34th Australasian Association for Engineering Education Conference
3 - 6 December 2023 – Gold Coast, QLD

# Honours project marking of seminar and expo assessments with a robust statistical approach

William S P Robertson; Wui Kei Yung; Hannah Vine Hall; Hong Gunn Chew; Dorothy Missingham

Faculty of Sciences, Engineering and Technology, The University of Adelaide

Corresponding Author Email: *will.robertson@adelaide.edu.au*

## ABSTRACT

### CONTEXT

Marking of capstone project oral presentations (in either seminar or expo formats) is often performed by panels of assessors with varying discipline and/or academic backgrounds. When multiple panels are used to mark large cohorts of students, these variations can lead to inconsistencies in the marking processes and can lead to a lack of confidence in the marks awarded. At The University of Adelaide we have developed an approach which uses robust statistical analysis to automatically account for systematic and random biases in assessment.

### PURPOSE

To provide effective student feedback, we believe that transparency in assessment is the best way to ensure consistency and robustness. Compared to written assessments, marking of communication-based assessments can have stronger elements of subjectivity, unconscious bias, and be influenced by the closeness (or lack thereof) of the expertise of the assessor with the discipline of the project being assessed. We aimed to develop a robust approach to processing marks for these assessments that required minimal intervention from the coordination team while automatically addressing discrepancies between assessors where possible.

### APPROACH

During the presentation events, marks were input and collated with an electronic survey tool such as SurveyMonkey, Google Forms, or Microsoft Forms (ideally 4+ assessors per group). These marks were post-processed in Microsoft Excel into a format that allows each set of marks awarded to each project group to be analysed while also considering each set of marks awarded by each individual assessor. An 'assessor weighting' was calculated that corrects for systematic differences between assessors. Within each project group, a robust statistical calculation (location and scale, instead of mean and standard deviation) was performed that automatically mitigated the effects of outlier marks, and if desired could reject outlier marks entirely.

### OUTCOMES

Large-scale processing of seminar and expo marks is now possible, with 50+ assessors marking 100+ project groups. Feedback from students was positive, with only one student group raising concerns about the marking process, where an outlier 'high mark' was rejected. Such edge cases may still require manual consideration.

### CONCLUSION

It is challenging to build an assessment process that is transparent, scalable, and fair. The statistical approach presented has been an effective solution to this problem.

### KEYWORDS

Honours project, Assessment, Robust statistics

## Introduction

The University of Adelaide undertook a major restructure of its engineering curriculum in 2019–2022, to replace twenty-two named programs with nine named programs which include associated majors. The new program structure included a 'mostly common' first year and a professional practice core throughout levels one and three, leading to the final year honours course ENG 4001 Honours Research Project. The new cross-disciplinary honours project course includes students across most engineering disciplines, and was designed to improve efficiency of management logistics and processes, maintenance and delivery of teaching material, and consistency of assessment structure and marking schemes. Prior to this change, each school managed their own individual honours project course.

The new Research Project course is undertaken over two consecutive semesters (25% load). It involves approximately 350 students per year, the majority starting in Semester 1 and a smaller group commencing their project in Semester 2. Alongside project management and written report deliverables, there are two verbal communication-based assessments: concluding Semester 1 is a two-day seminar event, intended to be a semi-formal event to a technical audience with multiple parallel sessions; concluding Semester 2 is the two-day expo event 'Ingenuity', which is open to the public and doubles as both outreach to high school students as well as to local business and industry. Both seminar and Ingenuity presentations are marked via a panel of assessors. In the last ten years, we have shifted from onerous and error-prone paper-based assessment to electronic means of assessment, and added sophistication to the methods used to calculate the final mark awarded to each student.

This paper discusses the general approach used for assessing students, then presents the methods used to account for systematic and random errors when collating marks across a disparate group of assessors. A general discussion of the benefits and limitations of these methods follows.

## Assessment tasks

Research projects are undertaken by students in groups of one to six. For the seminar assessment, students present in their groups with part of the marks awarded towards the group as a whole and the remainder of the marks assigned individually. Groups present for around 10 to 30 minutes depending on the number of students. One chairperson and up to four academics are invited to assess each session, with a teaching assistant present to aid with technical issues and provide another assessment data point. Each assessor marks with a rubric with three whole-group criteria; project plan, engineering content, and outcomes; and two criteria for each individual; preparation and verbal/non-verbal communication. Each criteria could receive an integer mark between zero and ten.

Ingenuity assessment is performed via face-to-face discussions with assessors from industry and academia; these are awarded to the group as a whole to keep the assessment process as streamlined as possible. Representatives from industry are invited and randomly allocated around eight projects to assess over two to three hours at the event. The random allocation is spread across all engineering disciplines to ensure that a variety of projects are seen by each assessor.

For both seminar and Ingenuity events, marks are input by assessors in an online survey created in Google Forms. Likert-style rubrics are used, with a single free-text comment field for general feedback. To minimise the chance of erroneous input by the assessors, custom forms are created using the Google Forms API. Drop-down boxes and hard-coded student names are used to streamline the process for assessors as much as possible. As the marks are input, they are downloaded via Google Forms as a series of CSV files for subsequent data analysis. For the seminar assessment this is done after the event's conclusion, however for the Ingenuity event, this is done in semi-real-time as prizes are awarded at the end of the event to the top projects. Continuous review of marks allows student groups who have not yet been marked by enough assessors to be targeted for additional assessment.

The median number of marks received by each group are four for the seminar and six for Ingenuity, with variances due to factors such as conflict-of-interest, late-minute unavailability, and statistical sampling differences.

## Key assessment issues

The goal of these assessments is a process that is as transparent and fair as possible for students. However, from analysis of marks and student feedback from previous events, it is clear that discrepancies between assessors occur. While a simple Likert-scale rubric was used to address these issues, much like methods discussed by Littlefair and Gossman (2008) and Henderson et al. (2009) for similar assessments, student feedback and analysis of marks from previous events highlight that use of a rubric alone is insufficient for presentation assessments. This is likely due to the time-pressures during marking (Kim 2014) and subjective interpretations of the rubric criteria. As such, it is clear that when collating a set of marks from multiple assessors for assessments in this format, systematic and random differences arise from assessor discrepancies that must be managed through statistics rather than purely assessment design. Although we have not retrospectively analysed the rate at which these discrepancies substantially affected student grades, our experience suggests that subjective inconsistencies can lead students to perceive the assessment as unfair or poorly managed. Attempts to manage this in the past have included: not releasing individual marks to students, manual review and screening of marks, and simpler versions of the statistical methods discussed following.

The first stage of the data analysis is intended to compensate for systematic differences between assessors caused by different interpretations of the marking rubric and biases towards certain oral communication styles (Henderson et al. 2009; Kim 2013; Kim 2014). If all assessors were assigning marks to the same set of students, this would be a reasonably simple task, since the mean mark awarded by each assessor could be normalised to the mean of all marks. However, each assessor awards marks to a sparse set of groups, which only overlaps in part with the groups assessed by others. In this case it is likely that the average quality seen by each assessor may not be consistent. In other words, if one session of talks happens to be of higher quality than another session, we do not wish to penalise them in an effort to normalise the overall distributions of marks awarded. With sufficiently experienced assessors and appropriate rubrics, we should expect the variance between assessors to be relatively small and that only a minority of assessors should need substantial scaling.

The second stage of data analysis is intended to identify and remove outlier marks. Outliers are associated with a number of factors: errors in data entry; unidentified conflicts of interest; very close or very far discipline knowledge that heavily rewards or penalises students, depending on the context (Kim 2013; Kim 2014). Prior to the introduction of automatic outlier detection, we instructed academics not to assess their own students, as we found their marks were sometimes biased. This led to other problems, such as not having enough assessors in a session, and assessors marking their own students anyway. Allowing any assessor to mark any student has simplified the logistics of assessment.

Despite attempting to adjust for certain factors that influence the marks, as discussed above, there are other factors which we hypothesise can impact the marks awarded. These factors include, but are not limited to: the degree of familiarity and/or appreciation the assessor has with the subdiscipline of the project; any prior exposure the assessor has with other projects from the same supervisor or lab group; and within-session factors such as a mediocre project following an excellent one; assessor fatigue, with a higher chance of inconsistent marks after a long day of assessment. While some of these factors may be indirectly remedied by assessor scaling or automatic outlier detection, they are not directly addressed in the following solution; it is unlikely that any quantitative solution to remedy all such factors is possible.

## Methodology of marks processing

This section will outline the mathematical detail of the data analyses we have developed. These analyses use robust statistical methods (Maronna et al. 2018) to estimate the central measure of data and data spread around that measure, *location* and *scale* respectively. Such methods are more robust for small sets than using common measures such as *mean* and *standard deviation*, potentially making them more suitable for this application. The subsections to follow will cover: an introduction to the theory; an outline of the robust statistics used; the approach used for assessor marks weighting; and, the approach used for outlier detection.

## Introductory theory

When collating a set of marks $M = \{m_1, \ldots, m_n\}$ from multiple assessors, the common approach to calculate the awarded mark $\bar{m}_i$ for student $i$ would be using the mean $\mu$ of the marks awarded by each assessor $j$:

$$\bar{m}_i = \mu = \mathrm{mean}(M) \equiv \operatorname*{mean}_j(m_j) = \frac{1}{n}\sum_{j=1}^{n} m_j \,. \tag{1}$$

The spread of the marks around the mean value is often represented by the standard deviation $\sigma$:

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(m_j - \mu)^2}\,, \tag{2}$$

These calculations are not considered 'robust' in the sense that an outlier may skew the results substantially (mathematically, outlier mark $m_o \to \infty$ results in $\mu \to \infty$). Similarly, the standard deviation grows large in the presence of outliers. This makes the standard deviation problematic if it were being used to attempt to identify outlier terms.

Two common metrics are used to replace mean and standard deviation when robustness is required; these are the *median* and the *median absolute deviation* ($\mathrm{MAD}$): (where $|\cdot|$ is the absolute value function)

$$\mathrm{MAD}(M) = \mathrm{median}(|M - \mathrm{median}(M)|)\,. \tag{3}$$

While robust, these measures have unusual behaviour when applied to a small number of marks. For instance, the set of marks $M = \{5, 6, 6, 6, 7\}$ has a median of $6$ and $|M - \mathrm{median}(M)| = \{0, 0, 0, 1, 1\}$. Therefore, $\mathrm{MAD}(M) = 0$, which leads to subsequent problems in the numerical calculations. In this work, we have found that an alternative to $\mathrm{MAD}$ is more appropriate, the *average absolute deviation* ($\mathrm{AAD}$):

$$\mathrm{AAD}(M) = \mathrm{mean}(|M - \mathrm{median}(M)|)\,. \tag{4}$$

Although theoretically affected by outliers, we find that $\mathrm{AAD}$ provides a better initial estimate of the spread since in practice marks awarded by assessors are always bound to a finite range. (In the example from the previous paragraph, $\mathrm{AAD}(M) = 0.4$.) Other alternatives are possible and should be explored in the future (Rousseeuw and Croux 1993).

## Robust statistics

The approach described here has been adapted from the work of Rousseeuw and Verboven (2002), whose work on analysing very small sample sizes was instrumental in identifying an appropriately robust methodology.

The *location* $T_n$ of a set of marks $M = \{m_1, \ldots, m_n\}$ is solved via iterative numerical techniques by first robustly estimating the scale (denoted by $S_n^*$) then finding the solution of

$$\operatorname*{mean}_j\left(\psi\Big(\frac{m_j - T_n}{S_n^*}\Big)\right) = 0\,, \qquad \psi(x) = \tanh\frac{x}{2} = \frac{e^x - 1}{e^x + 1}\,. \tag{5}$$

After making an initial guess of the location, $T_n^{(0)}$, the iteration equation for solving Eq. (5) is

$$T_n^{(k)} = T_n^{(k-1)} + \frac{S_n^*}{0.4132}\operatorname*{mean}_j\left(\psi\Big(\frac{m_j - T_n^{(k-1)}}{S_n^*}\Big)\right) \tag{6}$$

which can be applied sequentially with $k = 1, 2, \ldots$ until $|T_n^{(k)} - T_n^{(k-1)}|$ is less than a predefined tolerance $\epsilon$.

 

```
=LAMBDA(Tn,SnEst,x,LET(
      tol, 0.0001,
  psi_arg, (x-Tn)/SnEst,
  psi_log, (EXP(psi_arg)-1)/(EXP(psi_arg)+1),
       Tp, Tn + SnEst*AVERAGE(psi_log)/0.4132,
  IF(ABS(Tp-Tn)>tol,RLoc(Tp,SnEst,x),Tp) ))
```

```
=LAMBDA(TnEst,Sn,x,LET(
      tol, 0.0001,
  psi_arg, (x-TnEst)/Sn/0.3739,
  psi_log, (EXP(psi_arg)-1)/(EXP(psi_arg)+1),
       Sp, Sn*SQRT(2*AVERAGE(psi_log^2)),
  IF(ABS(Sp-Sn)>tol,RSca(TnEst,Sp,x),Sp) ))
```

**Figure 1: Excel functions for implementing Eqs** (6) **and** (8) **to calculate robust location and scale. These are added to a spreadsheet using the 'Name Manager', named `RLoc` and `RSca` respectively, and can subsequently be used in a formula like any other Excel function. `RLoc(Tn0,SnEst,x)` calculates location, where `Tn0` is the initial estimate $T_n^{(0)}$, `SnEst` is the scale estimate $S_n^*$, and `x` is the set of marks; `RSca(TnEst,Sn0,x)` is similar for calculating scale.**

The scale $S_n$ is solved similarly, using robust estimate of location $T_n^*$, with

$$\underset{j}{\text{mean}}\left(\rho\Big(\frac{m_j - T_n^*}{S_n}\Big)\right) = \frac{1}{2}\,, \qquad \rho(x) = \psi^2\left(\frac{x}{0.3739}\right)\,. \tag{7}$$

The iteration equation for solving Eq. (7), with initial guess $S_n^{(0)}$, is

$$S_n^{(k)} = S_n^{(k-1)}\sqrt{2\,\underset{j}{\text{mean}}\left(\rho\Big(\frac{m_j - T_n^*}{S_n^{(k-1)}}\Big)\right)}\,. \tag{8}$$

The regularisation functions ($\psi(x)$ and $\rho(x)$) and constants ($0.4132$ and $0.3739$) in Eqs (5) to (7) are based on analytic expressions to ensure well-behaved theoretical behaviour (Rousseeuw and Verboven 2002).

Equations (6) and (8) require estimates of the location and scale. With some experimentation, we have found best results by first calculating location using $T_n^{(0)} = \text{median}(M)$ and $S_n^* = \text{AAD}(M) + \epsilon/2$, followed by calculating scale using $T_n^* = T_n^{(k)}$ (i.e., the converged location calculation) and $S_n^{(0)} = S_n^*$. The $\epsilon/2$ offset in $S_n^*$ is used to avoid numerical problems when a set of all equal marks results in $\text{AAD}(M) = 0$.

Equations (6) and (8) were implemented using Microsoft Excel's recent 'Lambda functions' (Lasak and Králová 2023). An example of the code for this is shown in Figure 1. While it would be possible to use a programming language instead, we find Excel to be a useful tool for this purpose as the results can be well visualised in tabular form during the analysis phase. It also has a lower barrier to entry for new members of the course coordination team.

The benefits of using the robust statistical measures location and scale, instead of the more traditional mean and standard deviation, are twofold. The location is less affected by outlier marks, so even if there are cases where outlier marks are not removed there is still confidence that the outlier would not overly shift the result and, as seen further below, the scale allows a semi-automated approach for identifying outlier marks.

## Assessor weighting

From experience, we know that some assessors mark more generously or more critically than others and this becomes even more apparent when the pool of assessors is broadened to include industry representatives. The use of the robust location to calculate a final mark is an improvement on the mean, but with small sample sizes it will not account for such systematic biases. The approach taken to calculate an assessor weighting to account for systematic differences is as follows, illustrated with a minimal example with six students and three assessors that do not mark all students. We have not been able to identify similar practices in the literature; if this approach is not novel, it has been developed independently.

 

**Table 1: Assessor weighting, step 1: Calculate the robust location of the scores awarded to each student ($\mathrm{loc}(S_i)$).**

| | Assessor 1 | Assessor 2 | Assessor 3 | $\mathrm{loc}(S_i)$ |
|---|---|---|---|---|
| Student 1 | 60 | | 54 | 57 |
| Student 2 | 70 | 77 | | 73.5 |
| Student 3 | 80 | 88 | | 84 |
| Student 4 | | 55 | 45 | 50 |
| Student 5 | | 77 | 63 | 70 |
| Student 6 | 60 | | 54 | 57 |
| $\mathrm{loc}(A_j)$ | 67.1 | 75.4 | 54 | |

**Table 2: Assessor weighting, step 2: Normalise each mark by the respective location of each student ($A_j \oslash \mathrm{loc}(S_i)$). The robust location of these normalised values for each assessor defines assessor weighting $w_j = \mathrm{loc}(\hat{A}_j)$.**

| | Assessor 1 | Assessor 2 | Assessor 3 |
|---|---|---|---|
| Student 1 | $\frac{60}{57} = 1.05$ | | $\frac{54}{57} = 0.95$ |
| Student 2 | 0.95 | 1.05 | |
| Student 3 | 0.95 | 1.05 | |
| Student 4 | | 1.1 | 0.9 |
| Student 5 | | 1.1 | 0.9 |
| Student 6 | 1.05 | | 0.95 |
| $w_j = \mathrm{loc}(\hat{A}_j)$ | 1 | 1.075 | 0.925 |

Define $S_i$ as the set of marks awarded to the $i$-th student, and $A_j$ the set of all marks awarded by the $j$-th assessor. First, the marks are tabulated and the respective locations are calculated for each student ($\mathrm{loc}(S_i)$) and each assessor ($\mathrm{loc}(A_j)$). This is shown in Table 1.

Note that in this case with only two assessments per student, the location $\mathrm{loc}(S_i)$ is identical to the mean. The purpose of $\mathrm{loc}(A_j)$ is simply to identify the 'robust average' mark per assessor, which aids manual checking and review of marks; however, this value is not used in the subsequent mathematical steps. The raw marks for each assessor are normalised against the location $\mathrm{loc}(S_i)$ for each student,

$$\hat{A}_j = A_j \oslash \mathrm{loc}(S_i), \tag{9}$$

where $\oslash$ refers to element-wise ('Hadamard') division. The normalised marks are tabulated again (Table 2). These assessor weighting results indicate where assessors have marked higher or lower for each student based on the other marks that each respective student recieved. In this example, it appears that Assessor 2 has been consistently generous, Assessor 3 consistently critical, and Assessor 1 is neither consistently generous nor critical. The resulting assessor weightings $w_j$ are used to normalise the original raw marks awarded by each assessor:

$$\bar{A}_j = A_j \oslash w_j = A_j \oslash \mathrm{loc}(\hat{A}_j)\,. \tag{10}$$

Defining $\bar{S}_i$ as the set of normalised marks awarded to the $i$-th student (i.e., considering the marks row-wise in the tabulated example), if there are no outliers to be removed, the final mark awarded $\bar{m}_i$ to each student is calculated accordingly with $\mathrm{loc}(\bar{S}_i)$, shown in Table 3. The statistics of each assessor (their weighting $\mathrm{loc}(\hat{A}_j)$ and their final 'robust average' $\mathrm{loc}(\bar{A}_j)$) can again be used qualitatively for feedback purposes.

**Table 3: Assessor weighting, step 3: Weight each mark by the respective assessor weighting, and re-calculate the robust location for each student to calculate the mark awarded $\bar{m}_i$. Outlier removal would occur during this step if necessary, but is not illustrated here (four or more marks per students required).**

| | Assessor 1 | Assessor 2 | Assessor 3 | $\bar{m}_i = \mathrm{loc}(\bar{S}_i)$ |
|---|---|---|---|---|
| Student 1 | 60 | | 58.4 | 59.2 |
| Student 2 | 70 | 71.6 | | 70.8 |
| Student 3 | 80 | 81.9 | | 81 |
| Student 4 | | 51.2 | 48.6 | 49.9 |
| Student 5 | | 71.6 | 68.1 | 69.9 |
| Student 6 | 60 | | 58.4 | 59.2 |
| $\mathrm{loc}(\bar{A}_j)$ | 67.1 | 70.1 | 58.4 | |

## Outlier removal

With the process in place for calculating the awarded mark $\bar{m}_i$, additional techniques from robust statistics can be used to identify outliers. In classical statistics, the z-score, $z_i$, for an individual mark $m_i$ is calculated with

$$z_i = \frac{m_i - \mu}{\sigma}\,, \tag{11}$$

where a threshold for $|z_i|$ would indicate a mark that is considered outside the expected range. Equivalently, following the treatment above, the robust score for the mark $m_{i,j}$ ($i$-th student, $j$-th assessor) can be considered instead with

$$r_{i,j} = \frac{m_{i,j} - \mathrm{loc}(\bar{S}_i)}{\mathrm{scale}(\bar{S}_i)}\,, \tag{12}$$

where $\mathrm{scale}(\cdot)$ is the robust scale measure. The value of the scale in some cases within our data varied substantially between students within some individual groups. This was particularly problematic when some students received very similar scores from their assessors, causing their scale values to approach zero. To remain conservative, the maximum scale within the group $G$ was used for the outlier detection calculation:

$$r_{i,j} = \frac{m_{i,j} - \mathrm{loc}(\bar{S}_i)}{\max\limits_{i \in G}(\mathrm{scale}(\bar{S}_i))} \tag{13}$$

A threshold of $|r_{i,j}| > 2$ sufficiently identified outliers in our data, which identified around 2.5% of the total number of marks as outliers. This threshold was found through manual tuning based on an analysis of marks which were post hoc identified as potentially problematic. The threshold was reduced until most manually identified cases were marked as outliers, but not so much as to introduce too many spurious positives in the outlier detection.

In an ideal scenario, it should be enough remove all marks with a robust score based on Eq. (13) outside the specified threshold. However, we identified edge cases where marks from an assessor were identified as outliers for some but not all students in the same group. These cases typically arose either when one student's performance was inconsistent with the rest of their peers and assessors varied in their assessment of this difference, or when each student obtained large $r$ values, but only some exceeded the threshold. In some of these cases, when outliers were removed, the rank order of the student marks within the group changed — for instance, it would be possible for the mark of a poor-performing student to be increased to an unfair degree, possibly higher than their peers. Since in such cases it was not possible to automatically identify whether the outlier was due to the student or the assessor, outlier marks were only discarded for a group if *all* marks from one assessor to that group were outliers.

 

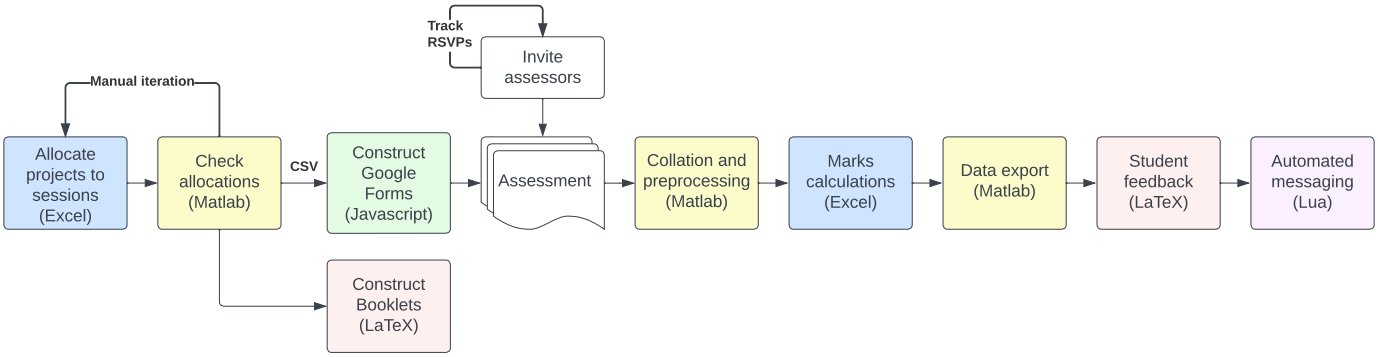


**Figure 2: Flow chart of high-level processes involved with organising the seminar event.**

## The workflow

The process for calculating student marks is the most involved element of a larger workflow that we use to organise the seminar and expo events (Figure 2). Although a full breakdown of this larger workflow is beyond the scope of this paper, a number of features are worth noting:

- It is a considerable amount of work to coordinate such events; a reusable process is invaluable for efficiency and maintaining quality.
- Despite incremental improvements over many years, the steps in the workflow are still quite disparate, requiring code/scripts written in five separate programming languages (Matlab, Excel, Lua, Javascript, LaTeX). If we were to reimplement the workflow from scratch, some consolidation would be possible.
- Manual iteration and ad hoc code is used to allocate projects to sessions to ensure: zero clashes for supervisors (e.g. none of their students present in parallel rooms simultaneously); 'fully booked' sessions to minimise event length; and clustering of projects by discipline and subdiscipline. If these requirements are not met key assessment issues identified earlier are likely to be exacerbated.
- Electronic assessment at scale proved more challenging than expected. At time of writing, only Google Forms provides an accessible API (using Javascript) for dynamically creating form questions, allowing student group details to be hard-coded to minimise the assessor effort needed, and data entry errors, by the assessors.
- The use of the Canvas LMS (Instructure, Inc.) REST API allows for automation of tasks such as uploading customised feedback documents and messaging students with the documents attached.

## Discussion

The workflow presented allows the current coordination team to manage the honours project seminar and expo assessment processes across multiple engineering disciplines. In 2023, the seminar presentation involved four parallel sessions over two days (eight sessions in total), with almost 350 students marked by over 60 assessors. Some 1600 individual marks were collated, with around 20 outlier marks rejected. Assessor weightings $w_j$ were distributed as follows: >25% of assessors were within $1.0 \pm 2\%$, >50% within $1.0 \pm 5\%$, and >80% within $1.0 \pm 10\%$. This relatively large spread of assessor weightings indicates that marks normalisation is playing an important role in helping improve equity to students, who receive varied pools of assessors. The robust scale of the individual assessor weightings also varied from 5% to 10% for most assessors, so despite normalisation there is still assessor-to-assessor variation to an equal order of magnitude that cannot be accounted for using summary statistics.

After marks normalisation and outlier removal, the absolute difference between mean and location for each set of student marks were distributed as: approximately 90% of students had a difference of 1% or less, 95% of students had a difference of 2% or less, and 99% of students had a difference of 5% or less. This relatively low difference between the calculation methods suggests that the use of robust statistics is of limited value in the calculation of the final mark itself, most of the time. However, the critical purpose of the robust statistics approach is an automated and reliable method for identifying outliers, which was well achieved.

The assessment and feedback process aids in ensuring effective assessment of individual students within the team-based presentation. Eliot et al. (2012) states effective assessments must be transparent, outcomes orientated, provide opportunities for individual students to demonstrate their learning, and strategically develop learning intent. This was initially addressed in the rubric, as 40% of an individual's mark relates to their own presentation skills rather

than the technical content, thus ensuring the marking criteria aligns with the assessment type (Kim 2014) and that each student has the opportunity to demonstrate their learning. During the feedback phase, each student received details of individual assessor marks and comments, including outliers marked as rejected, alongside a simplified explanation of the defensible process outlined in this paper. This was particularly important as we suggest that there is a subjective or even emotive element to the interpretation of grades awarded to honours project assessment and therefore such communication ensures transparency and an outcomes focus that provides assurance to the students that their marks are being calculated fairly.

Through this newly developed process, outlier detection was able to successfully identify cases such as: a supervisor assessing their own students, and being overly generous; an academic awarding marks significantly below the other assessors present for a cross-disciplinary project; and, a PhD student in the field of research being overly critical based on the technical material rather than on the communication of the work as a whole. While these could be screened for in other ways (e.g., we could omit marks from supervisors towards their own projects), the solution presented here has proven to be robust at identifying such cases. The assessor weighting process has been similarly effective. While the effects of the statistical calculations are small, the robustness and transparency of the approach have reduced the rate of negative student feedback. We suggest for those who wish to adapt such methods, that assessment activities are structured with at least four markers to allow outliers to be robustly detected.

In this paper we have presented in detail the method we have developed to robustly calculate marks from disparate assessors, which is able to correct for systematic (on average) differences between assessors and eliminate outlier marks automatically. It is a key component in a larger workflow we are developing to manage the process of running seminar and expo events for honours and masters research projects. As this is a new process, in the future we plan to longitudinally assess the need for and effectiveness of this approach.

## References


Eliot, M, P Howard, F Nouwens, A Stojcevski, L Mann, JK Prpic, R Gabb, S Venkatesan, and A Kolmos (2012). "Developing a conceptual model for the effective assessment of individual student learning in team-based subjects". In: *Australasian Journal of Engineering Education* 18.1, pp. 105–112.

Henderson, Alan, Marcus Guijt, Michael Breadmore, Anna Carew, and Rosanne Guijt (2009). "Honour thesis assessment: The role of guidelines in achieving inter-rater agreement". In: *Proceedings of 20th Annual Conference for the Australasian Association for Engineering Education*.

Kim, Ho Sung (2013). "Quantification for complex assessment: uncertainty estimation in final year project thesis assessment". In: *European Journal of Engineering Education* 38.6, pp. 671–686.

Kim, Ho Sung (2014). "Uncertainty analysis for peer assessment: oral presentation skills for final year project". In: *European Journal of Engineering Education* 39.1, pp. 68–83.

Lasak, P. and M. Králová (2023). "Lambda Function In Excel In Response To Challenges Of Nonprogrammers In Modern Corporate Practice". In: *INTED2023 Proceedings*. 17th International Technology, Education and Development Conference. Valencia, Spain: IATED, pp. 931–937. DOI: 10.21125/inted.2023.0288.

Littlefair, Guy and Peter Gossman (2008). "BE (Hons) final year project assessment — leaving out the subjectiveness". In: *Proceedings of 19th Annual Conference for the Australasian Association for Engineering Education*.

Maronna, Ricardo A., R. Douglas Martin, Victor J. Yohai, and Matías Salibián-Barrera, eds. (2018). *Robust Statistics: Theory and Methods (with R)*. John Wiley & Sons Ltd. DOI: 10.1002/9781119214656.

Rousseeuw, Peter J. and Christophe Croux (1993). "Alternatives to the Median Absolute Deviation". In: *Journal of the American Statistical Association* 88.424, pp. 1273–1283. DOI: 10.1080/01621459.1993.10476408.

Rousseeuw, Peter J. and Sabine Verboven (2002). "Robust estimation in very small samples". In: *Computational Statistics & Data Analysis* 40, pp. 741–758. DOI: 10.1016/S0167-9473(02)00078-6.


## Copyright statement