The Effects of Feedback Elaboration on the Giver of Feedback

Ryan S. Wooley (rwooley@kent.edu)
Educational Technology, 207 Moulton Hall, Kent State University
Kent, OH 44242

Christopher A. Was (cwas@kent.edu)
Educational Psychology, 205 White Hall, Kent State University
Kent, OH 44242

Christian D. Schunn (schunn@pitt.edu)
Psychology, Intelligent Systems, Cognitive Studies, LRDC Rm 821, University of Pittsburgh
Pittsburgh, PA 15260

David W. Dalton (ddalton@kent.edu)
Educational Technology, 213 Moulton Hall, Kent State University
Kent, OH 44242

Abstract

This study was designed to measure the effects of varying the degree of elaboration and the use of prototypical examples during reviewing activities on reviewers' subsequent writing. The assignment was facilitated by SWoRD, an online peer review system for student writing. Results indicate that students who provided elaborate forms of feedback, which included free-form comments, performed significantly better on their own writing than students who provided numerical ratings only. In this context, the use of examples did not have significant effects on reviewers' subsequent writing quality.

Background

Writing is regarded by many as a sophisticated process of constructing knowledge. In Zinsser's (1988) words, writing is a "form of thinking" (p. vii). James Britton's (1970) publication of *Language and Learning* drew much-needed attention to the probable connection between writing and thinking. His work launched the early phases of the *Writing Across the Curriculum* (WAC) movement (Bazerman et al., 2005) which casts writing as an essential method of learning in every discipline. As writing models evolved in tandem with the early WAC movement, particularly those proposed by Flower and Hayes (1980) and Bereiter and Scardamalia (1987), they supported the notion that writing can be a recursive knowledge-construction activity.

Research outside the context of writing has supported the notion that the articulation of ideas in language leads to better learning. For example, Chi and her colleagues at the University of Pittsburgh have identified cognitive benefits associated with self-explanations of students as they solve physics problems (Chi, Bassok, Lewis, Reimann, & Glaser, 1989). Several subsequent studies of self-explanation have suggested similar correlations with learning (Aleven & Koedinger, 2002; Bielaczyc, Pirolli & Brown, 1995; Neuman & Schwarz, 1998; Schworm & Renkl, 2006).

To argue that the explanations embedded in the act of teaching are cognitively beneficial may border on the perfunctory, yet the cognitive benefits of peer teaching and cross-age tutoring have been well-supported in research since Cloward (1967) found that tenth- and eleventh-grade students with difficulty in reading and text comprehension benefitted from tutoring struggling fourth and fifth graders. Several subsequent studies produced similar results across a variety of contexts (Bargh & Schul,1980; Palincsar & Brown, 1984; Wheldall & Colmar, 1990). In fact, peer teaching and cross-age tutoring have even been supported in meta-analyses (Britz, Dixon, & McLaughlin, 1989; Cohen, Kulik, & Kulik, 1982). Leelawong et al. (2002) assert that teachers' knowledge structures are organized as much by providing feedback as by preparing materials to teach.

Research in self-explanations and reciprocal teaching has suggested that students stand to benefit cognitively by articulating explanations to self and others, but this research has been conducted primarily in math and science domains. There have been few, if any, investigations of the effects of articulating feedback for others on one's own subsequent writing. In other words, what effects might come from writing about writing? Traditional feedback research has tended to favor feedback elaborations, but such research has generally examined the receivers of feedback rather than the givers of it, and little, if any, of this research has been applied to complex tasks such as writing. Even Kulhavy and Stock (1989), authors of the early definitive feedback model, acknowledged that their model was suited more to simpler recognitive tasks rather than complex tasks. Further, results from other domains, such as the math and science problem-solving most often studied in self-explanation research, may have little application in the study of writing processes.

Peer review, a common method for engaging students in the evaluation of writing, often requires students to explain the merits or problems with peer writing. Intuitively, peer review would seem to harbor some of the benefits associated with peer teaching and cross-age tutoring. Yet, few, if any, researchers have examined the impact of articulating feedback for peers on one's subsequent writing. Little is known about the effects of treatment variations in the context of preparing students to articulate feedback, and what particular aspects of peer review might be most beneficial to students. For example, does mere exposure to peer papers benefit writers? Is it important to put students into the role of evaluator? Are there different degrees of benefit associated with different degrees of feedback elaboration?

This study was designed in two phases aimed at answering the latter two questions above. In the first phase, this investigation sought to identify the effects of reviewing on one's own subsequent writing. The hypothesis was that students who engaged in review activities would outperform students who did not. In Phase II we predicted there would be benefits associated with the use of prototypical examples and benefits associated with the use of more elaborate feedback. For the first question, participants were divided into review-first and write-first conditions. For the second question, participants from the review-first condition were divided into a rating-only group and a rating-plus-commenting group.

Method

Participants

Participants included 114 students selected from 10 sections of a sophomore-level Educational Psychology course at a large, Midwestern state university. The 10 sections of the course were distributed across four different instructors. Participants range in age from 19-44. Males composed 46% of the sample (n=53), and females composed 54% of the sample (n=61). Grade-Point-Average at the time of the study ranged from .75-4.0, with mean 3.21 and median 3.27.

Apparatus

SWoRD (Scaffolded Writing and Rewriting in the Discipline) is an asynchronous web-based reciprocal peer review system designed to facilitate student peer review of writing in content classes of any field. The system manages the submission of student papers and their subsequent distribution to a set of peer reviewers. Students using the system assume pseudonyms in order to maintain anonymity. For each assignment, students are required to engage in writing and reviewing activities across two drafts of a paper. After each draft submittal deadline has passed, papers are randomly assigned to 5-6 peer reviewers. Therefore, writers become reviewers of peer papers, and each student's paper is scored through the aggregation of peer scores with an internal proprietary algorithm. All students are graded on both their writing and reviewing. As reviewers, students are guided to evaluate writing on three dimensions: *flow*, *logic*, and *insight*. Each of these dimensions is rated by reviewers on a seven-point scale. Trait-based assessments of this type are common and have been proven effective (Bellamy, 2006). For more information about the SWoRD system and its design see Cho & Schunn (2007).

Participants were given directions through both a written instructional packet and verbal direction from the instructor These printed materials supplemented the online directions provided by the SWoRD system. The additional directions were provided as part of the experimental variations and to situate the assignment within the context of this particular course. Student review activities required access to a computer with Internet connectivity, a web browser, a word processing program capable of generating a Microsoft Word document (.doc), a Rich Text Format (.rtf) document, or a Portable Document Format (.pdf). The system is designed to guide novices through the process of submittal, and therefore does not require any special skills beyond the use of standard web browsers and simple web-based forms.

Instrumentation

Though the validity and reliability of SWoRD (student-generated) scores have been well-established (Cho, Schunn, & Wilson, 2006), expert scores were favored here due to the potential for bias in SWoRD scores created by the context of the study. Three experts were asked to evaluate student papers on the same criteria and with the same Likert-type scale used by student reviewers during the SWoRD process. Each of the experts have several years of experience in evaluating student writing in an academic setting, either at the secondary or university level. Each expert also has a minimum of a master's degree in either education or in English. Experts were asked to perform blind evaluations, by evaluating papers independently with no evidence of previous evaluations. Experts were given the same rubrics and assignment descriptions given to student participants. IntraClass Correlations (Shrout & Fleiss, 1979) were used to establish inter-rater reliability of expert ratings. Average measures ICC ranged from .766 to .794.

Procedure

The assignment for participants was to write two drafts of a 5-6 page academic position paper on an educational psychology learning theory and to submit these drafts to the SWoRD system. In addition, participants were asked to review 5-6 peer submissions of each draft. While students received writing and reviewing scores for each draft, only writing scores on the first draft are relevant to the questions of this study.

The 174 students enrolled in Educational Psychology for the Spring 2005 semester were randomly assigned into four groups using a random number table. Even though the course is naturally divided into ten sections, the researcher circumvented these naturally-formed groupings in favor of

randomly assigning treatments across the entire pool of 174 students. Therefore, students within each section received instructional packets placing them into one of four groups. The nature of the assignment offers the necessary flexibility to have students within each section of the course working on different processes at different times.

To minimize the potential for teacher effect and to ensure consistency of procedures, the researcher introduced the assignment to each of the ten sections of the class. During the 45-minute introduction, instructional packets were distributed to participants in manila envelopes. Each envelope included the assignment sheet, a list of possible topics, the specific instructional packet for a given treatment, a consent letter, and a printed copy of the SWoRD Version 3.0 Student Manual, available electronically at http://sword.lrdc.pitt.edu/getstarted.aspx.

Treatments

One of the goals of this investigation is to measure differences in writing quality between participants who engage in reviewing activities prior to writing and those who write without having reviewed peer papers. Because the review-prior-to-writing condition does not occur naturally in the standard SWoRD cycle, approximately one third of the 174 students were assigned to write drafts earlier than the other groups. The first drafts generated by this group (Group One) provided review material for the entire group (n=57) of review-first treatments. To avoid a time-of-semester confound, none of these students were part of the actual subject pool for this investigation. Therefore of the six logistical groups, only five provided data for this analysis. The remaining 114 students were divided equally between a review-first (n=55) condition and a write-first condition (n=59). This grouping forms the basis of Phase I of the study. Phase II concerns only the review-first group. The 2x2 factorial distribution of examples and feedback elaboration treatments was nested within the review-first group, breaking this larger group into three groups of n=14 and one group of n=13.

The hypotheses were tested in two phases. Phase I is a single factor design, where the independent variable is reviewing experience prior to writing with the following two levels: (a) reviewing prior to writing and (b) no reviewing prior to writing. Phase II is a completely crossed 2 x 2 factorial design, where the independent variables are: (a) use of examples to guide student review (included versus excluded) and (b) level of feedback elaboration required by reviewers (elaborate versus simple). All four treatment groups reviewed peer papers prior to writing first drafts of their own.

The dependent variable is a composite writing score derived from combining three separate writing dimension scores. Dimension scores were averaged Likert scores on a seven-point scale in the following dimensions of writing: (a) flow, (b) logic, and (c) insight. These scores were added and converted to a total writing score. The dependent variable was measured on first draft only and represents the average of three expert scores. While both composite and individual scores were available for analysis, composite scores were deemed most appropriate due to the high correlations of the separate dimension scores. Flow and logic were correlated at .861; flow and insight were correlated at .889; and logic and insight were correlated at .880.

Results

Preliminary analyses revealed that a few students had markedly low scores on the assignment. Upon examination of student artifacts for these students, it was discovered that three students had submitted assignments too incomplete to be evaluated by the dimensional criteria. For example, one student submitted a single-page bulleted list, when the assignment was to write a 5-6 page paper. Based on this discovery, three students were dropped from the analysis.

Phase I – One Factor ANOVA

The research hypothesis for Phase I was that articulation of feedback by reviewers using a web-based peer review system for writing would result in higher quality of reviewers' subsequent writing. The hypothesis was analyzed through a single factor ANOVA with writing scores as a dependent variable. Means differ by only 1.41 on a 100 point scale (review-first, M = 47.04, SD = 19.32; control, M = 48.45, SD =18.90). ANOVA results indicate no significant differences for writing scores based on this treatment: $F(1,112) = .16$, $p = .69$.

Phase II - Factorial ANOVA

Hypothesis 1 predicted that mean writing scores in the with-examples treatment group would be higher than those in the without-examples group. As shown in Table 1, mean differences were in the opposite direction from those predicted (with-examples, $M = 43.71$, $SD = 17.72$; without-examples, $M = 50.50$, $SD = 20.61$). As shown in Table 2, these differences were not statistically significant: $F(1,55) = 2.18$, $p = .14$. These results do not support the hypothesis that providing examples of prototypical helpful and unhelpful reviews would lead to higher quality of reviewers' subsequent writing.

Hypothesis 2 predicted that mean writing scores in the with-elaboration treatment group would be higher than those in the without-elaboration group. As shown in Table 1, mean differences were consistent with the predicted direction (with-elaboration, $M = 54.79$, $SD = 21.78$; without-elaboration, $M = 39.57$, $SD = 13.09$). As shown in Table 2, these differences were statistically significant: $F(1,55) = 10.21$, $p = .00$. Partial Eta^2 was .17. These results support the hypothesis that providing elaborate comments in addition to simple numeric ratings would lead to higher quality of reviewers' subsequent writing.

Table 1: Means and Standard Deviations by Factorial Cell

	With Examples	Without Examples	Total
Rating Only	*M* = 36.85 *SD* = 13.50 *n* = 14	*M* = 42.29 *SD* = 12.56 *n* = 14	*M* = 39.57 *SD* = 13.09 *n* = 28
Rating and Commenting	*M* = 50.57 *SD* = 19.19 *n* = 14	*M* = 59.34 *SD* = 24.21 *n* = 13	*M* = 54.79 *SD* = 21.78 *n* = 27
Total	*M* = 43.71 *SD* = 17.72 *n* = 28	*M* = 50.50 *SD* = 20.61 *n* = 27	*M* = 47.04 *SD* = 19.32 *n* = 55

Table 2: Factorial ANOVA –
Main Effects and Interaction for Examples and Elaboration

	MS	*df*	*F*	$P\eta^2$	*p*
Examples	692.61	1	2.18	.04	.15
Elaboration	3251.65	1	10.21	.17	.00
Interaction	38.02	1	.12	.00	.73
Error	318.39	51			
Total		55			

Discussion

Expected Results

The results of the study indicate the strongest effects were related to the elaboration treatment. Those familiar with the SWoRD system know that reviews include both numerical Likert ratings and comments. Elaboration groups were simply asked to follow this standard format, which includes commenting. Rating-only groups were asked only to rate on the Likert scale and to type the word "blank" into the comments field. The variation allowed effects of providing comments to be separated from the effects of mere exposure to peer writing, even when it includes evaluative rating. It was hypothesized that the articulation of comments on specific dimensions of writing during peer review would lead to stronger cognitive organization for reviewers which, in turn, would translate to stronger writing of their own. Results of the study support the hypothesis. Members of the elaboration groups performed better than members of the without-elaboration groups in each of the evaluated dimensions of writing. These results support the theoretical positions of Britton (1970), Zinsser (1988) and others regarding the strong correlation between articulation and thought. These results also expand explanation literature, which has tended toward examining effects of self-explanation, most often in science and math. In developing comments for peers, it seems, students are forced to organize their evaluative facilities enough to improve their own subsequent writing. In the current context, 17% of the variability was attributable to elaboration.

Unexpected Results

Review-first Versus Write-first

The two-group analysis of Phase I, comparing the writing performance of students who reviewed prior to writing with performance of those who did not review prior to writing, revealed that no differences could be attributed to the act of reviewing. At first glance, these results may seem to contradict support for the articulation hypotheses cited above. In other words, if reviewing is largely an act of articulating feedback, we would expect that students who reviewed papers prior to writing would perform better than those who did not. However, a closer look at the design of the study as well as the results from the factorial Phase II, provides some plausible explanations for the discrepancies.

First, it is important to consider that the composition of the review-first group includes each of the four groups from the 2 x 2 factorial analysis of Phase II. Recall the following treatment groups from the review-first condition: (a) with-examples and with-elaboration, (b) with-examples and without-elaboration, (c) without-examples and with-elaboration, and (d) without-examples and without-elaboration. Of these four groups, only the first group (with-examples and with-elaboration) follows the standard set of SWoRD reviewing activities that include exposure to prototypical examples of feedback as well as providing ratings and comments on peer papers. Each of the other three factorial groups nested in the larger review-first group is subtractive in at least one aspect from this first group. These three groups either forgo examples, elaboration, or both. Therefore of the 55 participants in the review-first group, only 14 engaged in the full set of standard review activities.

Given the results of the factorial analyses, it is clear that variations in review treatments have an influence on the effects of reviewing. For example, since elaboration treatments in this study showed positive effects on reviewing, the 28 members of the review-first group who did not provide comments likely lost at least some of the benefits associated with reviewing. This effect may have been compounded by the fact that examples treatments in this context were shown to have negative effects on writing, though not at a significant level. Thus, only the without examples/with elaboration group (n=13) would have seen obvious advantages from the reviewing activities. Some of the other nested groups may have even acquired disadvantages. The fact that with-elaboration groups outperformed without-elaboration groups suggested that mere exposure to peer writing was much less powerful than when it involved the articulation of comments. These findings illuminate flaws in the design of the two-group analysis of Phase I, specifically that mere exposure to peer papers was the only common bond among the nested groups. In light of the data, it is not so surprising that the review-first group (n=55) as a whole did not outperform the

write-first group (n=59). Future researchers in this area may favor a clean two-group comparison over nested treatments.

Examples

The most puzzling results from this study involve the examples treatment. Results of the factorial ANOVA of Phase II indicate that the members of the without-examples group performed better than members of the with-examples, though not at a significant level (p=.15). Recall that with regard to examples, members of the review-first group were placed in either a with-examples group or a without-examples group. The with-examples group was instructed to visit the following website: http://peerfeedback.net where they would watch a multimedia presentation about how to give good feedback on writing. Each slide contains only text, and many are quite lengthy, the longest being nearly 300 words. Students were asked to complete a reaction sheet after viewing the presentation. The reaction sheet was considered sufficient proof that students followed the treatment.

One possible explanation for results of the with-examples group is that the activity was a burdensome additional requirement embedded into a very taxing assignment. It is possible that many students were nearing a mental effort threshold with the larger SWoRD assignment when this additional activity was introduced. Up to this point, students had been asked to process and accept the following: (a) an unusual approach to managing an assignment, (b) controversial grading mechanisms which are supported by complex statistics, (c) two sets (one electronic and one printed) of very detailed and sometimes disparate sets of instructions, (d) a multi-page student manual about how SWoRD works, (e) the responsibility of accurately evaluating the work of one's peers, (f) implications for involvement in a research study, (g) interactions with a support structure and support personnel external to the class, and (h) the usual requirements of a writing assignment, such as reading research, planning, drafting. It is possible that with-examples group members may have reached limits as to how much energy they could dedicate toward this one assignment in this one particular class.

Salomon (1983) described mental effort as controlled, non-automatic elaborations applied to materials being learned. Since such effort involves choice, student motivation has a causal relationship with mental effort—a relationship sometimes informed by attributions and expectations for success in learning (Cennamo, 1989). According to Weiner (1979), students will persist in a task if they consider success or failure to be related to unstable causes, such as effort or luck, as opposed to stable causes, such as task difficulty. In other words, if a task is perceived to be so difficult that students become convinced that even effort or luck will not influence outcomes, they will not likely persist. Weiner (1979) suggests that students work best at tasks of intermediate difficulty.

Implications of Study

Results of the study clearly support the use of explanation activities requiring students to articulate their evaluative observations of peer work. The with-elaboration groups out-performed without-elaboration groups on the writing assignment, and this treatment accounts for about 17% of the total variability. As noted in the literature on self-explaining (Chi et al., 1989) teach to learn (Bargh & Schul, 1980; Cloward, 1967; Palincsar & Brown, 1984;), and write to learn (Bereiter & Scardamalia, 1987; Zinsser, 1988), the act of articulating observations through language seems to organize conceptual information in a way that is accessible to students during later activities. In this case, articulations made during review activities led to differences in writing activities that were not completed until two weeks later.

It seems, when students engage in evaluative activities normally reserved exclusively for the teacher, they stand to see cognitive benefits. Articulation marks a commitment to an observation, which, as Kelly (1963) noted, allows one to predict, test and revise conceptions. While some researchers have argued for training students in peer review (Hu, 2005) or explanation strategies (Bielaczyc, Pirolli & Brown, 1995) prior to their engagement in these activities, results from the current study suggest there are some raw benefits associated with the activity of articulating comments, even in the absence of extensive training.

Given the unexpected results in the study, specifically the lack of differences between the review-first and write-first groups of Phase I, more research is needed to determine if these results carry beyond the context of this study. For example, a follow-up two-group study without factorial treatments nested in the review-first group might provide cleaner information about an overall effect of reviewing activities. Given the inability to control for diffusion effects in the examples treatment, it might be appropriate to perform future related investigations in laboratory settings where one can exert more control over diffusion effects

References

Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based Cognitive Tutor.

Bargh, J. A. & Schul, Y. (1980). On the cognitive benefits of teaching. Journal of Educational Psychology, 72, 593-604.

Bazerman, C., Little, J., Bethel, L., Chavkin, T., Fouquette, D., & Garufis, J. (2005). Reference guide to writing across the curriculum. West Lafayette, IN: Parlor Press.

Bellamy, P. C. (2006). Research on writing with the 6+1 traits. Retrieved on July 19, 2007, from NREL website: http://www.nwrel.org/assessment/pdfGeneral/Research_On_61traits_Bellamy.pdf

Bereiter, C. & Scardamalia, M. (1987). The psychology of written composition. Hillsdale, NJ: Erlbaum.

Bielaczyc, K., Pirolli, P. L., & Brown, A. L. (1995). Training in self-explanation and self-regulation strategies:

Investigating the effects of knowledge acquisition activities on problem solving. Cognition and Instruction, 13, 221-252.

Britton, J. N. (1970). Language and learning. Coral Gables, FL: University of Miami Press.

Britz, M. W., Dixon, J., & McLaughlin, T. F. (1989). The effects of peer tutoring on mathematics performance: A recent review. B. C. Journal of Special Education, 13(1), 17-33.

Cennamo, K. S. (1989, February). Factors influencing mental effort: A theoretical overview and review of literature. Presented at the annual meeting of the Association for Educational Communication and Technology, Dallas, TX.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. Cognitive Science, 13, 145-182.

Cho, K., Schunn, C., & Wilson, R. (2006). Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives. Journal of Educational Psychology, 98(4), 891-901.

Cloward, R. D. (1967). Studies in tutoring. The Journal of Experimental Education, 36, 14-25.

Cohen, P. A., Kulik, J. A., & Kulik, C. C. (1982). Educational outcomes of tutoring: A meta-analysis of findings. American Educational Research Journal, 19(2), 237-248.

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), Cognitive processes in writing (pp. 31-50). Hillsdale, New Jersey: Erlbaum.

Hu, G. (2005). Using peer review with Chinese ESL student writers. Language Teaching Research, 9(3), 321-342.

Kelly, G. A. (1963). A theory of personality. New York: Norton.

Kulhavy, R.W., & Stock, W.A. (1989). Feedback in written instruction: The place of response certitude. Educational Psychology Review, 1(4), 279-308.

Leelawong, K., Davis, J., Vye, N., Biswas, G., Schwartz, D., Belynne, T., Katzlberger, T., & Bransford, J. (2002). The effects of feedback in supporting learning by teaching in a teachable agent environment. In P. Bell, R. Stevens, & T. Satwicz (Eds.), Keeping Learning Complex: The Proceedings of the Fifth International Conference of the Learning Sciences (ICLS) (pp. 245-252). Mahwah, NJ: Erlbaum.

Palincsar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. Cognition and Instruction, 1, 117-175.

Salomon, G. (1983). The differential investment of mental effort in learning from different sources. Educational Psychologist, 18, 42-50.

Schworm, S. & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. Computers & Education, 46 426–445.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing reliability. Psychological Bulletin, 86, 420-428.

Sweller, J., & Cooper, G. A. (1985). The use of worked examples as a substitute for problem solving in learning algebra. Cognition and Instruction, 2, 59–89.

Weiner, B. (1979). A theory of motivation for some classroom experiences. Journal of Educational Psychology, 71(1), 3-25.

Wheldall, K., & Colmar, S. (1990). Peer tutoring for low-progress readers using "pause, prompt and praise." In H. C. Foot, M. J. Morgan, & R. H. Shute (Eds.) Children Helping Children. (pp.117-134). New York: John Wiley and Sons.

Zinsser, W. K. (1988). Writing to learn. New York: Harper & Row.