# Assessing Semester-Long Student Team Design Reports in Large Classes to Provide Individual Student Grades

*Filippo A. Salustri and W. Patrick Neumann*
Department of Mechanical and Industrial Engineering, Ryerson University
salustri@ryerson.ca, pneumann@ryerson.ca

***Abstract*** *This paper presents a method and tool to achieve a trade-off between workload on assessors of semester-long team-based design projects in large classes, with the need for fair and comprehensive assessments of each student individually. Students "book time" throughout the semester, recording their level of input into each project element. They each provide totals for time spent on each element of their final reports. The instructor assesses each design report as if one person wrote it. These data are combined into a single rubric/spreadsheet. The rubric scales report assessments to accommodate differences in team size, and generates a unique grade for each student in a team. Examples are given in the paper, as are details from the implementation of the method in a Fall 2015 introductory design course. There is anecdotal evidence that the method works, but there is always room for improvement. Several ideas for future modifications to method are discussed. All spreadsheets, documentation, and examples are freely available via the Web. Links are provided.*

***Keywords:*** engineering design, teamwork, project, assessment, individual grading.

## 1. INTRODUCTION

This paper summarizes some 15 years' effort by the authors to find an "optimal" method and tool for assessing semester-long, team-based undergraduate engineering design projects. While the historical journey through those years may be quite interesting, this paper will focus mainly on the most current grading approach, and some ideas that will be tried in 2016. The work presented here is all in the context of a 2nd year course in the Department of Mechanical and Industrial Engineering (MIE) at Ryerson University, *MEC325 - Introduction to Engineering Design*, which is mandatory for all mechanical and industrial engineering students.

Since we are talking about design projects here, the paper itself is organized along the lines of a design report.

## 2. DESIGN BRIEF

Teamwork is a skill that engineering students must learn. Accurate, fair, and comprehensive assessment of *individual students* in course projects is both important and difficult to achieve, especially in large classes with semester-long design projects.

The authors have been teaching introductory engineering design to large (200-350 student) 1st and 2nd year classes for over 15 years in Mechanical and Industrial Engineering at Ryerson University. Classes are divided into Sections of 25-30 students. All teams in a given section get the same project design brief; different sections get different projects. While the instructors try to attend every lab and tutorial, the lion's share of student assistance is provided by Teaching Assistants (TAs), whose experience and knowledge with the subject matter can vary substantially.

Students working in teams regularly complain about having to "carry" academically weaker teammates. Since teamwork itself is not a topic covered in the course, it is necessary to disconnect *team* performance as much as possible from the assessment of *individual* students' design skills.

One must also balance the workload on assessors (both TAs and instructors) against the need for accurate assessment of each student individually.

## 3. REQUIREMENTS

The following requirements are proposed for any suitable solution to the Design Brief:

- The instructor(s) must be able to grade 50-60 design reports in 10-12 days. This constraint results from enrollment and timetabling rules enforced by Ryerson University.
- The grading method must result in potentially different grades for each student in a team, in a consistent, traceable, and justifiable way. This is a requirement in response to certain features of Ryerson University's "culture" and its course management Policies.
- The grading method must allow for adjustments based on various extenuating circumstances such as

variable difficulty of different projects, variability in TA skills, etc.

- The method must help ensure "fairness" of grading; i.e., every team and every student must be graded consistently. When grading so many reports, it is possible for biases to develop over time during the grading process; these biases are to be avoided/minimized wherever possible. In particular, one would prefer the assessor to be "blind" to the team being graded and the individual students in that team.
- The method must accommodate CEAB program assessment learning objectives as implemented at Ryerson MIE, insofar as it is possible to do so.

## 4. CONCEPT DESIGN

To allow the assessor to focus on actual grading (versus data management of rubrics, individual grades, spreadsheets, etc.), and especially in light of the time constraints noted above, the method must be highly automated; thus the authors use Google Spreadsheets (under the Google Apps for Higher Education program, of which Ryerson is a member) extensively. The most important feature of "GApps" in this context is the ability for multiple users to edit single documents simultaneously and from any device while also maintaining a complete revision history. This means the assessors can work independently without having to email spreadsheets to one another, maintain multiple versions, merge different versions, etc.

The courseware is divided into fairly conventional ***modules*** on: product strategy, problem analysis, systems design, concept design, and detailed design. These units plus a few others (CAD drawings, overall report preparation, etc.) provide a natural collection of general deliverables around which students can organize their work and assessors can organize grading. Thus, lectures, assignments, the project itself, and assessment thereof all align to this same general deliverables-based structure. The authors consider this internal consistency to be an important feature of providing a meaningful educational experience for students.

Since students are expected to collaborate in a rough simulation of "real world" projects, their group reports are single, monolithic structures in which individual contributions cannot be distinguished based solely on the report content or structure. Thus, *it is only possible to grade a project report as a whole.* While this establishes a baseline grade for that team, something else must be done to distinguish each student's performance within a team and thus provide per-student assessment.

In keeping with the real-world practice of engineers "booking time" on projects, the authors look to the students themselves to provide some measure of their own contributions to each of the modules listed above. These measures are then applied to the baseline report assessment to develop grades for individual students. In particular, our approach is to use the self-reported measures to allocate a fixed number of points among team members. The overall number of points is based on the overall project report assessment. Thus, students whose self-reported measures indicate that they did not fully contribute their share of the work on the project will give up points to those team members who did more work. That is, points lost by one team member are gained by one's teammates.

Furthermore, this measure had to be reported separately for each module (as defined above) of the project, to capture the varying degree of contribution of each student to each project component. This also facilitates data gathering for the sake of CEAB program assessment.

A "classic" example of why this is necessary is the execution of CAD drawings. Typically, only one or two students in each team have particular skill and enthusiasm for using CAD software; these one or two team members will end up generating all the CAD drawings for the team's project. The instructors do not mind this, so long as the other team members are doing other pertinent project-related work. If the CAD drawings are particularly good, then the grades of only those students who developed them should benefit. Conversely, if the CAD drawings are particularly bad, then the grades of only those students who developed them should suffer.

For this approach to work, detailed instructions and support would have to be provided to students. Extensive courseware has been developed, as well as special lectures dedicated to this "project management" component of the course.

For several years, the student-reported measure that Salustri used was "effort" measured on a 3-point scale (little or no effort / average effort / significant effort). While this scale seemed to work for several years, the authors started noticing increasing problems as time went on. One typical problem was that some students reported effort expended only insofar as the final project report was concerned (rather than cumulatively over the whole semester), thus leading to artificially low grades for those students, and artificially high grades for their team mates. In some cases this lead to individual student grades of less than 0% or more than 100%.

Another typical problem was that different students would have different senses of what "effort" means regardless of the amount of support and direction provided by the instructors and TAs. It seems that, to some people, "effort" is a nearly meaningless term. Again, this leads to artificially inflated or deflated grades.

Three years ago, we decided to change the self-measure from effort to "booked hours." We believed a significant source of difficulty at that point was the qualitative nature of "effort." We hoped the more quantitative measure of actual time spent on each project module throughout the semester would reduce the problems students were having.

Over the last three years, we have found that (a) overall there were fewer instances of problems resulting from booking time than from specifying "effort," however (b) some new types of problems began to occur and (c) problems that did occur tended to be far more pronounced.

In particular, many students – especially those who tended to get higher marks generally – argued that hours worked did not account for actual ability; that is, a "good" (fast) student could do in one hour what a "bad" (slow) one might need two hours to do, thus resulting in the "bad" student accumulating more points and therefore getting a higher mark. There are many counter-arguments to this, none of which seemed to reduce the friction between team members that resulted from these arguments being made.

Another problem with booking hours was that some students were clearly over-estimating their hours. There were cases, for instance, of students booking in excess of 60 hours on their projects – or about 6 hours per week, while taking five other courses, notwithstanding the instructors' recommended target of approximately 30 hours total. Once one adjusts for commuting to/from school and other necessary activities, students spending that much time on all their subjects would have little or no time for sleep, which is highly improbable.

Clearly, booking time to distinguish contributions is also flawed, though not as badly as using "effort." The authors will be again altering this aspect of our assessment in 2016 (see Section 6 for details).

## 5. IMPLEMENTATION IN 2015

In this section, the authors provide details on the implementation of our assessment method in the Fall 2015 offering of MEC325. The implementation is described along the timeline of the 13-week semester and the three-week period after the end of classes by the end of which final grades must be submitted.

In the first lecture of the course, the nature of the project is described. Three pages of Salustri's wiki are dedicated to explaining the project management aspects of the design project: the Design Project page[1] describes the overall project and details of the expected deliverables; the Grading Team Reports page[2] describes the method by which final design reports are assessed and individual student grades generated; and the Workload Distribution Form (WDF) page[3] describes how students are to report their specific contributions to the project. Of particular importance is that the WDF represents a contract among the team members such that they agree that *all* data reported therein (i.e., the hours reported by all students on all modules) is accurate. To indicate the seriousness of the contract, the WDF must be signed by every team member to be considered valid. Reading and understanding those wiki pages is given as a reading assignment. The focus in the lecture is to impress on students the importance of recording in their design journals[4] the time they spend on each module of the project.

Booking time in 2015 was done simply by expecting students to track the time they spent on the projects, on a daily basis, in their design journals. The intention was that students would simply provide totals per project module at the end of the semester, in the WDF for their project.

In Week 2, the students are reminded of the material presented in Week 1, and any immediate questions about those points are answered. The specific steps that each student needs to take to book project time properly are reviewed.

In Week 3, design teams are announced and the projects kick-off during the regular tutorial times.

At regular intervals during the remaining 10 weeks of the semester, teams submit four project milestones. Each milestone has its own rubric. The instructors remind students that the final project will be graded as the milestones were – except that all four milestone rubrics will be used together for the final report. The instructors take these opportunities to remind students to book time properly.

Approximately in Week 8 or 9 (depending on timetabling for the particular year in question), Salustri gives a lecture covering the method by which the reports themselves are assessed and how those assessments are transformed into individual student grades. The importance of a properly completed WDF is again stressed.

During Week 13, the last week of class, student teams present their projects and submit their reports for grading. The WDF, signed by all students in each team, must be included in the team's report. By the end of that week, a softcopy of the WDF must be provided to the instructor.

Once classes are over, the instructors grade the reports. This involves applying an overall rubric and assessing the report as if a single person wrote it. A sample of the rubric for a hypothetical team is available at https://goo.gl/ptMg3m. Since the rubric is a live spreadsheet that calculates students' grades, the rubric is copied once per team. Since the rubric is a single sheet in a Google Spreadsheet, each team's rubric is stored in its own sheet in one file.

The rubric has two important areas where instructors input data.

The first is the column (column E) in which assessments on a 0-10 scale for various features of each module of the project are entered. For instance, the report's Abstract (or Executive Summary) is assessed with respect to length, problem definition, solution overview, grammar,

[1] Available at http://goo.gl/1aqmQ6.
[2] Available at http://goo.gl/f624L1.
[3] Available at http://goo.gl/COi7d1.

[4] A description of design journal expectations is available at http://goo.gl/rs5AOV.

spelling, and composition. Each module has a weight (the Element Weight in column B). Each feature within a module is individually weighted too (Column D); specifying the weights is a task done before the course starts. This way, the assessor need only enter an assessment value on a 0-10 scale for every feature of every module. The spreadsheet takes care of weighting and summing all the features of all the modules (the Weighted Score in column F). This helps the assessor focus on the hardest and most important task: assessing each part of the report accurately and consistently. The semi-automated nature of the spreadsheet also lessens the cognitive burden on the instructor and, we presume, helps eliminate bias by separating the individual assessments made from the grade ultimately generated.

The second area for assessor input is an area at the bottom left of the rubric, for WDF data. After the reports are graded, the instructor copies/pastes each team's WDF from the softcopy submitted by the team to this area of the rubric. The WDF is designed to match exactly with the corresponding rubric region so that copying and pasting is quick, easy, and not prone to error.

The instructor provides two other data: the number of students in each team, and the average team size in that section of the class. The purpose for these data is explained below. We normalize by section rather than over the entire class because different sections have different projects. The different projects can be of slightly different levels of difficulty, despite the instructors' best efforts to find equivalent projects. Since project difficulty can skew grades, we need to normalize only against similar projects to help ensure consistent grading.

Once these data are all provided, the spreadsheet produces individual student grades. This is done as follows:

1. The report's grade is calculated as the weighted sum of all the components. This is shown in cell F98 of the sample rubric.
2. Row 100 of the sample rubric reports the total weighted scores for each module of the project. These weights are the sum of the individual weighted scores (column F) for each module. These values combine the assessment of the work reported by the team with all pertinent weights.
3. For each student in a team, and for each module of the product, the number of hours reported in the WDF is normalized with respect to the total number of hours spent by all team members on the project. The normalized values are then summed for each student. This sum is reported in the CUM column of the rubric.
4. The accumulated points per student (CUM) are normalized to a z-score. This is reported in the ZC column of the rubric.
5. The cell U100 of the sample rubric contains the report average, scaled to accommodate differences in team size. We recognize that a larger team will, ceteris paribus, tend to do better than a smaller team. Given the constraints on team formation in the course, team sizes can range between 4 and 6 students. To ensure this size difference does not inappropriately penalize smaller teams, the report grade is scaled in proportion to how much larger or smaller a given team is than the average. In the example, the average team size is 5.75 while the team represented in the rubric has 6 members. Thus, the report grade is scaled back slightly to make up for the larger than average team size. This is reflected in the value of cell U100 with respect to the report grade (cell F98).
6. The z-scores for each student are then adjusted such that the average student grade is the report's size-adjusted grade, and the standard deviation of student grades is scaled according to the proportion of the raw average grade (cell S108) and the size-adjusted report grade (cell U100). This scaled standard deviation is reported in cell U109.
7. Finally, the GRD column reports the grade out of 10 to be assigned to each student, such that no student can get more than 100%.

The data in the sample spreadsheet is intentionally unrealistic and intended to "test" the calculations that the rubric performs. There are several points of interest upon which we comment below.

Student A has both worked more than any other member of the team by far, and significantly exceeded the recommended 30 hours over the entire project. Student A's raw score is 142.7%. This is clearly a problem, not with the rubric but with the team's performance.

In practice, over the last three years, there have been four cases of individual students (out of nearly 1,000 students total in that time) with grades exceeding 100% but for the artificial truncation performed by the rubric. In all cases, investigation by the instructors has led to identification of either (a) an honest data entry error by students, the resolution of which corrected the problem, or (b) some type of team dysfunction that was resolved through mediation with the instructors, and the eventual development of a new, more reasonable WDF.

Student B reported having spent virtually no time at all on the project. His grade is 1.5%, which is entirely reasonable under the circumstances. We will post this grade and leave it to the student to come to us.

This situation has happened several times in recent years; in roughly half the cases, the student raised the matter with the instructors, who investigated and resolved the situation as in the cases of grades over 100%. In the other half of cases, it turned out that the students had dropped the course.

Students C and D present an interesting situation. They worked the same number of hours on all modules, and the same total overall hours, ***except*** for the concept ideation

(PCSg) and systems design (PAS) modules. Where Student C reported 7 hours on systems design and 2 hours on ideation, Student D reported only 2 hours on systems design and 7 hours on ideation. As a result, Student C received a final grade of only 80.4% while Student D received a final grade of 96%. This is explained by the grades given by the assessor to those two modules: while ideation (range E47 to E55 in the sample rubric) was rated as perfect, the systems design module (E37 to E45) was quite weak. Since Student C dedicated more time to systems design while Student D dedicated more time to ideation, Student C received an overall lower mark because the systems module was poorly done.

*This is precisely the kind of effect we want the rubric to capture*: the instructor ***only*** assesses the work as reported by the students; the students ***only*** report the time they spent. This helps keep the instructor detached and unbiased with respect to individual student performance. The rubric combines the data without human intervention or bias. Because of this, the rubric is very "fair" in that it applies the same rules to all students uniformly. Opportunities for unfairness are narrowly restricted to (a) the assessment of specific features of the submitted work – independent of all other considerations, and (b) the hours spent by the students. In situations where students appeal their grades, it is relatively easy to justify all decisions made in the assessment process.

Student E spent relatively little time (only about 1/3 the recommended number of hours) on the project. His grade, 33.7%, is correspondingly low.

Student F worked less than expected, but still reported a respectable 22.5 hours work. His grade is 80.9%, which is quite high. This is because he booked time largely on modules that were assessed as having been quite well done. Furthermore, he scored above the team average because two of the team members scored very, very low. This is another instance of how the rubric balances the grades of students: as one student's grade drops, those of the other team members will increase.

Finally, we note that if every student in this hypothetical team had spent exactly the same amount of time on every module, then all of them would have gotten exactly the same grade: the report's size-adjusted grade of 72.3%. This is also exactly what one should expect in such a case.

## 6. FUTURE WORK

As noted in Section 4, while requiring students to book time generally worked well, it was not without its problems.

Some problems were organizational. For instance, it was very difficult to verify the hours each student reported in the WDF. (Sometimes, students disillusioned with their final project grade claimed that one or more of their teammates had lied about their hours.) To do this, we would have to find the student's design journal (in a pile of approximately 300 journals), then hunt through it for all entries containing statements of hours spent, and add up the hours per module ourselves. For small classes this is quite simple; for large classes, however, this sort of work quickly becomes utterly intractable.

The authors are working on a more automated system for booking time using Google Forms as supported by the Ryerson GApps facility. This will allow students to book time via any web-enabled device whenever they have hours to report. The WDF will be automatically generated at the end of the semester for the team. This will at least make it easier to trace a student's time to a specific place, day, and time, which should in turn make verifications easier.

Other problems with booking time are conceptual, as described in Section 4.

In late 2015, Neumann hit upon the idea of "responsibility" as a self-reported measure of student contribution. Simply put, students would "claim responsibility" for each module of their project on a coarse scale: little or no responsibility for a module; moderate responsibility; or substantive/most responsibility. The greater their responsibility, the more of that module's grade would contribute to that student's grade.

While this may seem at first largely equivalent to "effort," the authors believe there are substantive differences. We believe, and have anecdotal evidence, that students tried to quantify "effort" with respect to other related measurable parameters such as hours spent and quality of work produced (as reflected by grades). We do not think that this is possible with "responsibility." Furthermore, we believe there is a qualitative difference between "effort," which one's claim to can easily be argued implies a relative inadequacy in others who make a lesser claim, and responsibility, which seems more easily disconnected from the relevance of another claiming equal responsibility. In other words, Student 1 claiming responsibility over a task does not preclude Student 2 also claiming responsibility for the same task, nor does it imply lesser responsibility by Student 2 unless Student 2 claims it so.

The danger of using "responsibility" is that a student claiming responsibility for a module will be seen by other students as volunteering to do all the work for that module. This would clearly undermine the fundamental course goal of providing a team-based experience for students. To avoid this, there are a few possibilities. One would be to set a very low maximum number of modules for which a student can claim no responsibility; e.g., we might require that each student claim at least "some responsibility" (the middle value on the three-point scale) on at least six of the eight major project modules. Another possibility is to equate "full responsibility" with assuming a leadership role on a given module, thus partitioning the kinds of work the leader would do and leaving opportunity for other team members to contribute.

Yet another way to address the risk of misusing "responsibility" is to combine it with hours spent. That is, each student would continue to report hours spent via the WDF (hopefully through some online form-based system as described above), and also report overall responsibility on a three-point scale. The two measures would be combined automatically by the rubric.

If booking time is kept as a component of project assessment, we will also try to provide more dynamic immediate feedback to students when they actually book time. Specifically, we will use a "conditionally formatted" Google Spreadsheet to render booked time for the whole team, highlighting items of possible concern. Such items might include weeks where substantially more or less than the recommended number of hours were booked, or cases where estimated end-of-semester totals are either too low or high based on trends extrapolated from existing data.

The authors have noticed over the years that many problems of team dynamics are easily addressed by requiring a dysfunctional team to develop, agree to, and follow a "team contract" that very specifically describes how, when, and how often team members will communicate, specific procedures for resolving open issues, etc. We have noticed that this type of problem has been occurring with increased frequency, so we will institute a required team contract for every team at the outset of the semester, rather than just using contracts as remediation instruments.

Finally, we hope to pursue funding to conduct a proper study of this assessment method and its tools, for the sake of defining with more scientific certainty whether or not there benefits to students, instructors, and assessors accrue from its use.

## 7. CONCLUSIONS

The authors have presented a method for assessing student team design projects the allow for individual level performance to be identified. All material needed to replicate the method is available freely via the links provided in this paper.

We have found anecdotally that this method allows us in the role of assessors of student work to focus on grading the actual design reports, and frees us of a variety of biases and significant manual organizational and clerical work. Furthermore, compared to years past, when this method was not used, we find (again anecdotally only) that the number of student complaints and appeals has decreased. Once the method is fully and carefully explained to student, more of them seem to agree with the claim that their reports are being graded fairly.