T. Koschmann, D. Suthers, & T.W. Chan (Eds.), *Computer Supported Collaborative Learning 2005: The Next 10 Years!*
Mahwah, NJ: Lawrence Erlbaum Associates.

Design Principles for Online Peer-Evaluation: Fostering Objectivity

Yael Kali
Technion – Israel Institute of Technology
Israel
yaelk@tx.technion.ac.il

Miky Ronen
Holon Academic Institute of Technology
Israel
ronen@hait.ac.il

Abstract. Peer-evaluation is a powerful method for fostering learning in a variety of contexts. Yet challenges of application in contexts involving personal values received little attention. This study used a design-based research approach to explore such challenges in an undergraduate educational-philosophy course. The study was organized in three design-and-implementation iterations of a peer evaluation activity. Discrepancies between student and instructor scores were explained by bias due to non-objective student personal stands. Refinements to the design, based on emerging design principles a) assisted students to better differentiate between objective criteria and personal opinions, b) increased learning gains, and c) decreased tensions between different cultural groups.

Keywords: Online peer-evaluation, Design, Undergraduate Education, Educational Philosophy

INTRODUCTION

Peer-evaluation is an educational strategy in which students are required to evaluate the work of their peers. The evaluation can focus either on a learning product, or on the process. Many studies have shown that peer-evaluation is a powerful method for leveraging learning processes in a variety of contexts (e.g., Falchikov, 2003; McConnell, 2002; Suthers, Toth, & Weiner, 1997; Topping, 1998). Learning outcomes from peer-evaluation are related to: a) leveraging student understanding of evaluation criteria, and thus supporting students in creating improved artifacts, b) learning by reviewing peers' work, c) consideration of a wide range of feedback, and d) development of evaluation skills (Ronen and Langley, 2004; Zariski, 1996; Dominick et al., 1997; Miller, 2003). There is a debate concerning the legitimacy of using peer-evaluation scores as replacement of instructor's scores. In such cases, the outcomes of the peer-evaluation are usually validated by comparison with the instructor's evaluation (e.g., McGourty et al., 1997).

One of the main obstacles in the implementation of peer-evaluation is that it demands a great deal of management, organization and analysis work. Technology can provide powerful tools to reduce this workload, either by using generic online environments including forums and email (Mann, 1999), or by using targeted environments developed specifically for online peer-evaluation (e.g., Davies, 2000; Cuddy et al., 2001). Another obstacle of peer-evaluation is the issue of bias (Topping, 1998). Approaches that have been used to minimize bias in many cases are solved by anonymous evaluation. However, there is another aspect of bias that has received very little attention in the literature. This aspect, rather than being related to the *people* who are evaluated, is related to the *contents* that are being evaluated. When these contents are related to values, and are socially or culturally sensitive, designing peer-evaluation activities becomes a special challenge, and solutions such as anonymity are not sufficient to help students provide objective, non-biased evaluation to their peers' work. Our main goal in this research is to explore the challenges of peer-evaluation in a context in which personal values, morals and ethics are involved. An additional goal is to provide a set of design principles that immerge from this study, and apply to other contexts that involve similar challenges.

CONTEXT

This research took place in the context of a compulsory course in educational philosophy for undergraduate level at a university in Israel, taught by the first author of this paper. The main goal of the course is to help students develop their own perceptions about fundamental issues in education and schooling (e.g. what is the goal of schooling? What contents should be taught in school? What should be the role of the teacher?). In order to understand the social dynamics in the class it is important to note that the student population of compulsory courses in undergraduate level at that university is typically heterogeneous and includes about one third of

Jewish students who were born in Israel, one third of Jewish students who are relatively new immigrants from the former USSR and one third of Israeli Arab students (Moslem and Christian).

A main theme in the course is the "ideal school" project, in which groups of 3-4 students construct a conceptual model of a school that meets their evolving educational perceptions. Toward the end of the semester each group gives a short presentation of one day in their ideal school. For this purpose, most students use PowerPoint, but other less-conventional means, such as drama-performances were also used. The presentations took place in three class meetings, with three or four presentations in each session. One challenge we faced was how to ensure that students make the most out of these meetings. Prior teaching experience in similar contexts reveals that students tend to be focused on accomplishing the course's requirements (their own presentations in this case) and less interested in their peers' projects. This challenge was addressed by designing a peer-evaluation activity, in which students were involved in the assessment of their peers the "ideal school" presentations. The rationale for engaging students in this activity was: a) to ensure their involvement in their peers' projects, b) to create a framework for them to learn from each others' projects, c) to help them develop evaluation skills that they would need as future educators, and d) to reinforce criteria for building their products. The analysis of this peer-evaluation activity by the instructor involved the integration of hundreds of assessments (35 students, times 10 groups, times about four criteria). To help facilitate that analysis we decided to use a computerized system, which would enable gathering, presenting and analyzing these assessments in a productive manner. The activity was therefore performed online with the CeLS environment (Collaborative e-Leaning Structures), a novel system that allows the instructor to create and conduct a variety of online structured collaborative activities (http://www.mycels.net)

METHODS

In order to explore the challenges of peer-evaluation in this context we used a design-based research approach. Barab and Squire (2004) describe design-based research as: a) resulting in the production of theories on learning and teaching, b) interventionist, and involving some sort of design, c) takes place in naturalistic contexts, and d) iterative. In this spirit, the study was organized around three design-and-implementation iterations that took place in successive semesters with a total of 144 students (Iteration 1: fall 2003 with 80 students in two groups; Iteration 2: spring 2004 with 29 students; Iteration 3: fall 2004 with 35students). Each iteration was followed by data analysis and refinements to the design of the online peer-evaluation activity. Data-sources included:

- Peer-evaluation data (numeric grades and textual explanations) gathered in the CeLS environment.
- Artifacts created by each group (PowerPoint slides of the "ideal school" project and online discussions used by each of the groups for developing the conceptions for their project).
- Students' responses to an attitude questionnaire administered at the end of the course.
- Students' spontaneous online discussions in a virtual "coffee corner" at the course's site.
- Instructor's reflective journal including remarks about the events that took place during class.

The outcomes from each iteration were defined as Design Principles, according to a framework defined in the Design Principles Database (http://design-principles.org). This database is a public infrastructure funded by the National Science Foundation (NSF) and developed by the Technology Enhanced Learning in Science (TELS) center. One of the main goals in the database is to enable designers to build on the successes and failures of others rather than reinventing solutions that others have struggled to develop (Kali et al., 2004).

THE EVOLUTION OF THE DESIGN

First iteration: Initial design

The initial online peer-evaluation activity was designed according to the following design principles that were abstracted from the literature concerning peer-evaluation:
Design Principle 1: Involve students in the development of evaluation criteria
Design Principle 2: Make evaluation anonymous as possible
Design Principle 3: Use an overall global score rather than scoring individual dimensions
Design Principle 4: Use scores generated from the peer-evaluation only after validation
Design Principle 5: Minimize workload for instructors
The initial design of the peer-evaluation activity included criteria that were derived from students' suggestions in a classroom discussion that occurred prior to the presentations and included the following: a) is the uniqueness of the school apparent? b) is the rationale clear? c) are the activities that take place in the school demonstrated clearly? The activity included an online form in which students were required to grade each of the group-presentations between 1 (poor) to 7 (excellent). The form also included text fields for students to justify their grading according to the three criteria. Students used prints of these forms to take notes during the

presentations, and entered their grades and justifications to the online environment in the next few days. At the end of the activity all students were able to view a histogram of the scores for each group, statistical data (sample size, mean, median, and standard deviation), and the individual scores and the justifications for each score (presented anonymously) (figure 1). All this information was automatically generated by the CeLS environment without requiring any extra work of the instructor.

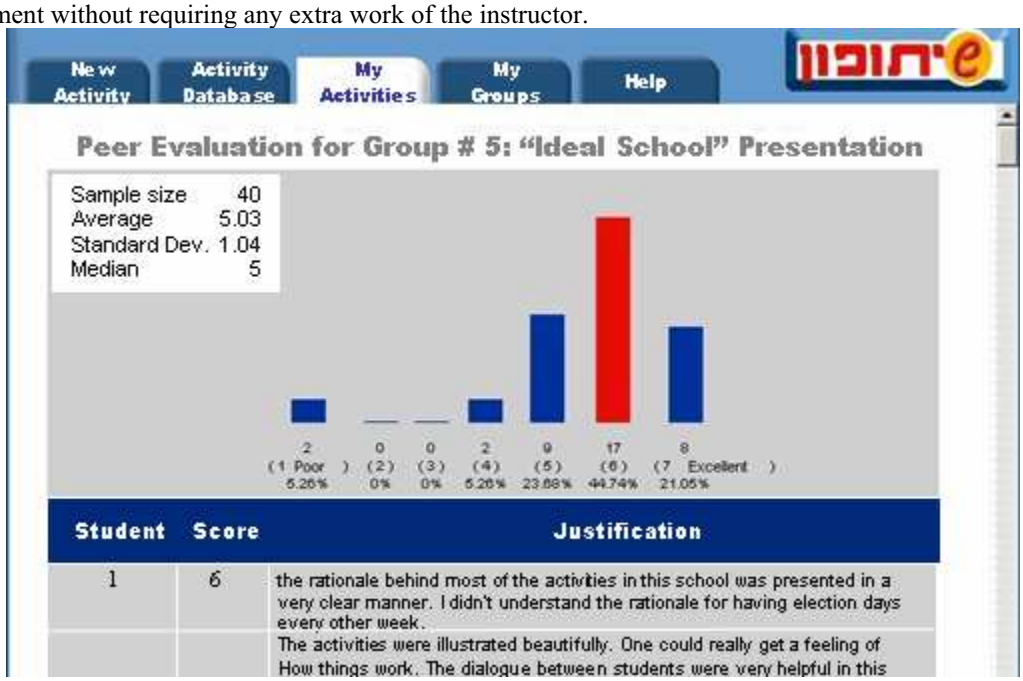

Figure 1: Interface of the peer-evaluation activity in the CeLS environment

In order to assess the validity of student scoring, the set of mean scores that were given by students for each of the 10 presentations was compared with the set of scores given by the instructor for these presentations. We refer to the instructor's grading as standard reference, and used it to validate students' grading (as in Falchikov & Goldfinch, 2000). The analysis indicated that though there was a moderate positive correlation between students' scores and the instructor's scores ($r=0.43$), it was not significant ($p=0.1$). A detailed examination of the qualitative data enabled us to identify the cases in which large discrepancies were found between students and instructor's scoring. Such discrepancies were especially apparent in presentations that introduced educational perceptions that were relatively "extreme" according to views held by many students. Though students were specifically instructed to try to ignore personal viewpoints in their grading, it seems that they found it difficult to so. An example can be seen in Figure 2. The "ideal school" presented by Group #2 was based on a somewhat existentialistic rationale; elementary students were entitled to have many choices, including the choice not to participate in any lesson. According to data analyzed from the course's online discussions, and from ideas presented in other groups' projects, most students' perceptions about schooling were more conservative. Comparison of the scores provided by the instructor, and those provided by students, shows that the largest difference was found in the scores for this presentation. The justifications that some of the students gave for lower scores, indicate that their scoring for Group #2 was biased due to their objection to the educational perception presented. For example, one student justified a low grade by saying "…students are too young at this stage and shouldn't be given such responsibilities…" Other students justified low grades by using the supposedly objective criteria, but in a biased manner. Justifications such as "the rationale wasn't at all clear" or "the activities that take place in the school weren't explained well", which were in complete contradiction with the view of the instructor and the other students, indicate that they were probably biased. In order to use the scores generated by students for grading their "ideal school" projects (15% of the final score in the course), scores that seemed biased were omitted from the statistics.

Second iteration: Differentiating between objective criteria and personal stands

Based on the outcomes of the first iteration, and in order to foster objectivity, we decided to refine the design of the online peer-evaluation activity so that it would provide students with a way to differentiate between objective aspects of the presentation and their personal, non-objective viewpoints. Our rationale was that if students would be given a chance to express these views in a neutral area, which does not affect the score, they would be more aware of their personal values and emotional stands, and thus, provide a more objective score. Therefore, we defined the following design principle and added it to the Design Principles Database:

Principle 6: Enable students to state their personal, non-objective viewpoints about their peers' work.

As in the first iteration, a class discussion about evaluation criteria preceded the activity. To engage students with the issue of personal viewpoints in peer-evaluation, we decided to seed the class-discussion with ideas for criteria, including a criterion about the degree to which a student is in agreement with views introduced in the presentation. Following the classroom discussion, four text areas for justifying scores were defined. The first three were similar to those defined in the first iteration (referring to uniqueness of the school, rationale, and demonstration of activities), but a forth area to was added, named "My personal opinion about this school". As suggested by students, this field was *not* considered a criterion that should effect scoring. Rather, it was intended to provide general feedback for presenters as to the degree of acceptance of their ideas among other students. Another design principle was therefore added it to the Design Principles Database:

Principle 7: Foster discussion about non-objective evaluation criteria

Outcomes indicate that the refined design, which enabled students to express their personal viewpoints, assisted students to better differentiate between objective criteria and personal stands. This was evident from a higher correlation between the set of scores provided by the instructor for each of the groups, and those provided by students ($r=0.62$, $p=0.03$) compared to the first iteration. Furthermore, the learning gains from the peer-evaluation activity, as indicated from the attitude questionnaire, seemed to be higher in the second iteration. This can be seen in a comparison between answers to a question regarding the extent to which students felt that the peer-evaluation activity contributed to their learning (Figure 3).

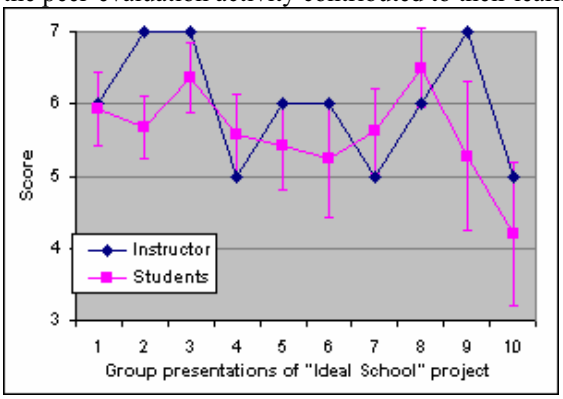

Figure 2: Comparison between scores provided by instructor and by students for each of the groups.

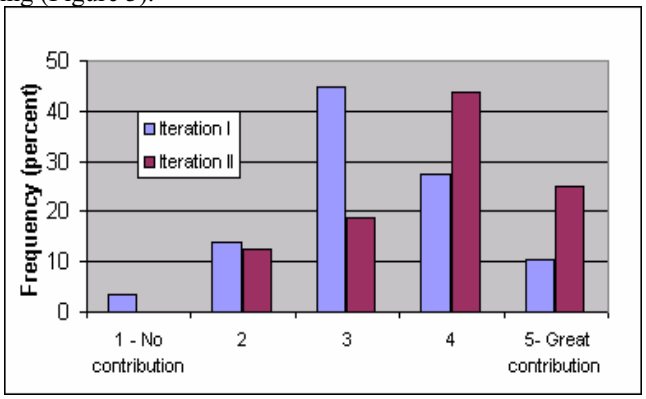

Figure 3:Distribution of student responses concerning the degree to which the activity contributed to their learning.

However, further revisions for the activity were suggested following an incident that occurred during the peer-evaluation of a certain group's presentation. The main rationale for the "ideal school" presented by that group was to bridge between religious and non-religious students in a certain cultural group. At the end of the presentation, a discussion was held between students as to whether such a school could be applied to bridging between other religious and non-religious groups. The presenters claimed that the problems that they dealt with in their school were unique. This answer, in the context of a complicated political situation in Israel, created tension in the discussion, which eventually found its way to the peer-evaluation activity, as inappropriate and even offending justifications, and biased scoring provided from a few of the students in the evaluation for that group. Following this incident, a spontaneous online discussion took place between several students and the instructor at the "coffee corner" of the course's site. In their postings, all students, no matter which sector they represented, were empathetic toward the presenters of the project, praised the quality of their presentation and criticized the biased scores and offensive justifications. They also questioned the appropriateness of the peer-evaluation activity, and discussed ideas for changing it. Students seemed to agree that the learning outcomes were tremendous, but did not like the fact that other students, who might be biased, might affect their final grade for the course. It is important to note that except for this event, the multi-cultural characteristic of the student population provided a source of richness to discussions, and to "ideal school" projects. Several of the groups were mixed (by their own choice), and introduced conceptions that fostered highly tolerant ideas.

Third iteration: Evaluating students as evaluators

Based on the findings of the second iteration, and in order to further foster objectivity, classroom norms, and tolerance, we designed the third iteration of the activity according to the following design principles.

Principle 8: Do not grade students according to peer-evaluation results.

Principle 9: Evaluate students as evaluators using results from peer-evaluation.

According to these principles, 15% of students' scores in semester fall 2000 were derived from the peer-evaluation activity and indicated how well they served as evaluators. The score was comprised of: a) number of

evaluations provided, b) respecting classroom pre-defined norms, c) quality of justifications, and d) degree of correlation with instructor's score. Outcomes indicate that implementation of the redesigned activity enabled students to better exploit the vast advantages of peer-evaluation; tensions were decreased, and higher correlation with instructor ($r=0.7$, $p=0.02$) were found.

SUMMARY

This study builds on the body of knowledge created by many studies that have designed, applied and analyzed peer-evaluation activities in a variety of contexts. We translated this knowledge into design principles and used them for designing a peer-evaluation activity for an undergraduate educational-philosophy course, taught to a multi-cultural population. Implementation in three iterations, careful analysis and tailoring of the design in a design-based research approach, enabled us to identify and confront challenges in peer-evaluation, which arouse when the evaluated contents involve personal non-objective values and morals. The following design principles emerged from this study, and apply to peer-evaluation in such contexts: a) enable students to state their personal, non-objective viewpoints about their peers' work, b) foster discussion about non-objective evaluation criteria, c) do not grade students according to peer-evaluation results, and d) evaluate students as evaluators using results from peer-evaluation. These design principles were contributed to a public online resource, the Design Principles Database, for further enhancement of the design field.

REFERENCES

Barab, S. A. & Squire, K. D. (2004). Design-Based Research: Putting Our Stake in the Ground. Journal of the Learning Sciences, 13(1), 1-14.

Cuddy, P., J. Oki, (2001). Online peer evaluation in basic pharmacology. Academic medicine, 76(5): 532-3.

Davies, P., (2000), Computerized Peer Assessment. Innovations in Education & Training International, 37(4), 346-355.

Dominick P. G., Reilly, R. R., & McGourty J. (1997). The effects of peer feedback on team member behavior. Group and Organization Management, 22, 508-520.

Falchikov, N. (2003). Involving Student in Assessment. Psychology Learning and Teaching, 3(2), 102-108.

Falchikov, N. & Goldfinch, J. (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. Review of Educational Research. 70 (3), 287-322.

Kali, Y., Spitulnik, M., and Linn M. (2004). Building Community using the Design Principles Database, Proceedings of the first joint meeting of the EARLI SIGs Instructional Design and Learning and Instruction with Computers. Tuebingen.

Mann, B. (1999). Web course management. "Post and Vote" Peer Assessment Using Generic Web Tools. Australian Educational Computing 14(1).

McConnell, D. (2002). Collaborative assessment as a learning event in Elearning environments. Proceedings of CSCL '02, January 2002, Boulder, CO.

McGourty, J. Sebastian, C., & Reilly, R. (1997). Incorporating student peer review and feedback into the assessment process. Paper presented to the Best Assessment Processes in Engineering Education: A Working Symposium, April 1997, Terre Haute, Indiana.

Miller, P. J. (2003) The Effect of Scoring Criteria Specificity on Peer and Self-assessment. Assessment & Evaluation in Higher Education, 28(4), 383 – 394

Ronen, M., & Langley, D. (2004) Scaffolding complex tasks by open online submission: Emerging patterns and profiles. Journal of Asynchronous Learning Networks (in Press).

Suthers. D.D., Toth, E.E., & Weiner, A. (1997). An integrated approach to implementing collaborative inquiry in the Classroom. Proceedings of CSCL '97, December 1997, Toronto, Ontario.

Topping, K. (1998) Peer assessment between students in colleges and universities, Review of Educational Research, 68(3), 249-276.

Zariski, A. (1996). Student peer assessment in tertiary education: Promise, perils and practice. In Abbott, J. and Willcoxson, L. (Eds), Teaching and Learning Within and Across Disciplines, p189-200. Proceedings of the 5th Annual Teaching Learning Forum, Murdoch University, February 1996. Perth: Murdoch University.