Student Evaluation in an International Collaborative Project Course

Arnold Pears, Mats Daniels, Anders Berglund
Uppsala University
Department of Computer Systems
Box 325, 751 05 Uppsala, SWEDEN
arnoldp@docs.uu.se

Carl Erickson
Grand Valley State University (GVSU)
Dept. of Comp. Sci. and Info. Sys.
Allandale, MI, USA
erickson@csis.gvsu.edu

Abstract

Grading is a frequently discussed and contentious issue. There are several views on how best to do grading and deciding how to grade students who participate in a joint international project-oriented course is far from trivial. This paper examines some in-situ observations and concerns, here referred to as myths, which arose during the project. Some statistical information extracted from the assessment data is used to examine the truth and relevance of these myths.

1 Introduction

Courses that involve multiple institutions and which span cultures and languages present a unique challenge when it comes to designing student assessment strategies. Factors include staff and student learning culture, perceptions of inequity based on the marking scales used at the participating institutes, and concerns about objectivity and parity in the application of marking schemes by staff in several locations. A key aspect of addressing such concerns is to build up trust within the teaching team, and also between the instructors and the students.

This paper describes the processes and tools used to implement assessment of a project based collaborative course offered by Uppsala University (UU) in Sweden and Grand Valley State University (GVSU) in MI, U.S.A.. The project is known as the Runestone project [2, 1, 3]. The assessment outcomes and teaching experiences gained in delivering this course provide useful insights into truths, myths, and fallacies associated with assessment of courses that span multiple cultures.

The remainder of the paper is structured as follow. Section 2 describes the course, and facilities provided to students taking the course. The assessment process and student feedback mechanisms are described in 3. Two significant issues or "myths" concerning the assessment scheme and its interaction with institutional cultures and student expectations are described and contextualized in 4. Section 5 presents statistical data collected during assessment and in section 6 the myths are revisited with the gathered data in mind. Conclusions and outcomes are presented in 7.

2 Course Context

A practical programming project forms the core activity of the course. Groups of six students (three from UU, and three from GVSU) collaborate to solve a problem involving control of a motorized marble game. In Spring 2000 there were 18 groups working on this project. Because all the groups work on the same project topic, collaboration and interaction between the groups is expected and encouraged. Apart from the purely technical goals, major aspects of the course include learning how to build up virtual working groups in an international collaboration, and using Internet communication facilities for group work.

Students are provided with course information and support hardware and software in the two course locations. Course information is provided over the WWW as text, programs, and pictures. Communication between the groups is supported by providing access to internet videoconference and chat facilities, as well as e-mail, standard file transfer and web site hosting on departmental file servers.

Communication and computer systems at both Universities are located in dedicated laboratory facilities, which are equipped with high end PC workstations, internet access, digital cameras and communication headsets. Two video conference stations were available in each location, as well as four experimental hardware setups (motorized marble games). Each Swedish student was also equipped with a Laptop computer and

an ether-net interface card which provided a portable work environment. The USA students were not provided with laptop computers.

3 Assessment

The project was both large and complex, and team progress was a major concern. The fundamental problem is how to measure and record both team and individual input towards the final product, in order to carry out fair and egalitarian assessment over the student cohort. To do this milestones (seven in total) were defined using information derived from a Gantt chart of the total project. Figure 1 shows a sample Gantt chart with progress lines representing the group's achievements at each milestone date. The group and individual achievements associated with each of the milestones are the subject of presentations to the course instructors, each roughly a week apart.

A major part of the assessment was based around these seven milestone report presentations. Half the teams reported to an instructor in the USA, and half to an instructor in Sweden. This meant that half of the Swedish students were given a grade by the USA instructor and vice versa. For each weekly milestone the teams prepared a report based on a standard template, and then used IRC chat to present the highlights of the report to the team supervisor. After each milestone meeting the supervisor prepared a feedback document for the team making comments and suggestions related to the progress towards that milestone. Two marks were allocated for each report. One mark was allocated to the individual who prepared the report (based on report quality). A second mark was allocated to the whole group assessing the manner in which the group as a whole had made progress towards that milestone, and the manner in which they managed the presentation of their work.

Each milestone meeting was conducted using a combination of IRC chat, and report documents written in html with pictures and screen snapshots. The documents for each meeting were produced by the team member responsible for that report. Live, in-lab, demos were conducted during three of the meetings. The report documents had to be available from the team web site for each weekly milestone meeting. The objectives for each milestone were clearly defined in documents provided on the course web site. The milestone reporting outline, and the topics for each report were defined at the start of the project and were based on the project Gantt chart, and the key functional aspects of the final product code.

The responsibility for preparing each of the first six reports was allocated to group members in turn. Consequently all group members had equal reporting responsibility and were required to demonstrate skills in report writing, communication using web tools, and leadership of discussions. The final presentation was a product demonstration, conducted using IRC, NetMeeting and the hardware setups. These presentations were conducted in the dedicated laboratories at each location, with full web based video conference support. The final presentation was a group effort and was conducted by the USA and Swedish team halves in collaboration. This final presentation concentrated on demonstrated functionality and presentation skills, including demonstration of international collaboration and use of the web communication tools provided throughout the course.

The Swedish and USA instructors were both present for each final presentation, and a grade for the team presentation was agreed on by both instructors at the end of the session. In this final milestone two separate grades were given. One for presentation and one for demonstrated functionality. Final group grades were based on the presentation of the final product, and the demonstration of the software functionality, as well as progress over the semester. Finally individual grades were computed using the group grade as a basis, with modifications to that grade based on peer evaluations and the grade achieved by each student for preparing their milestone report.

Peer feedback was assessed by asking each member of the group to allocate 100 USD between the members of the group on the basis of their contribution to the project. Each member was also asked to justify their allocation on the basis of technical and non-technical contributions to the overall success of the team. This information was used to adjust individual marks by up to $\pm$ 10%.

It is relevant to point out that all formal communication between instructors and students was via email or IRC, and informal contacts between an instructor and local students were intentionally minimized. The idea behind this was to help students to see IRC and other Internet tools as a primary method of communication. By doing this students were forced to become familiar with the use of the communication tools necessary for other aspects of communication throughout the course.

To minimize divergence in the application of marking during grading we defined clear guidelines on how each milestone was to be graded. Despite this the method used to apply this set of marking guidelines evolved when it was used by each instructor. In addition the tools used to give feedback to the students

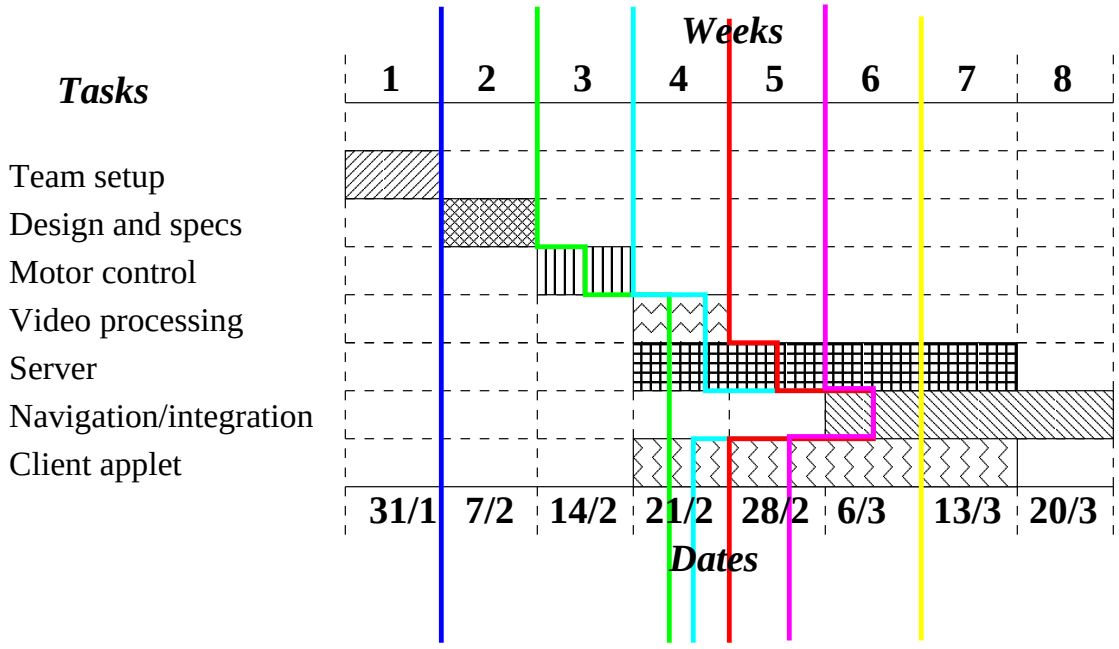

Figure 1. Sample Gant Chart

differed. Despite regular discussions on the application of the marking scheme and the outcomes of marking for the groups these disparities in the application of the scheme did not become apparent until we tried to devise a formula for computing an overall grade. At this point it became clear that the two instructors had adjusted the focus of the marking as they went along. This is discussed in more detail in section 5 in conjunction with techniques for addressing this problem.

4 Myths

Grading is an ever-present issue in educational discussions, and there are many opinions about how it should be done. Differences in opinions are especially visible in project based courses, and the international setting with the inevitable cultural differences, as well as being restricted by availability of, often rather limiting, technical means for communication, has intensified the concerns in our case. In this paper we will address two of the most significant concerns.

4.1 Myth 1

Different grading scales in USA and Sweden cause different levels of motivation in the students. The pass/fail grade structure for Swedish students mean that they don't work as hard as their USA counterparts. Statements from an email discussions include:

- "...there were quite a few Swedes who made the rational decision to work on other courses during the project, knowing theyd do well enough on the Brio project to get a passing grade."
- "This coarser final grading means that in some sense Swedish students have less incentive to work to the maximum of their ability, and this is further reinforced by their exposure to the USA grading system through the collaboration process."
- "it's not fair to the GVSU students to put them in teams with people who have different motivations because of the way grades are recorded. .. 'perception is reality', if one group perceive that their partners are not contributing as much as they should because of a different grading, then this is their reality and it could poison the Runestone waters."
- "The criticism that *think* I have understood from Swedish students (note, I am *not* sure about this) is that the US students worked more for themselves and less for the group. Some Swedes complained that *some* US individuals worked too little - but in general they were not unhappy about the volume of work done by the

Americans. My understanding for the moment is instead that they had a feeling that the US students were less eager to share their results, but wanted to show what they had done, more as individuals."

4.2 Myth 2

Grading strategies in which instructional staff in a remote location allocate marks to students they have never met in person will create an environment in which marking will be seen to be unfair.

This perception will influence the ability of staff to agree on a marking process, especially in the presence of different cultural expectations. The outcome will have a negative impact on students who will feel that marking was arbitrary and potentially unfair. The expected result of such an approach is likely to be an unhappy student cohort in each location, and significant level of dispute over the validity of awarded marks.

5 Research Data and Analysis

During the course several types of data were collected that allow further examination of the issues raised in the previous section. The discussion in this section uses the raw student evaluation data and the final course grades. These data give some insights into the actual experiences of the instructors and students. Time study results from previous course instances have also been examined.

Myth 1 relates to perceptions of work contribution and the connection of that work contribution to differences in the manner of grading students in the teams. Swedish students receive a Pass/Fail grade, while students in the USA receive marks in the range A, A-, through to F. Both students and instructors had the impression that this difference in marking schemes would have a detrimental effect on the amount of effort Swedish students would contribute to the project.

Since this topic was particularly evident in informal communication between the instructors and students it also became a topic of discussion between the educators involved in the instruction and evaluation of the course. There are two measures of effort that were collected during the course instances which can be used to shed some light on this question.

A starting point can be found in the results of time study data from previous course instances, in 1998 for example the total time devoted to work on the project per student was 120 hours for Swedish and 100 hours for USA students. The opposite result to that which we had expected.

Of course time spent on a course is not a true measure of contribution to the eventual goal. However, student perceptions of contribution gathered during peer evaluation can be used to investigate that aspect of work practices. Using the data gathered from 93 students we can compute figures which show how each group of students perceived its own contribution and the contribution of its international collaborators. This data is presented in the following table. Since each student was given 100USD to distribute over a team of 6 an even allocation of effort to all members of a team should result in 16.67USD per member. A quick examination of the average USD value for all cases convinces us that all the students perceived each other to have made an approximately equal contribution. Note that these averages are slightly above the predicted due to the fact that some groups had only 5 members.

USD allocated to Swedish and USA students averaged over the whole student cohort are presented in the first two rows of the table. If anything, an average of 17.53USD given to Swedish students overall compared to the average of 16.97USD given to USA students indicated that the student body as a whole felt that Swedes had made a marginally greater contribution. Further investigation in rows 3 and 4 shows that Swedes gave slightly more credit to their foreign counterparts(17.32USD) in contrast to the Americans (17.05), however this difference is so slight as to be nearly irrelevant. The data reveal no indication of favoritism on nationalist lines in rows 5 to 8.

While the average evaluations are consistent with a perception of equal work contribution from all students an examination of the variance and standard deviation in the data yields some other interesting information. Standard deviations show fairly narrow spread within the USD allocations, which supports the contention that most students contributed equally. The significant difference in variance in rows 1 and 2 can be attributed to the relative levels of ability in the Swedish and USA class members. Students in Sweden were perceived by the instructors to be of a more uniform level of ability than those in the USA class, and that is reflected in the higher variance in Swedish allocation of work credit to the Americans. Levels of certainty about contribution can be inferred from the variance data in rows 5 to 8 with both Swedish and American students having less variability in assessing local students (28.27 and 25.23) compared to their international classmates (30.34 and 34.61). This is consistent with the observation that Internet communication tools remain limited in permitting free communication.

The second myth concerned allocating marks to students in a course where half the students will be given

Number of US	46
Number of Swede	47
Number of students	93

Group	Average	Variance	Std Dev
Av. to Swede	17.53	20.47	4.52
Av. to US	16.97	14.04	3.75
Av. fr. Swede	17.32	29.14	5.4
Av. fr. US	17.05	29.69	5.45
Av. fr. Swede to Swede	17.66	28.27	5.32
Av. fr. US to US	16.88	25.23	5.02
Av. fr. Swede to US	16.91	30.34	5.51
Av. fr. US to Swede	17.22	34.61	5.88

Figure 2. USD Distributions (Statistical Data)

a mark by an instructor at the foreign institute. To answer this question, especially since marking processes diverged over the duration of the course, we have taken data from the grading process and divided the grades according to the nationality of the instructor. All grades were given in the range 0 to 5.

While in some categories (notably Project Functionality) the Swedish grades are lower, overall the grades allocated to teams and individuals by both instructors vary by an amount that can be attributed to normal variations in ability within the teams that were being assessed. The lower functionality ratings for Swedish teams in this study are due to greater difficulty in getting the marble boards to operate correctly in the Swedish laboratory.

As a consequence it seems likely that both instructors had a very similar view of the ability levels of students, and the intentions underlying the common grading scheme that had been devised, despite individual variability in the application of the grading scheme as the course progressed. Further evidence of student perception of fair and equitable grading is that only 5 students in total, 3 from the USA and 2 from Sweden (from a cohort of 96) made any complaint about the mark that they had been awarded. Indeed of the 3 Swedish students who failed the course, only one questioned the grade that had been awarded. Overall, this evidence seems to indicate that the majority of students accepted the marks allocated as fair, despite previous concerns that the inclusion of peer evaluation, and the use of interviews to grant progress marks, might increase student concerns about fairness in grading.

6 Myths and Perspectives

It is possible to unify grading and assessment outcomes in an international course offering, but is this a desirable outcome? Recognizing that such differences and tensions will emerge raises a question for educators. Of course we need to provide some uniformity in order to have consistent and meaningful grading. However, we should also address the issues of fairness and uniformity actively. It might perhaps be valuable to spend time and effort explaining why the world is not equal and egalitarian, and that this is a part of working with teams of people with differing agendas and interests. As a result the slight variations in hardware configurations and marking processes become a part of learning about working process.

Thus, rather than responding directly to student pressure to unify grading and other aspects of the course by making them identical, we believe that this should not be the case at all. Part of the strength of an international project is the different cultural and educational perspectives, and the modes of cultural and learning interaction between disparate classes. The more we act to make the class environments uniform, the more we rob students of the ability to experience cultural difference first hand. Is this aspect of Runestone of sufficient importance that we should ask students to suffer in order to learn the lesson? This is an ongoing and interesting question, and we are interested in other responses to this question.

Establishing and applying a grading scheme to international course offerings requires consistent communication and a high level of trust between the instructors involved. In this course trust was established by engaging in extensive collaboration in developing the grading scheme, and consistent open communication between the staff involved. While this is not always possible, we consider it an extremely important aspect of collaboration, and one that cannot be emphasized enough in cases where foreign instructors will be directly responsible for the final grade of a local student. A student that they might never meet in person.

Final Presentation		Proj. Functionality		Team Milestone		Ind. Milestone	
USA	Sweden	USA	Sweden	USA	Sweden	USA	Sweden
4.25	4.51	4.5	3.87	4.1	4.42	4.04	4.51

Figure 3. Final Assessment Task Grades

7 Conclusions

We have addressed two concerns that arose during an international collaborative project course. The first concern was that uneven work contributions would be made by students of different nationalities based on cultural factors and significant differences in the final grade between the USA and Sweden. The key point identified in this paper is that although reports and discussion seemed to indicate that the USA and Swedish grade systems would generate different levels of work input, in the end the grades for all teams and the allocations of peer effort show that participants did not feel that one side of the other had done significantly more work. The Swedish and USA team halves also agreed strongly on how the contributions had been made by their team members irrespective of physical location. In addition few complaints indicates that students accepted these assessments and the overall grading system as having been fair and rewarded effort appropriately.

The second issue was perceived differences in the application of the grading scheme by the two instructors. This was a concern to both the instructors as it transpired that their methods of assigning grades had diverged despite close and continuous communication throughout the course. In reality, the impact of such divergence was minimized by the manner in which the grading scheme had been devised. This reinforces the importance of pre-agreed grading methods and procedures. An examination of the marking data shows that variations in the grades assigned to teams by the Swedish and American instructor are no greater that those that would normally be expected due to random variation in the ability of the students allocated to each assessor. Fears that students would challenge marks that were determined using interviews and peer assessment, and that students would have less confidence in the fairness of marks allocated by a foreign instructor proved groundless. Only 5 students out of 93 made an appeal over their mark, and none of the complaints were related to unfair grading by a foreign instructor, or the methods used to obtain grades. Complaints were in fact based on students arguing that the grades did not accurately reflect the effort that they had put into the project.

Even though the concerns raised in this paper were found to be largely unfounded it is important to address them. They should be taken seriously and responses are needed in some cases, even if a myth is totally unfounded. One way is to avoid the situation, e.g. perceptions of unequal effort due to final grade scales, is to make students aware that international collaboration involves differences and change. In addition one must take care to keep the situation coupled with a follow-up to look at the actual situation. The benefit of the latter approach is that it can be used to increase the understanding across the cultures.

References

[1] M. Daniels. Runestone, an international student collaboration project. *NyIng report*, (11), 1999.

[2] M. Daniels, M. Petre, V. Almstrum, L. Asplund, C. Björkman, C. Erickson, K. Klein, and M. Last. Runestone, an international student collaboration project. *IEEE Frontiers in Education conference*, 1998.

[3] M. Last, V. Almstrum, D. M., and B. Erickson, C.and Klein. An international student/faculty collaboration: The runestone project. *ACM SIGCSE/SIGCUE Conference on Innovations and Technology in Computer Science Education (ITiCSE'00)*, 2000.