reviewers submit written comments and then grade the quality of writing on the seven-point rating scale. Here the order of giving written comments before numerical ratings is thought to add validity to the ratings and avoid holistic grading problems.

4.6. Reviewer support

4.6.1. Accuracy feedback

We have developed three accuracy indices to calibrate inaccuracy of student reviews: systematic differences, consistency, and spread. All three measures depend upon a comparison of a given reviewer's ratings to the mean ratings for that set of papers across (typically six) reviewers – i.e., it is assumed that the group view is a reasonable approximation of the truth. For papers written to peers as the audience, this assumption is not as controversial as it might be for writing for expert audiences.

Systematic differences concerns the extent to which each reviewer systematically tends to be overly generous, overly harsh, or unbiased in assessing papers (see Fig. 4). It is a variation of a *t*-test between the given reviewer's ratings for their set of papers and the ratings that should have been given. Consistency concerns the extent to which each reviewer systematically discerns good papers from poor papers. It is a variation of a correlation between the given reviewer's ratings and the ratings that should have been given. Fig. 5 presents an example of the visualization that illustrates the consistency score. Finally, spread concerns the extent to which each reviewer distributes scores too narrowly or too widely (see Fig. 6). It is essentially the relative difference in standard deviations of reviewer ratings versus the ideal ratings for that set of papers. In all cases, the measures are transformed into 0 to 1 scales with 0 being the worst performance and 1 being the best performance.

The three indices are computed in each of the three dimensions (flow, logic, and insight), producing nine accuracy measures. The nine numbers are used to measure each reviewer's overall accuracy of ratings, for SWoRD to generate feedback to reviewers on their review accuracies, to compute reviewers' grades, and finally to weight the contribution of each review to the writing grades.

4.6.2. Back-review

Reviewers also can learn how to be constructive as well as critical in generating feedback by receiving feedback from the authors to which the reviewers gave written feedback comments. Back-reviews are intended to serve this purpose (see the right column of Fig. 3).

4.7. Publication

As shown in Fig. 2, student drafts are published to their classes. When their reviews are done, all of the drafts are sorted in the order of their writing qualities stamped with a number of stars from 1 to 7. Students can tell what papers are good and poor, which often reveals some intrinsic information about what could constitute good writing.

4.8. Instructors actions

In the beginning of the semester, instructors create their SWoRD course and add teaching assistants if available. Then, instructors create a pool of topics for writing and reviewing with guidelines and specific due dates. Next, policies are set up for writing and reviewing assignments in SWoRD. For example, instructors set the number of topics upon which each student writes and reviews, the number of papers each reviewer needs to review, grace periods for late assignments, penalties for late assignments, and whether each writer needs to rate their own drafts.

5. Empirical evaluation

To show the effectiveness of the SWoRD approaches, here we provide a brief overview of a recent empirical evaluation of the value of peer reviewing in SWoRD (see Cho, 2004, for more details). We compared the quality of writing improvement of student writers who received feedback from a subject-matter expert with those who received feedback from either a peer or multiple peers.

5.1. Methods

5.1.1. Participants

Participants included 28 students and a domain expert. The students were enrolled in a 12-week Research Methods summer class at the University of Pittsburgh. They had completed an average of 3.4 college years (SD = 1.0). Seven were males and 21 were females. Individual students wrote first and final drafts on the topic 'informal science learning'. They reviewed six peers' first and final drafts. The domain expert was a Ph.D. on the topic area and had taught similar courses for the past eight years. She was not the instructor of the class but reviewed both drafts of all 28 students.

5.1.2. Design

The students took a test of basic writing skills before receiving writing/reviewing assignments. Based on the scores, the students were matched into blocks and then randomly assigned to one of three different conditions: a *Single-Expert* feedback condition, a *Single-Peer* feedback condition, and a *Multi-Peer* feedback condition. The writers assigned to the Single-Expert feedback condition received feedback and grades on their drafts from the expert. Those in the *Single-Peer* feedback condition received feedback and grades from a single best peer that achieved the highest review accuracy in each set of reviewers per writer. Those in the *Multi-Peer* feedback condition received feedback and grades from six peers. To get rid of possible reviewer's status effects, the writers did not know the status of their reviewers (i.e., whether they were student or expert). The writers were told that they would not receive writing grades from their instructors, but rather from their reviewers.

5.1.3. Procedure

The procedure of the experiment followed the built-in processes in SWoRD with some modifications for experimental purposes. After taking pretests on the basic writing skills, the instructor introduced the class to the writing and reviewing assignment topic with two required readings, and SWoRD. All of the remaining procedures were managed online by SWoRD. After two weeks, the writers turned in their first drafts. Then individual student reviewers received a set of six drafts that were randomly selected by SWoRD. During a 1-week period, the reviewers individually generated written comments on six peer drafts and evaluated the quality on the seven-point rating scale (1: Disastrous to 7: Excellent). During the same period, the expert reviewed all of the drafts. Note that even though all papers were reviewed by the expert and six peers, only some of those reviews were revealed to the author, depending on their conditions. Then, the writers received selected feedback based on their conditions, revised their writing over a one-week period. Then, writers turned in their final drafts, which were reviewed by the same reviewers who reviewed their first drafts. Then, the writers back-reviewed their reviewers' feedback on a five-point rating scale in terms of how helpful it was in revising their first drafts. The results of the back-review were not delivered to the reviewers unlike the normal SWoRD procedure because it would ruin this particular experimental contrast by revealing to the reviewers what condition they were in (i.e., whether

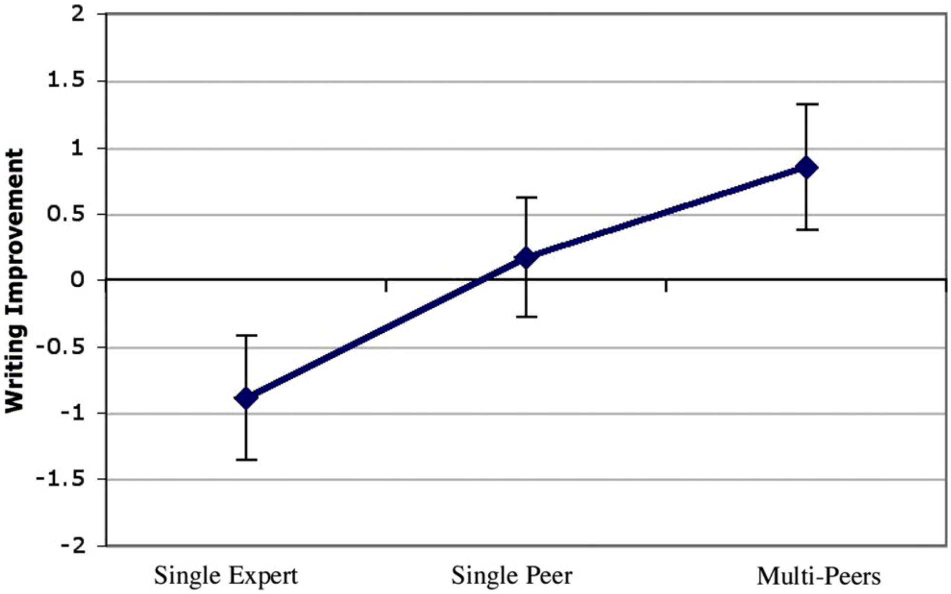


Fig. 9. Mean writing quality improvement between the first and second drafts as a function of condition and as judged by the expert.

their comment had actually been given to the author). Next, the reviewers reviewed the final drafts. Finally, the writers received the second round of feedback and back-reviewed the second feedback.

5.2. Results

Here we present one important result from a two-way mixed ANOVA on the writing quality improvement between first and final drafts based on the expert's evaluation. The expert blindly evaluated their qualities. The evaluation dimension was a within-subject variable. Note that the expert evaluated all the papers blind to condition.

As shown in Fig. 9, those in the *Multi-Peer* feedback condition showed the biggest improvement in writing qualities ($M = 0.85$, SEM[2] = 0.47) between their first drafts and final drafts, while those in the *Single-Peer* feedback condition showed only a slight improvement ($M = 0.17$, SEM = 0.45). Interestingly, those in the *Single-Expert* feedback condition performed worst ($M = -0.89$, SEM = 0.47). The writing improvement difference between the feedback conditions was statistically significant, $F(2,25) = 3.50$, $p < 0.05$. Tukey pairwiswe comparison showed that only the difference between the *Single-Expert* and *Multi-Peer* feedback condition was significant, $p < 0.05$. Thus, this result supported the SWoRD approaches in that student writers benefited from getting multiple peer feedback in that their writings improved significantly from that peer feedback. Even more strongly supportive of the SWoRD approaches, feedback from multiple peers produced especially strong improvements in writing relative to the more traditional feedback from a single expert.

[2] Standard error of the mean.

6. Conclusions

The goal of this paper was to demonstrate the SWoRD approaches to improve writing practice in large content classes where writing is needed but not included. SWoRD tries to integrate writing and rewriting practice into content courses by emphasizing the role of reciprocal peer reviews rather than instructor- or expert-based reviews. This approach also raises many challenges. One of the fundamental challenges is that peer reviewers are novices in their disciplines. Therefore, to improve the impact of novices' peer reviews potential drawback, various functions such as review accuracy indices and authors' back-evaluations about the helpfulness of reviewer feedback are implemented.

As a part of learning science research, it was found that the empirical evaluations strongly support the SWoRD approach. Also, the results of the empirical evaluations imply that without increasing instructors' workload related to writing practices, student peer reviews may successfully develop writing skills in SWoRD. The perhaps surprising performance of those in the *Single-Expert* feedback condition is consistent with prior research findings that experts or instructors often generate feedback that is unhelpful or sometimes harmful to novice writers' revision (Cohen & Cavalcanti, 1987; Coupe, 1986; Schriver, 1990) because experts tend to refer to their unique knowledge that they can use but novices cannot use (Camerer, Lowenstein, & Weber, 1989). As Sperling and Freedman (1987) suggest, instructors' "written comments are often misconstrued even when they are addressed to the most promising students in otherwise successful classrooms; they are misconstrued even when they are accompanied by teacher–student conferences, by peer response groups, as well as by whole class discussion focused on response" (pp. 3–4). A caveat is that although the writing quality in the *Single-Expert* feedback condition appeared to decline across drafts, this decline might be the result of the expert using more stringent criteria for the final drafts than the first drafts. Thus it might be that in fact the writing quality in the *Single-Expert* feedback condition improved rather than declined, and instead, the improvement was just relatively smaller than that of the other conditions.

Since the successfulness of reciprocal peer review systems depends up how to support writing practice based on peer feedback, theoretical as well as practical understandings on the nature of reciprocal peer feedback will provide very important design guidelines. From the theoretical point of view, for example, it seems necessary to understand what types of feedback student peers provide and how student writers respond to peer feedback would be very helpful compared to those of expert feedback. These understandings will improve theories of learning science that guide the development of effective learning systems.

From the practical point of view, for example, it would be important to understand how many evaluators are necessary to achieve acceptable reviewer accuracy and to maximize the impact of feedback on receivers' performance improvement. Although reciprocal peer reviews recently gain increasing popularity throughout education and training (Magin, 2001), few systematic research in this area has been done to guide the design of this issue except an informally referenced strategy called *the maxima strategy*. The strategy has based on an assumption that the more reviewers would achieve higher reliability as stated in assessment theories and better improvement of writers' performance. Considering that a primary advantage of RPE systems is providing multiple peer reviewers and hence more feedback, deducing the optimal number of evaluators warrants examination. However, few empirical studies have systematically examined this issue.

7. Note

SWoRD is free to use for non-commercial purposes. It is available at http://www.lady-bug.lrdc.pitt. edu/sword. Potential users are encouraged to visit the site or contact the author(s).

Acknowledgments

This project was funded by grants from the University of Pittsburgh Office, Korean Ministry of Science and Technology, and the Andrew Mellon Foundation. We thank Jennifer Dorfield and anonymous reviewers for their comments on earlier drafts.

References

Ackerman, J. M. (1990). Reading, writing, and knowing: the role of disciplinary knowledge in comprehension and composing. National Center for the Study of Writing Technical Report 40.

Camerer, C., Lowenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: an experimental analysis. *Journal of Political Economy, 97*, 1232–1254.

Cho, K. (2004). When multi-peers give better advance than an expert: the type and impact of feedback given by students and an expert on student writing. Unpublished doctoral dissertation, University of Pittsburgh.

Cho, K., & Schunn, C. D. (2003). Validity and reliability of peer assessments with a missing data estimation technique. In *Proceedings of ED-Media 2003*, Hawaii, USA.

Cho, K., Schunn, C. D., & Lesgold, A. (2002). Comprehension monitoring and regulation in distance collaboration. In *Proceedings of cognitive science conference*, George Mason University, Virginia, USA.

Cohen, A. D., & Cavalcanti, M. C. (1987). Giving and getting feedback on composition: a comparison of teacher and student verbal report. *Evaluation and Research in Education: The Durham and Newcastle Research Review, 63–73.*

Cohen, M., & Riel, M. (1989). The effect of distant audiences on students' writing. *American Educational Research Journal, 26*, 143–159.

Coupe, N. (1986). Evaluating teachers' responses to children's writing. In J. Harris & J. Willkinson (Eds.), *Reading children's writing: A linguistic view*. London: Allen and Uniwin.

Crampton, C. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science, 12*(3), 346–371.

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection diagnosis and the strategies of revision. *College Composition and Communication, 37*(1), 16–55.

Galegher, J., & Kraut, R. E. (1996). Computer-mediated communication for intellectual teamwork: An experiment in group writing. In R. Rada (Ed.), *Groupware and authoring*. New York: Academic Press.

Hartman, K., Neuwirth, C. M., Kiesler, S., Sproull, L., Cochran, C., Palmquist, M., et al. (1995). Pattern of social interaction and learning to write: some effects of network technologies. In M. Collins & Z. L. Berge (Eds.), *Computer mediated communication and the online classroom. Higher education* (Vol. 2). Cresskill, NJ: Hampton Press Inc..

Jonassen, D. H. (Ed.). (1996). *Handbook of research for educational communications and technology*. New York: Simon & Schuster Macmillan.

Kiesler, S., & Cummings, J. N. (2002). What do we know about proximity and distance in work groups? A legacy of research. In P. Hinds & S. Kiesler (Eds.), *Distributed work*. Cambridge, MA: MIT Press.

Magin, D. (2001). Reciprocity as a source of bias in multiple peer assessment of group work. *Studies in Higher Education, 26*(1), 53–63.

Markman, E. M. (1979). Realizing that you don't understand: elementary school children's awareness of inconsistencies. *Child Development, 50*, 643–655.

McCutchen, D. (2000). Knowledge processing and working memory: implications for a theory of writing. *Educational Psychologist, 35*(1), 13–23.

NAEP. (1998). Writing report card for the Nation and the States. U.S. department of Education, National center for Education Statistics. Available from www.nces.ed.gov/nationsreportcard.

National Commission on Writing in American Schools and Colleges. (2003). The neglected R. Available from www.writingcommission.org/report.html.

Otero, J., & Kintsch, W. (1992). Failures to detect contradictions in a text: what readers believe versus what they read. *Psychological Science, 3*(4), 229–235.

Rada, R., Michailidis, A., & Wang, W. (1994). Collaborative hypermedia in a classroom setting. *Journal of Educational Multimedia and Hypermedia, 3*(1), 21–36.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: methods and data analysis* (2nd ed.). New York: McGraw-Hill.

Schriver, K. A. (1990). Evaluating text quality: the continuum from text-focused to reader-focused methods. Technical Report No. 41, National Center for the Study of Writing and Literacy.

Sperling, M., & Freedman, S. W. (1987). A good girl writes like a good girl: written response and clues to the teaching/learning process. Technical Report No. 3, National Center for the Study of Writing.

Sproull, L., & Kiesler, S. (1991). *Connections: new ways of working in the networked organization*. Cambridge, MA: MIT Press.

Vera, A. H., Kvan, T., West, R. L., & Lai, S. (1998). Expertise, collaboration and bandwidth. In *Proceeding of CHI 1998* (pp. 503–510).

Webb, N. (1995). Group collaboration in assessment: multiple objectives, processes, and outcomes. *Educational Evaluation and Policy Analysis, 17*, 239–261.

Wess, R. C. (1980). Publishing student writing: an in-class model. Paper presented at the *Annual meeting of the conference on college composition and communication*, Washington, DC. Eric reproduction services number: ED 185 564.

Zammuner, V. L. (1995). Individual and cooperative computer-writing and revising: who gets the best results? *Learning and Instruction, 5*, 101–124.

Kwangsu Cho is a Research Associate at Learning Research and Development Center, University of Pittsburgh. His current research focuses on the role of peer-to-peer computing devices in improving learning and problem solving for science and engineering, including computer supported collaborative learning and problem solving and intelligent tutoring systems.

Christian D. Schunn is a Research Scientist at the Learning Research and Development Center, and an Assistant Professor of Psychology, Cognitive Studies in Education, and Intelligent Systems at the University of Pittsburgh. His current research focuses on understanding complex forms of expertise, building models of authentic practice in science and engineering, and applying those models of expertise and authentic practice to improve science education, K-20. His past research included the development of intelligent tutoring systems, collaborative technologies, and theories of strategy selection.