The Role of Community Feedback in the Example Annotating Process: an Evaluation on AnnotEx

Independent study report
Semester: Fall 2007
Course: INFSCI 2930
School of Information Sciences
University of Pittsburgh

Submitted by: Sharon, I-Han Hsiao
Instructor: Peter L. Brusilovsky

1 Introduction and Motivation

Example based learning in programming language is a common approach when mastering the art of programming. It encourages students to reuse the code of previously analyzed examples in solving a new problem [1][2]. Gomez-Albarran [3] in a synthesis report about teaching and learning of programming stressed that example-based learning is a natural way of learning. To support online learning from examples in programming courses, WebEx (Web Examples) System was developed to provide interactive access to examples enhanced with line-by-line comments [4]. It allows students to browse the comments at their own pace and order [5][Figure 1]. NavEx (Navigation to Examples) was presented in 2004 to provide adaptive navigation support [6].

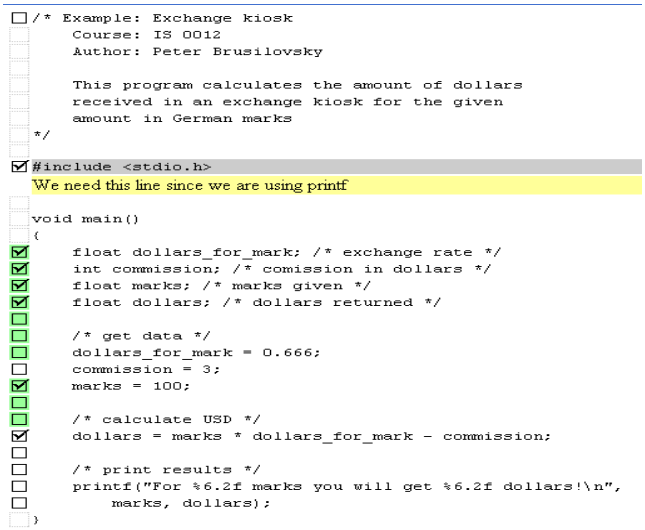


Figure 1. WebEx System

The problem addressed in this paper is that teachers usually have limited time to annotate the huge amount of examples. Examples are simply so many, but annotations are so few. In light of this problem, we explore the feasibility of an alternative authoring approach - community-based development of annotations. By harnessing students' power to create example annotations, it not only removes the burden from the instructors but also allows teachers to focus on other pedagogical tasks. In order to collect annotations from student community, we designed the system, AnnotEx (Example Annotator), as a community based authoring environment. We also conducted a pilot study to investigate this research issue [7]. The preliminary results show that community is capable to give ratings and indicate good or bad annotations.

Annotations were improved after re-annotation.

In addition, in the context of a programming course, authoring (rather than only using) examples could be considered as a useful learning activity. Jonassen and Reeves [8] contend that students are likely to learn more by constructing hypermedia instructional materials that by studying hypermedia created by others. Meanwhile, Chi et al. [9] showed that self-explanations in the context of learning about mechanics from worked-out examples had rather dramatic effects on participants' ability to solve problems on their own. Therefore, the problems we would like to look into are how the student community provides feedback and what the impacts on the example annotations are. Do students really learn from the annotating process and peer-review?

Hence, in this paper, we expand the pile of studies in two dimensions. Firstly, we design a clear-cut experiment with control and experimental groups and examine the thorough community effects. Secondly, we analyze the results and aim to answer the additional questions regarding to student learning activities.

The rest of this paper is organized as follows in order to answer the problems we are researching. In section 2, we firstly lay out the related work in collaborative example-based learning. In 3, we describe the system, AnnotEx, which is designed for catering to community-based collaborative authoring environment. In 4, the study design is presented. Main effects of the results and detail analyses will be presented in section 5. In 6, we report the further analysis in annotations and comments. In 7, we report the subjective analysis and then we summarize in 8.

2 Related Work

In order to examine the effects on community feedback on example annotations, we review two streams of work, community based peer-review and students' ability to explain examples. CPR (Calibrated Peer Review) and SWoRD (Scaffolded Writing and Rewriting in the Discipline) are two classic examples in the work of peer-review. Dr. Micki Chi has made the most exhausted research in learning activity on work-out examples. Therefore, we investigate and relate them to this study. The overall literature review in related work is presented as following.

CPR supports student learning by giving them writing assignments about important course topics [10]. Through the peer review process, students will be able to learn to read for content. At the same time, it's an exercise to develop reviewing skills. In the broader sense of education implication, perceived helpfulness is likely to mediate between the feedback and the revisions made in later writing [11]. SWoRD is a web-based peer review system. It supports the whole cycle of writing, reviews, back-reviews, and rewriting. SWoRD also examines review accuracy. It has been widely used in many courses and disciplines. The empirical evaluations of SWoRD have shown that it is effective in improving writing and helps students gain content knowledge as well as writing and reviewing skills [12].

According to Chi and her colleagues [9], students can learn a lot when attempting to explain examples. "Self-explanations," formulating the unwritten steps of an example or concept, help students understand examples and problems [9][13]. Other cognitive science studies have shown that students acquired less shallow procedural knowledge by specifically giving an explanation [14]. The benefits of generating self-explanations extend to explanations created in response to specific questions [15].

ExplaNet is a web-based learning environment where students can author and share explanations to of the questions which teachers provided. Students submit explanations and review explanations authored by their peers. Students then revise and resubmit their answers [16]. In the PhD thesis, Masters concluded that students can benefit from the process of viewing peer-authored explanations in an anonymous, asynchronous, web-based environment. The learning benefits that students receive from face-to-face peer instruction and collaboration can be extended to a virtual environment.

In our preliminary study [7], we used collaborative example authoring system to collect example annotations from students and observed the value of re-annotation based on community feedback. Students were initially assigned to annotate 2 examples. After annotating, they provided ratings and comments for 6 others' example annotations. Lastly, the low ratings group was randomly reassigned back to students. Study confirmed that community successfully filtered out good and bad annotations and the re-annotation process improved the quality of the annotations. In addition, the annotating example assignment was perceived highly helpful in understanding.

3 AnnotEx: Example Annotator System

AnnotEx, Example Annotator System (http://kt1.exp.sis.pitt.edu:8080/annotex/), was developed to support community based authoring. It allows students to author annotations to the examples as well as provide comments and ratings on the annotations. The environment creates the opportunity for students to practice collaborative authoring. The model is that students work in a group as a community. Each member from the community has three tasks to complete the example annotating process. The first task is to author the annotation of the example. The second task is to provide ratings/comments on the example annotations. The third task is to re-annotate the example annotations. AnnotEx is an online system, can be accessed anywhere through web browser with internet connection.

The AnnotEx interface [Figure 2] divides the screen into two sections. The upper section represents the tasks for the students; the lower section illustrates the example pool of the community. The tasks are sequentially arranged from left to right based on the process flow, annotating, rating/commenting and re-annotating respectively. Upon the completion of each task, s/he can continue the next task. The example pool of the community is available for all times in spite of which task s/he is doing. AnnotEx includes an evaluation prototype. Five stars rating mechanism is adopted into this system to display the evaluation in terms of quality indication. Ratings are collected from the second task. The average ratings of the example from the community will be shown on the main page.

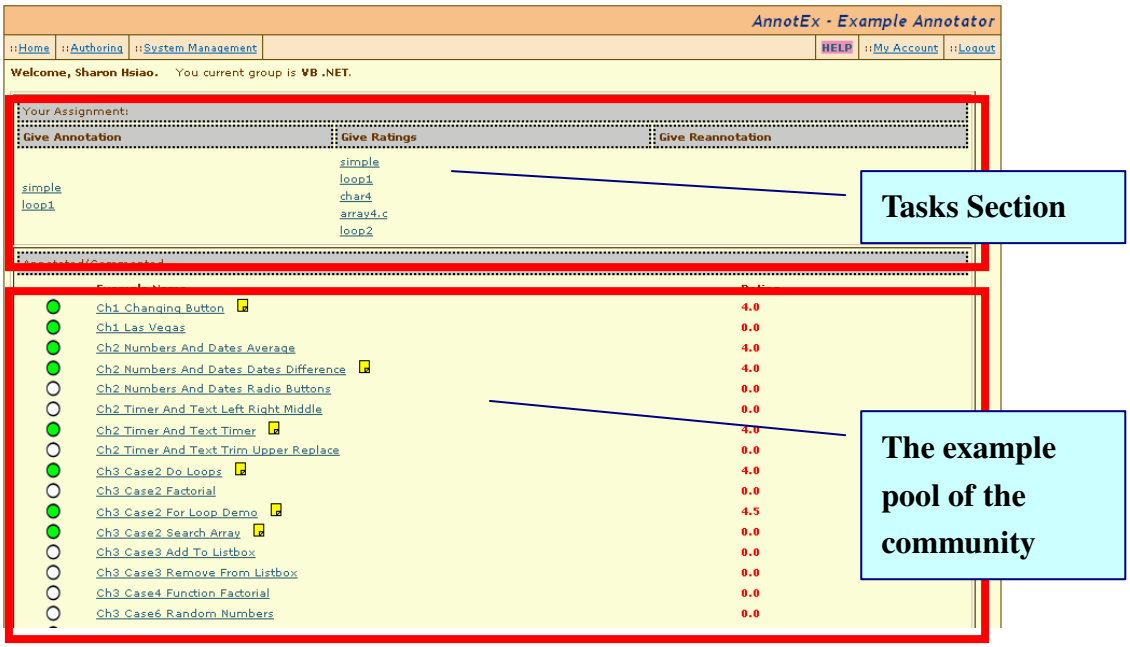


Figure 2. The main page of a community on AnnotEx

In Figure 2, the main page of a community on AnnotEx, the green circles show which examples are annotated, white ones are not. Yellow post-it icon shows comments on the annotations. Ratings are shown at the right. Figure 3 presents the first task, annotating task. The interface is divided into left and right. Left side indicates the example code line by line. Right side is the place for students to author annotations line by line correspondingly. Students can also click on the button at the top to copy the program codes. Figure 4 is the interface of second task, rating and commenting. The top of the screen is the ratings providing area. The ratings will be recorded once only through mouse over the stars and hit the submit button. There are three parts consist of the main body. Black letters on the left are the example codes; blue letters in the middle are the annotations corresponding to the examples codes line by line. The rightmost part is where students provide comments line by line accordingly. The third task, re-annotating, has the same interface as the first task.

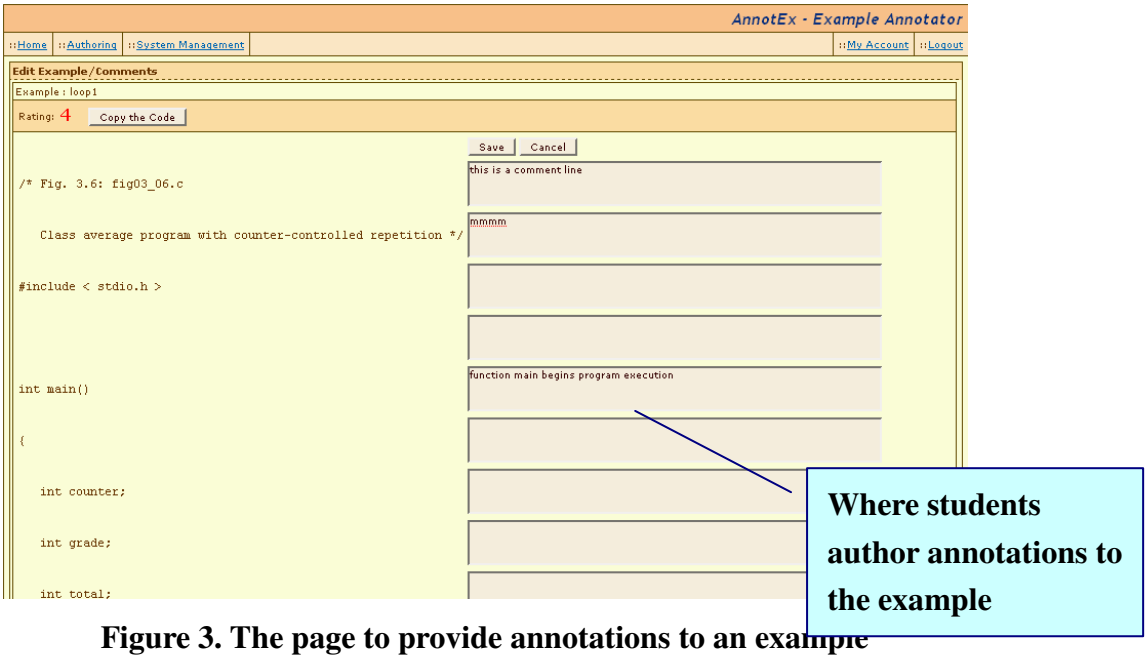


Figure 3. The page to provide annotations to an example

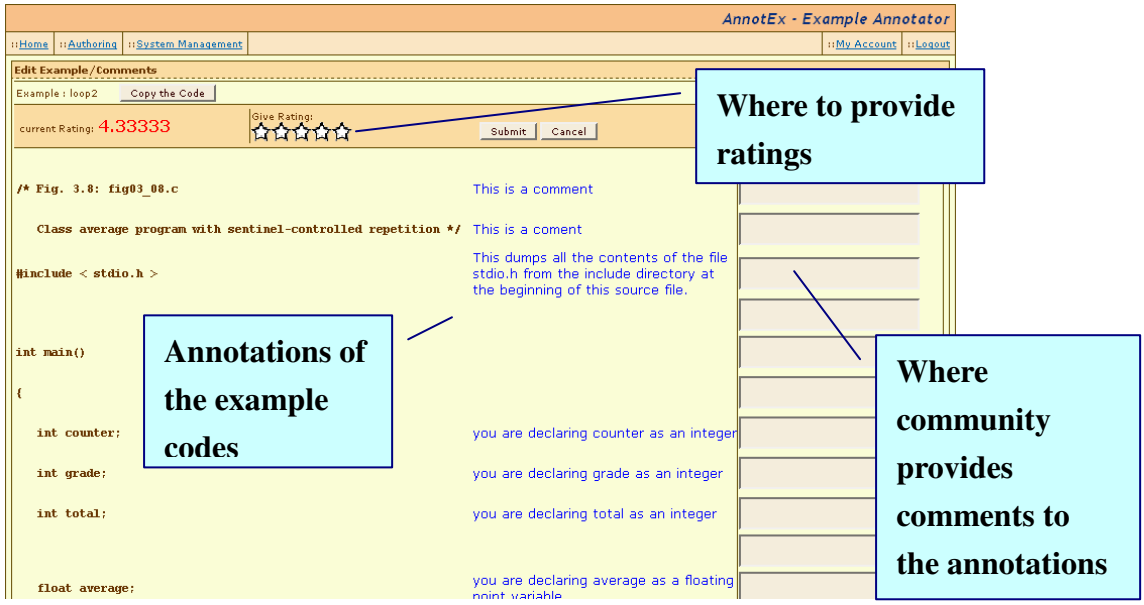


Figure 4. The page to provide ratings and comments to an example annotation

4 Study Design

The goal of this experiment was to assess the impact of community feedback and the influence of re-annotations. We aimed to investigate both the effects of re-annotations with and without community feedback. The hypothesis was that community feedback would help improve the annotation quality. Thus, the study is designed with control and experimental groups. Peer assessment and numeric ratings are commonly used to analyze the validity and reliability [17][18]. Therefore, we also used peer review

technique to examine the annotation quality, as judged by the community. Knowledge tests were also given before and after the experiment as one of the quantitative measures. The overall process flow is presented in Figure 5.

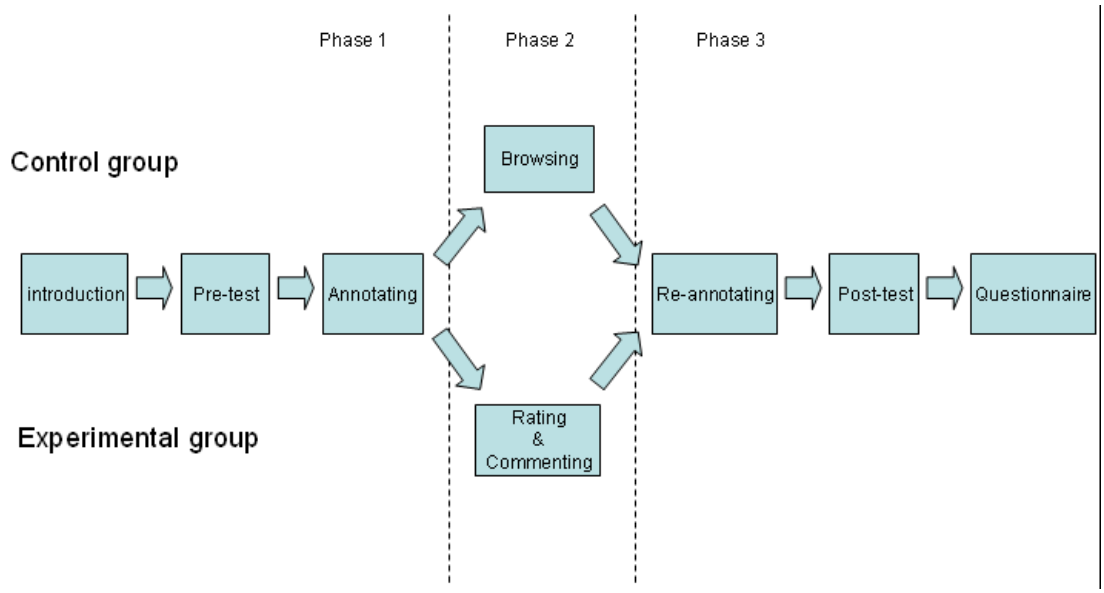


Figure 5. Study design process flow

4.1 Subjects

There are 30 subjects from National Taiwan Normal University. They are either freshman or sophomore of computer science students. Subjects were randomly divided into control and experimental groups, 15 subjects each. Each subject was rewarded a 200NTD gift card (about 6 USD) after completing the experiment.

4.2 Design

The experiment was undergone by separate introduction for each group with explanation of the experiment purpose and operation. The examples given in the experiment were topics which covered in the Introduction to Programming Language course. Examples were randomly assigned for student to annotate and to comment. This experiment lasted 90 minutes.

Pre-test: A quiz on topic Loops was given for both control and experimental groups. Question types are including answering the final value of the variables and what's the program printout.

Phase 1 (Annotating): Each student from both groups was asked to author

annotations to one example on the topic Loops.

Phase 2 (Rating and Commenting): control group was able to browse the whole examples with annotations from the community. However, they are not able to give comments or ratings to anyone of them. Experimental group could not only review the whole examples with annotations on the community basis but also specifically asked to give comments and ratings on the annotation for 6 examples per student. Ratings are scaled from 1 to 5, strongly negative to strongly positive.

Phase 3 (Re-annotating): the same example assigned at the first phase was re-assigned back for authoring re-annotations.

Post-test: A quiz on topic Loops which is similar to the pre-test quiz was given for both groups. Question types are including answering the final value of the variables and what's the program printout.

Questionnaire: 15 questions were asked for both groups based on a 5 point Likert scale (Strongly Disagree, Disagree, No Strong Opinion, Agree, Strongly agree). Free text remarks are also available.

5 Results

5.1 Dependent Measures

Annotation Rate: the ratio of annotated lines to total program lines.

Annotation Quality: the quality of annotation is measured by expert ratings.

Student Performance: the performance is measured by the pre and post knowledge test scores as the learning assessment.

Stronger student: expert ratings above 2.5.

Weaker student: expert ratings below 2.5.

Knowledge Test Score: scores is given according to the correctness, from 0 to 10. 10 means the quiz is correct.

Perceived Skillfulness: there are five categories, very bad, bad, moderate, good and very good.

5.2 Main Effects

For both control and experimental groups, the annotations collected after phase 1 and re-annotations collected after phase 3 were passed through Expert Review for quality examination. Every single given annotation was rated by two experts. They are both PhD students. One is from National Taiwan Normal University Computer Science and

Information Engineering Department with 2 years teaching assistance experiences in Introduction to Programming course. The other is from School of Information Science, University of Pittsburgh, with 6 years professional java programming experience.

First of all, based on the pre-test scores, there's no significant difference between control and experiment group before the re-annotation phase (p=0.22). This section will firstly report the overall effect to the community as a whole, and then describe the influence on students in depth. The data summary before and after re-annotation between control and experimental groups are provided in Table 1.

5.2.1 Community feedback increased annotated lines and annotation rate

According to the statistics [Table 1], both groups harness more annotation lines after re-annotation and annotating rate are increased. The raise of annotating rate contributes to the accomplishment of harnessing the explanations for the examples. Especially the experimental group, the final annotating rate is over than 50%, which suggests that there are more than half of the example codes are provided with explanations. However, without the comments as the community feedback in the control group, the annotating rate climbs up 7.25% ($p<0.1$). On the other hand, after re-annotation with community feedback interference in the experimental group, annotating rate increase 25.8% ($p<0.05$).

	Annotation		Re-annotation	
Group	**Control**	**Experimental**	**Control**	**Experimental**
Annotated Lines	8.3	9.2	11.3	18.8
Annotation Rate	27.03%	31.32%	34.28%	56.92%
Expert Ratings (all students)	1.99	2.40	2.11	3.69
Stdev(σ) of Expert Ratings	1.11	1.38	1.02	0.57

Table 1. Summary of Control and Experimental Group

5.2.2 Community feedback improves annotation quality

At the end of the study, the expert ratings increased for both two groups [Table 1]. For control group, it only increased 0.12. There are even 6 out of 15 actually scored lower than original annotation ratings [Figure 7]. The p-value (0.764) also shows the

increase is insignificant. The reasonable justification for the increase might be the additional annotations may be due to the reasons that subjects were trying to give as much annotations as they could at the third phase. For experimental group, the ratings climb up 1.29. The growth is significant ($p<0.01$). It explains that through the ability to access the community wisdom successfully promotes annotation quality in terms of higher ratings.

5.2.3 Community successfully distinguished good and bad annotations

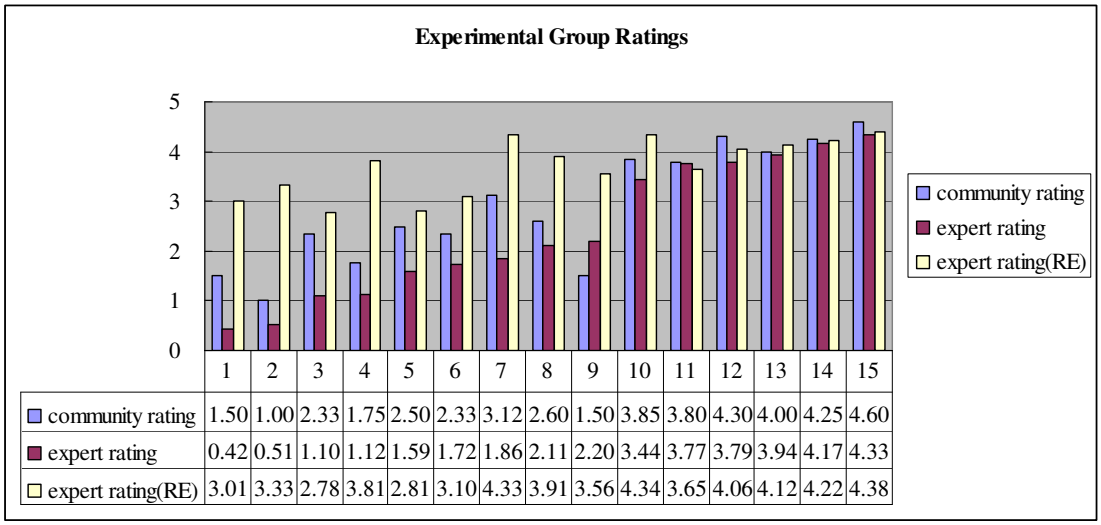
The average community ratings and average expert ratings in experimental group are respectively 3.01 and 2.40. The correlation between them is high ($r=0.93$). It indicates that community successfully distinguished good and bad annotations. It also shows community is capable to provide reasonable judgments. Figure 6 is a sorted figure based on average expert ratings. Although average expert ratings are slightly lower than community ratings, they are practically conformable.

5.2.4 Community-based re-annotation results in more coherent outcome

The ratings of experimental group after re-annotation are generally high. The standard deviation is 0.57 [Table 1], which also illustrates the coherence after community interference. On the contrary, we don't see such effect in control group. As you can see from Figure 6, student 10 to 15 performed generally well from the beginning to the end. Therefore, we focus on the rest of the pool, the weaker students. The growth of weaker students is from 1.40 to 3.40. It increased 2 points and 142.86% in total which are very substantial to the overall contribution.

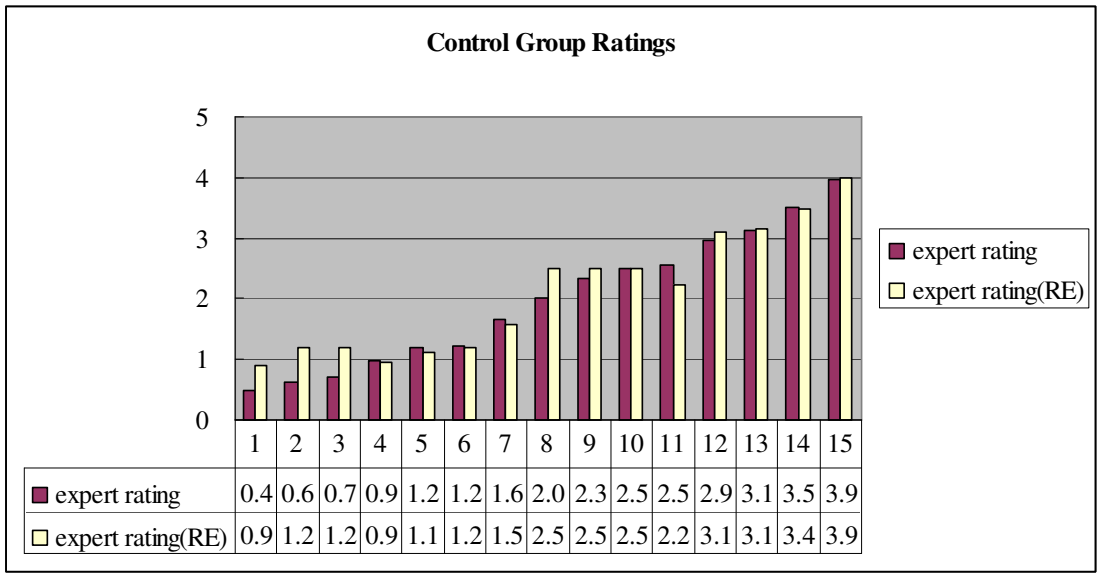
5.2.5 Good performance students help improve annotation quality

In control group, 9 out of 15 students who had good performance did actually contribute to the increase of the ratings in the end. The significance is high ($p<0.01$). It means students without the community feedback still gained knowledge after the three phases processes. In experimental group, the significance of good performance and contribution is also high ($p<0.01$). However, for control group, the overall scores decreased 4.3% in post test and for experimental group actually gained 9.9% [Table 2]. The difference once again assists in explaining that community feedback leads to the positive outcome, which influences students' learning and results.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
community rating	1.50	1.00	2.33	1.75	2.50	2.33	3.12	2.60	1.50	3.85	3.80	4.30	4.00	4.25	4.60
expert rating	0.42	0.51	1.10	1.12	1.59	1.72	1.86	2.11	2.20	3.44	3.77	3.79	3.94	4.17	4.33
expert rating(RE)	3.01	3.33	2.78	3.81	2.81	3.10	4.33	3.91	3.56	4.34	3.65	4.06	4.12	4.22	4.38

Figure 6. Each annotated example's ratings from experimental group: community rating, expert rating and after re-annotation expert's rating. This figure has been sorted by the expert rating.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
expert rating	0.4	0.6	0.7	0.9	1.2	1.2	1.6	2.0	2.3	2.5	2.5	2.9	3.1	3.5	3.9
expert rating(RE)	0.9	1.2	1.2	0.9	1.1	1.2	1.5	2.5	2.5	2.5	2.2	3.1	3.1	3.4	3.9

Figure 7. Each annotated example's ratings from control group: expert rating and after re-annotation expert's rating. This figure has been sorted by expert rating.

5.2.6 Community feedback positively affect weaker annotations

The ratings of weaker students in experimental group increased 2.0 ($p<0.01$), grew from 1.40 to 3.40 [Table 1]. We can summarize that the community feedback helps boost ratings for weaker students in the experimental group.

5.2.7 Subjective Perceived skillfulness does not provide indication to predict annotation outcome

Since community feedback positively affects weaker students, the growth for the stronger students is relatively little. If we knew this in advance, we could have left out the stronger students from the peer-review process and save some time. Thus, we expected to see whether subjective perceived skillfulness helps in predicting annotation outcome or not. In order to understand students' perceived skillfulness, they were specifically asked how good of their programming skills in the questionnaire [Table 2]. The average experience in Java are 14 months for both groups, none of them has professional experience ever. Neither of them reported as above average skillful.

We tried to correlate the perceived very bad skillful students with final ratings and with the pre/post tests scores, to see whether they contribute at all and check for the performance. Yet, no significance has been found to indicate that very bad skillful students contribute to the final increase of the ratings. In fact, both control and experimental groups have consistent results in saying that there's a negative correlation of very bad students in pre and post test, they are respectively -0.313 and -0.276. Therefore, we can only conclude that the perceived skillfulness may not be a good indicator to predict annotation outcome.

Perceived skillfulness	**Control group**	**Experimental group**
Bad	7	4
Very bad	8	11
Knowledge test scores	**Control group**	**Experimental group**
Pre-test	9.33	8.73
Post-test	8.93	9.60

Table 2. Students' Perceived Skillfulness and Knowledge tests Scores of control and experimental groups

6 Annotations & Comments Analyses

In order to find out the quality of the community comments and how it does associate with final ratings after re-annotation, the annotation lines are categorized into four types for analysis. The four types are accordingly (a) completely new annotations, (b) modified based on comments, (c) modified as exactly the same as comments and (d) the re-annotation is modified from original annotation.

Type	**(a)**	**(b)**	**(c)**	**(d)**	
	modification =new	**modification =:comments**	**modification =comments**	**modification =:original annotation**	**Sum**
Control	**4.2**			**0.8**	**5**
Experimental	**1.4**	**4.6**	**3.67**	**2.4**	**12.07**

Table 3. re-annotation types and composition statistics (average re-annotation per example)

For the control group, there are only type (a) and type (d). As you can see from Figure 8 and Table 3, in control group, 84% of the re-annotation growth mainly came from type (a). Once again, both groups were able to see everyone's example and annotations, but not the community feedback. In other words, subjects from control group were also under influence of whole community without feedback. Noteworthily, type (a) is highly correlated to the increase of the ratings (r=0.93). The increase of new annotations is primary falling in to 2 categories, declaration and block statement [Table 4]. These two categories are considered relatively simple in terms of annotating, which is less likely to be expressed wrongly. However, the improvement is shallow. Nevertheless, it explains the reason why the control group results in the trend of growth. Although the control group's growth of ratings is not significantly high as experimental group, it still suggests that the power of the community worked.

	The composition of new annotation	
Category	**Declaration**	**Block statement**
Concepts	class, type, method and variable declaration	close bracket functions
Control group	53.33%(8/15)	86.67%(13/15)

Table 4. the composition of new annotations in control group

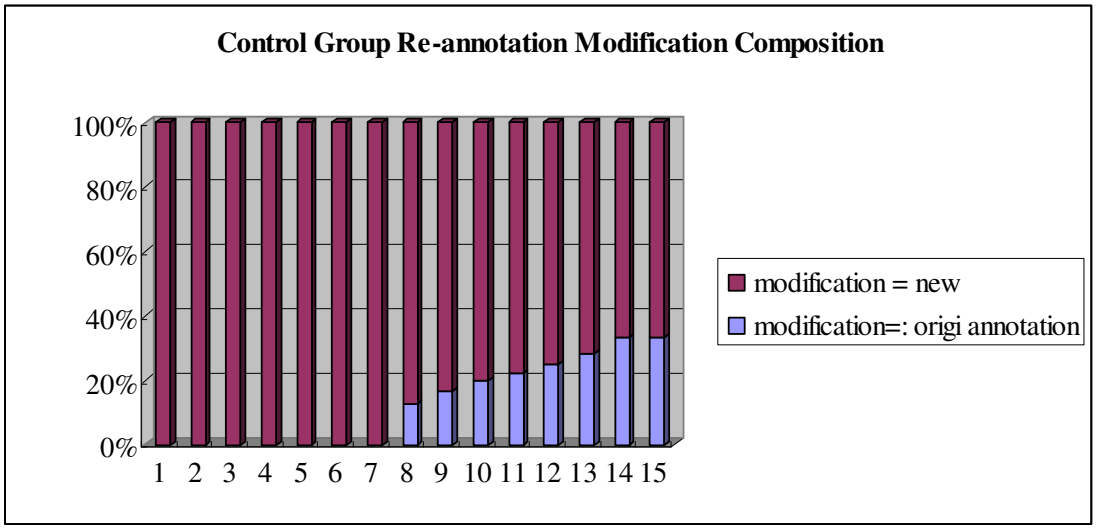


Figure 8. The composition of re-annotations of each example from control group

In experimental group, 68.5% of re-annotations were modified from community comments (type (b) and (c)). 11.6% is new (type (a)). Moreover, the re-annotations modified based on comments (type (b), (c)) is highly correlated to the increase of the ratings (r=0.94). Please refer to Figure 9 and Table 3.

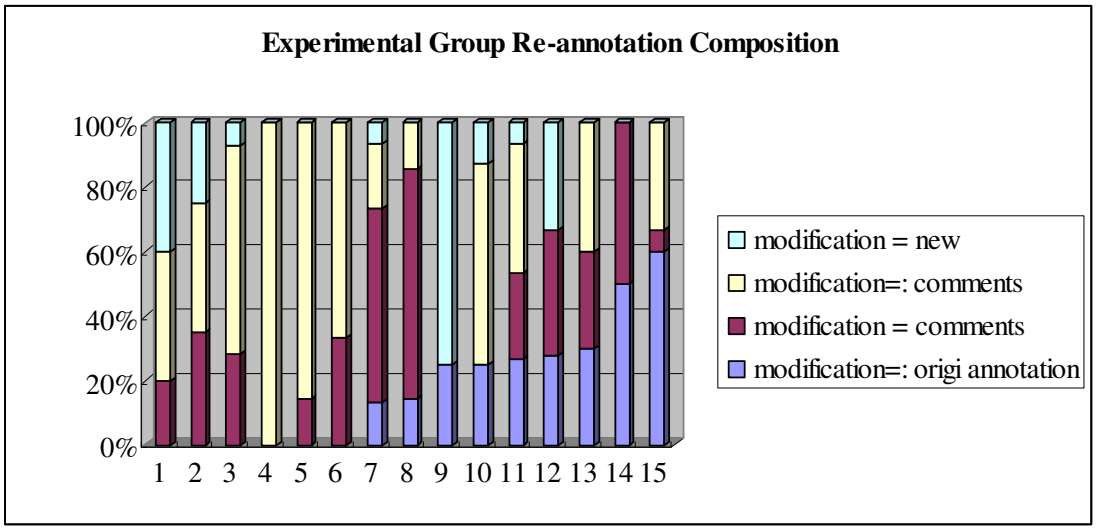


Figure 9. The composition of re-annotations of each example from experimental group

There's another interesting finding on the comments of the annotation. In our previous study [7], the task was part of the assignment of the course. During the commenting phase, students carefully checked through each program annotations. They provided as much comments as they can. However, in this study, it's no longer a grade-sensitive

assignment. Students did not specifically praise the annotation or agree to it as frequent as they were. The agreeable comments to specifically say "Good or This is correct!" are relatively fewer than the ones in our first study [Table 5].

Meanwhile, due to the experiment time limit, students were not prone to provide as much comments as they are expected to. Instead, they focused on giving comments on what have already been annotated or commented. Therefore, the similar comments were found in the end. It is also a key guidance for next phase, re-annotating, because he/she may consider the repeated points from the comments as more correct or more important. The most frequent supplemental lines provided in the comments are rather simple ones, such as closing loop annotations, end of function annotations etc. In spite of the fact that they are simple, they are still essential and sometimes very important to maintain the integrity of the annotation. It also fulfills one of our main goals which is to harness the annotations as complete as we can.

Comments/example	Study1(high rating group)	Study1(low rating group)	Study2(experimental group)
Praise/agreement	13	6.86	2.4
Supplemental annotations	0.29	8.7	19.87
Questions	0	0.72	0.07

Table 5. Comparison of previous and current studies: the composition of comments

7 Subjective Data Analysis

Opinions and suggestions on the features of the system were collected through the questionnaires after the experiment completed. As you can see from Figure 10, 86.67% of the students agree or strongly agree about the need for such tool in general, which responds to the high demand from our first study. Moreover, students also found it useful and complied with the scope of the learning activity. There're few disagreement among the questions, but there's not even a single strongly disagree point at all. Additionally, there's a section of question set designed in asking how helpful and useful of the re-annotations. 88.89% of the experimental group subjects found that community feedback is beneficial in terms of re-annotation quality and efficiency.

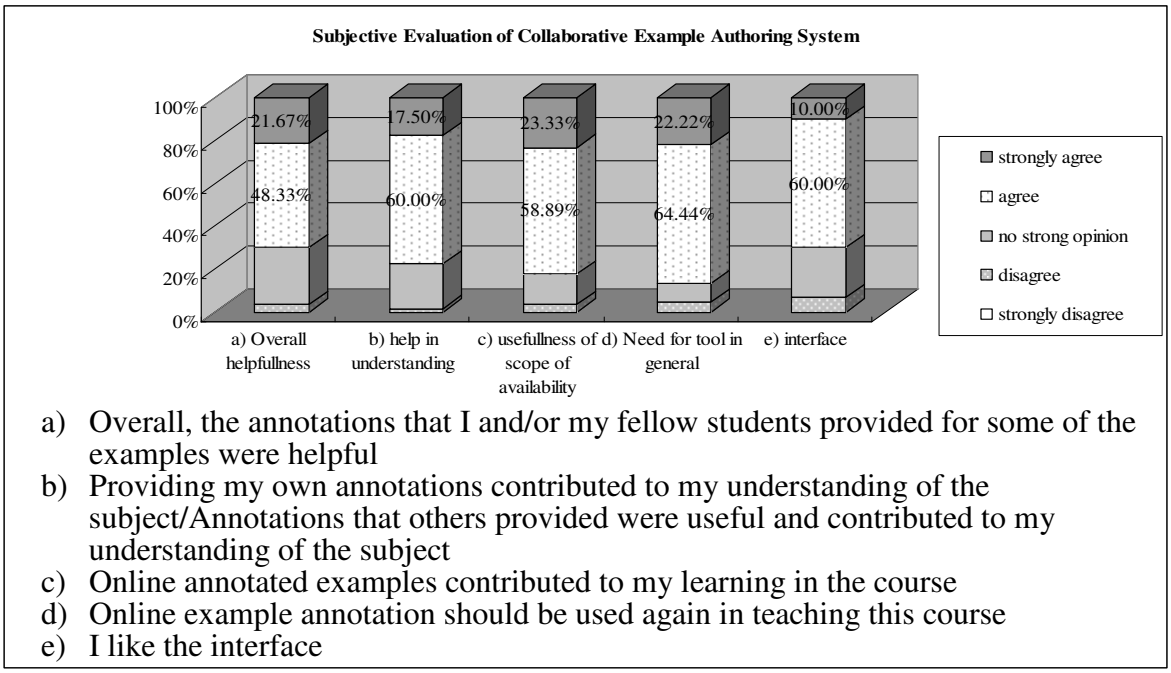


Figure10. Subjective Evaluation of AnnotEx System

8 Summary and Future Work

In this study, results show that community feedback not only increased the number of annotated lines and annotation rate, but also improved annotation quality. In addition, the ratings and comments on the annotation provide the efficient guidelines for re-annotation. It also positively affects poor students and results in a coherent good outcome. We also discovered that community feedback positively influenced weaker students and annotation outcome is hard to predict in advance. Furthermore, this study proves once more that community is capable to distinguish good and bad annotations through ratings and the comments on the annotations. The re-annotation improves the quality and helps students in understanding. Students have strong craving for the system itself. These are all consistent with our previous findings in the first study.

Since student authors are capable to provide valuable annotations within a community, we want to investigate two issues in the future. Firstly, whether students are able to create valuable examples as well as explain them. The retention after the re-annotation will be measured as well to provide more evidence in assessing learning activity. Secondly, we would like to find out how students perform according to various difficulty levels and how does the community help.

9 Reference

[1] Brna, P., Searching for Examples with a Programming Techniques Editor, Journal of Computing and Information Technology, 6 (1), 13-26, 1999.

[2] Weber, G. and Brusilovsky, P., ELM-ART: An Adaptive Versatile System for Web-based Instruction, International Journal of Artificial Intelligence in Education, 12, 351-384, 2001.

[3] Gomez-Albarron, M., The Teaching and Learning of Programming: A Survey of Supporting Software Tools, The Computer Journal, Volume 48, Issue 2, March 2005: pp. 130-144, 2005.

[4] Brusilovsky, P., WebEx: Learning from examples in a programming course. In: W. Fowler and J. Hasebrook (eds.) Proceedings of WebNet'2001, World Conference of the WWW and Internet, Orlando, FL, October 23-27, 2001, AACE, pp. 124-129, 2001.

[5] Sosnovsky, S., Brusilovsky, P., and Yudelson, M., Supporting Adaptive Hypermedia Authors with Automated Content Indexing. In: Proceedings of Second International Workshop on Authoring of Adaptive and Adaptable Educational Hypermedia at the Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH'2004), Eindhoven, the Netherlands, 2004.

[6] Brusilovsky, P., Chavan, G., and Farzan, R., Social Adaptive Navigation Support for Open Corpus Electronic Textbooks. In Nejdl, Wolfgang and De Bra, Paul (Eds.) Proceedings of the Third International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, (pp. 24-33). Eindhoven, The Netherlands: Springer LNCS 2004, 2004.

[7] Hsiao, I. & Brusilovsky, P., Collaborative Example Authoring System: The Value of Re-annotation based on Community Feedback, Proceedings of World Conference on E-Learning, E-Learn 2007,Quebec City, Canada, October 15-19, 2007.

[8] Jonassen, D. H. & Reeves, T. C., Learning with technology: Using computers as cognitive tools, Handbook of Research for Educational Communications and Technology (pp. 693-719). New York, NY: Mc Millan, 1996.

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

[10] Chapman, O., Calibrated Peer Review (CPR) White Paper, available at http://cpr.molsci.ucla.edu/ (last accessed on January 8, 2008)

[11] Rucker, M. L., & Thompson, S., Assessing student learning outcomes: An investigation of the relationship among feedback measures. College Student Journal, 37, 400-404, 2003.

[12] SWoRD: Scaffolded Writing and Reviewing in the Discipline, http://sword.lrdc.pitt.edu/ (last accessed on January 8, 2008)

[13] Recker, M., A model of self-explanation: strategies of instructional text and examples in the acquisition of programming skills. *Annual Meeting of American Educational Research Association*, Boston, MA, April, 1990.

[14] Aleven, V. & Koedinger, K., An effective metacognitive strategy: learning by doing and explaining with a computer-based Cognitive Tutor, Cognitive Science: A Multidisciplinary Journal, 26(2), 147-179, 2002.

[15] Pressley, M., Wood, E., Woloshyn, V.E., Martin, V., King, A. and Menke, D. Encouraging mindful use of prior knowledge: attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27(1):91–109, 1992.

[16] Masters, J., PhD dissertation: EXPLANET: A Learning Tool And Hybrid Recommender System For Student-authored Explanations, 2005.

[17] Cho, K., & Schunn, C. D., Validity and reliability of peer assessments with a missing data estimation technique. Proceedings of ED-Media 2003, 1511-1514, 2003.

[18] Cho, K., & Schunn, C. D., Commenting on Writing: Typology and Perceived Helpfulness of Comments from Novice Peer Reviewers and Subject Matter Experts, Written Communication, 23(3), pp 260-294, 2006.