Res. Lett. Inf. Math. Sci., 2005, Vol. **7**, pp 117-126
Available online at http://iims.massey.ac.nz/research/letters/

A control chart procedure for student grade monitoring

H. P. EDWARDS[1], K.GOVINDARAJU[2] & C. D. LAI[2]

[1] *Institute of Information & Mathematical Sciences*
Massey University at Albany, Auckland, New Zealand †

[2] *Institute of Information Sciences and Technology*
Massey University, Palmerston North, New Zealand ‡

This article reports an application of the control chart procedure for monitoring award of grades to students by the teaching staff in a large university. The chart procedure signals the presence of special cause variations if any in the award of grades. Implementation of the grade monitoring procedure saved considerable time and effort while ensuring that the reported special cause situations are justified. The mathematical derivations for the new control chart scheme are also presented.

1 Introduction

Massey University, New Zealand (www.massey.ac.nz) is organised around five academic colleges each composed of various Institutes, Schools and Departments which are spread across several campuses and sites. Each college is responsible for setting generic guidelines for the offer of grades (A, B, C etc) to its students. These administrative guidelines, which are similar across the colleges, prescribe that only a certain percentage of students can be offered top grades such A or A+. The colleges also have examination committees that monitor the tendency of some faculty who may offer very high or low grades to a large proportion of passing students. The administrative guidelines sometimes create tensions between faculty and college examination committees when a particular faculty member consistently offers a low (or a high) grade in a paper compared to other similar papers.

This article describes a how a statistical monitoring procedure can be put in place for grading control and improvement purposes. The procedure is based on a Markov chain formulation of the charting procedure and can be easily implemented using a spreadsheet package such as Excel.

2 Review of grading student marks and issues involved

The term *grade* used in student assessments has the same meaning in the quality literature as a category or rank indicator. As explained by Freund (1985), any grading should be able to reflect a recognizable difference. The planned difference in grading

† email addresses: h.edwards@massey.ac.nz; k.govindaraju@massey.ac.nz; c.lai@massey.ac.nz

students at Massey University is beyond the simple ordinal nature of the grades namely A+, A, A-, B+, B, B-, C+, C, (passing grades), R (a passing grade but Restricts the student from doing any advanced paper in the topic), D, E (failing grades) etc. In what follows, "A" grades refers collectively to A+, A and A- grades, and "C" grades refers collectively to C+, C and R grades.

One of the planned features of grading is that of requiring that the percentage of A+, "A" and "C" grades awarded in any particular offering should lie within certain prescribed limits over time. We will consider the Bachelor of Information Sciences (BInfSc) program in our discussion. The grade guidelines for a large first year undergraduate paper (taken typically by 400 students in a particular campus) are as follows:

Grade	**Allowable range**
A+	3 - 4% of all passing students
A+, A, A-	13 - 17% of all passing students
C+, C, R	45 - 50% of all passing students

It should be emphasised that the above ranges apply only to passing students.

The issue of quality in student assessment remains a matter of excellence rather than simple conformance. This is because students are in competition with each other in many respects: especially for jobs and scholarships. While students retain some freedom to choose which papers to study (and many tend to choose papers in subjects they are good at), they should not otherwise gain from or be penalized by their choice of papers. Hence the above grading scheme also provides equity. About one-half of the passing students being graded "C" and no more than 17% graded "A" are matters of quality rather than grading. This ensures a clear direct relationship between quality and grading in a given year. A student is typically assessed with assignments/home work, test/quiz and a final written examination. The scores in these components of assessments make up the final mark with varying weights for each components of assessment. The final marks are not usually standardized (or normalized), rather ranges of marks are decided for each of the above grades using the grade guidelines and grades allocated accordingly. These ranges of marks will vary from paper to paper and year to year as well.

We will now discuss the different source variations in student assessment and grading. The following is a list of causes of structural variation in marks in a particular batch of students:

- Two or more faculty teaching the paper leading to some assessment differences.

- A paper is taught in internal lecture mode for 13 weeks as well as off-campus or extramural mode where students attend only a very short campus course. Some are taught in a block mode with intensive lecturing for a short period of time. Even

though the courses are equivalent despite delivery modes, the assessment components may not be identical.

- Different faculties teaching the paper in different years or on different campuses – Massey University has campuses in three different cities.

- The components of assessment are not exactly the same for all offerings. That is, the duration of the exam, number of questions and choices available, number of assignments, number of questions in each assignment, test/quiz etc vary from year to year as well as between campuses in the same year. This is mostly at the discretion of the course coordinator of the paper in a particular campus. In continuous improvement in delivery (such as use of world-wide-web etc) as well as technology (such as software), the assessment procedures also undergo changes.

It is expected that the set limits for each of the grade groupings will take care of the above *common cause* type of variations occurring within a single offerings of the paper. For example, the 13 to 17% limit for A grade allows a *specification spread* of 5%. Common cause variations and rounding errors etc in marking etc are expected to be within this limit. However long-term trend variation in students' improved (or falling) performance is not accounted for in the above grade proportion specifications. For example, there is evidence that student performance is falling in mathematically intensive theory papers in the longer term in New Zealand but improving in computationally intensive papers. This means that two C graded students may possess very different performance or skill levels over a longer period of time such as 10 years.

Sometimes, especially in papers with small classes, the students in one year may form a particularly good class, or a particularly weak class, and examiners may recommend distributions further outside the grading guidelines. That is, grade distributions outside the guidelines may be accepted if there is evidence for an unusual class from the past record of the students in question or from a comparison with their grades in other papers in the same semester and/or previous semester. Over a reasonable time scale (say 5 years), the percentages of “A” grades and “C” grades are expected to scatter about the mid point of the guideline range. However there may be a tendency to consistently offer low or high grades for a particular paper. Hence a statistical monitoring procedure of the student grade distribution is needed to detect the presence of such *special cause* variations.

The distribution of grades for six different consecutive offerings of one particular paper is shown in Table 1. Evidently the percentage of “A” grades is above the guideline limit of 17% for the first three time periods. If the guidelines are not met in any particular year, the Examinations Committee has to decide whether there are acceptable reasons for the guidelines to be flouted or not (i.e. the presence of special cause(s) of variation). Verification of the presence of special cause variations in grading is a time consuming process, calling for discussion with the concerned faculty, verification of records etc, but the time available to approve student results in order to meet published

deadlines is very short. Hence it is desirable to have a statistical monitoring procedure in place.

Table 1: Distribution of student grades in a particular paper

Period	Number of passing students	"A" Grades	"B" Grades	"C" Grades
1	100	18 (18%)	31 (31%)	51 (51%)
2	104	22 (21%)	33 (32%)	49 (47%)
3	132	28 (21%)	52 (39%)	52 (39%)
4	152	24 (16%)	62 (41%)	66 (43%)
5	153	24 (16%)	55 (36%)	74 (48%)
6	177	29 (16%)	57 (32%)	91 (51%)

3 Control chart for monitoring grade distributions

In traditional statistical process control procedures, the false alarm probability of signaling the presence of special cause variation when it is not present is kept at a very low figure such as 0.0027. In other words, a single false alarm can be expected for every 370 points plotted on a chart. Such a low false alarm rate is not feasible for student grade monitoring because typically there are only few years available for monitoring. Courses and papers tend to evolve and change over a 5 to 10 year cycle. Furthermore the objective of the grade monitoring is largely in the shorter term of about 5 years. Hence the traditional Shewhart three sigma control limits are not usable. Ryan (1997) provides a discussion in a similar context and strongly recommends adoption of tighter control limits for certain applications.

For monitoring multinomial processes such as the one presented above, Marcucci (1985) illustrates a procedure and recommends use of a control statistic similar to a chi-squared statistic. The midpoint of the guideline grade ranges will be used to obtain the expected counts. For example, the expected number of "A" grades for 100 passing students is 15 (for the first period shown in Table 1). This expected count is then used to compute the chi-squared statistic for the Grade "A" category as $(18\text{-}15)^2/15 = 0.6$. A similar chi-squared statistic for the "C" categories is found to be $(51\text{-}47.5)^2/47.5 = 0.26$. Note that there is no grading guideline range for "B" grades and hence no chi-squared

statistic will be computed - indirectly, the control of the top and low grade proportions will lead to the control of B grades. Following Marcucci (1985), the sum of the chi-squared statistics for ”A” and “C” grade categories will be the control statistic for our monitoring procedures. For 2 degrees of freedom, the right tail area beyond 5.99 for the chi-squared distribution is only 5%. Hence a control chart with a false alarm probability of 0.05 can be drawn for Table 1 data as shown below:

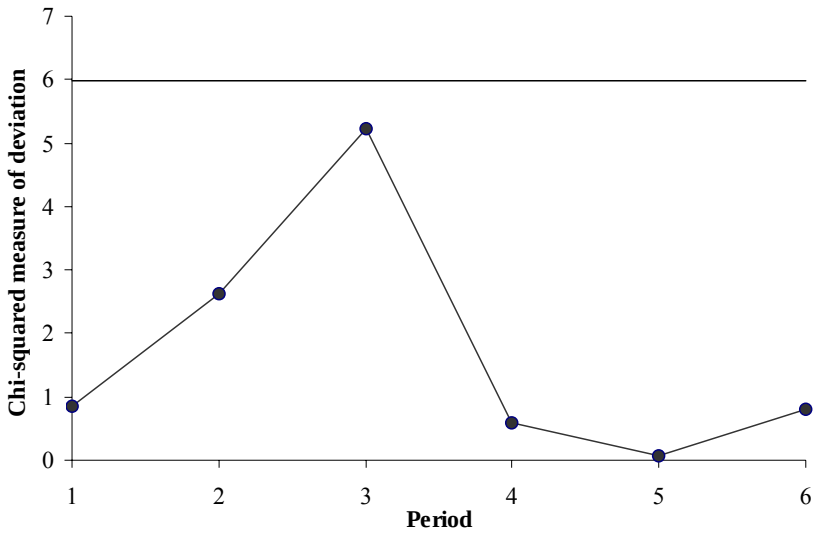

Figure 1: Chart for monitoring student grades

Obviously the above chart does not signal the presence of a special cause variation in grading for the paper after all.

Unfortunately, a signal rule based on a single plotted point is not particularly sensitive to small upward or downward deviations from the grading guidelines. Hence we need to introduce a warning line and an additional signal rule to detect small shift levels in grading. See Ryan (1989) for a discussion of control chart supplementary run rules. One of the supplementary signal rules relevant is that of two successive points above the warning limit line. This warning limit line is drawn at as a 77% Probability Limit which keeps the overall false alarm rate at about 1/20. The derivation is provided in Section 4. Figure 2 shows the control chart with the warning limits. The chart again did not issue any signal for the presence of special cause variations.

A slightly more complicated procedure is to have two warning limits say lower and upper warning limits. Two additional signal rules are employed: (i) two successive points above the upper warning line but below the control limit and (ii) three successive points between the upper and lower warning lines. The practical reason for using consecutive points for signals is related to context. That is, the faculty is not taking any corrective steps for successive paper offering has to be viewed seriously. Figure 3 provides such a control chart for Table 1 data.

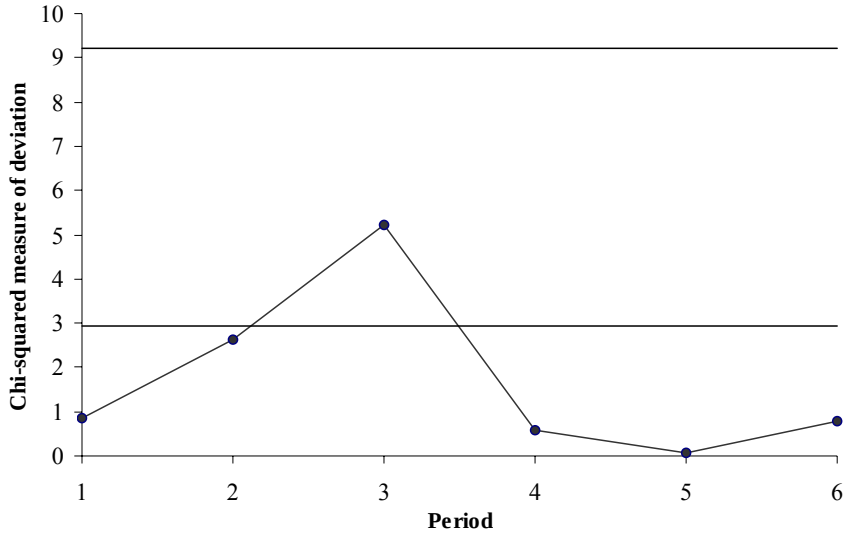

Figure 2: Control chart with warning limits

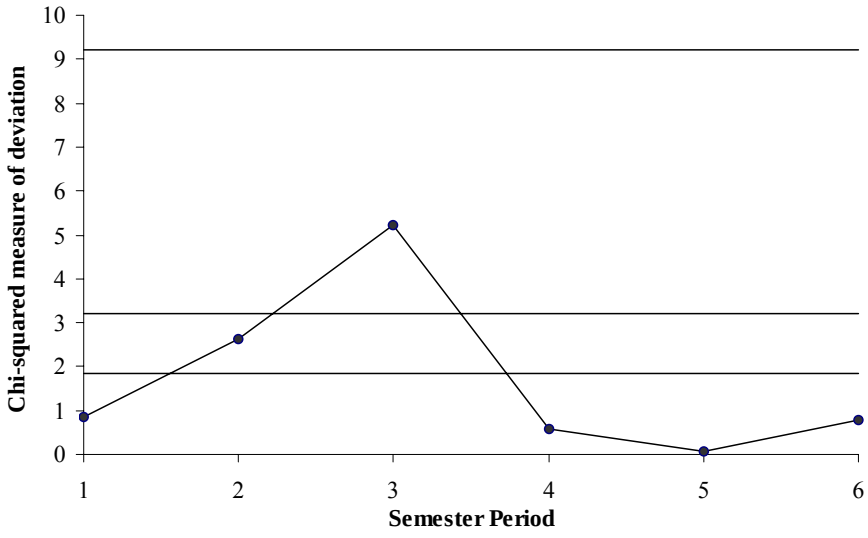

Figure 3: Control chart with two warning limits

The advantage of having three signal rules is that of detecting both small and moderate shifts. The false alarm probability is again fixed at 0.05 for the above chart.

The above monitoring procedure was tried for a large number of papers and found to work well. For confidentiality reasons, we are not providing any identifiable course details in Figures 4 and 5 where special cause signals are issued.

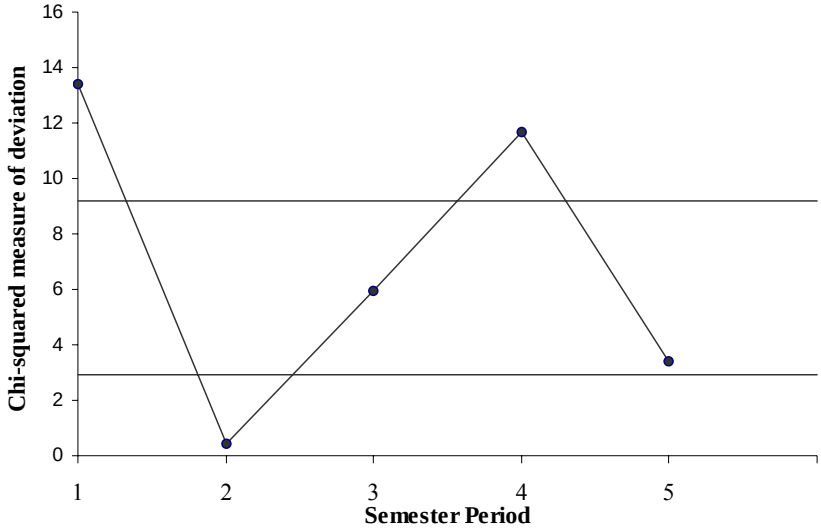

Figure 4: A sample chart giving a signal (single warning limit)

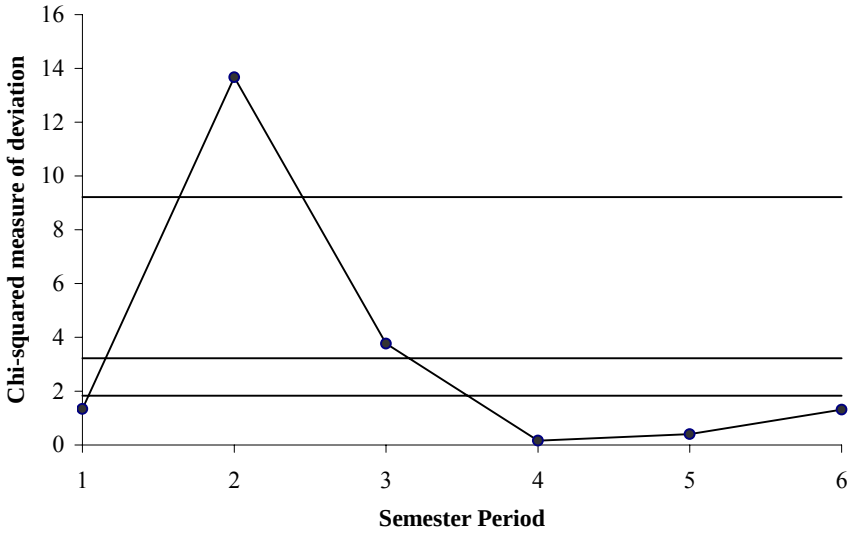

Figure 5: A sample chart giving a signal (two warning limits)

4 Conclusions

It is well known that every process involves both common and special cause variations. This is true for the process of grading students' performance in various papers of their study. A suitable control chart procedure is needed to monitor the variations in student grading. The charting procedure saves considerable time when grading does not conform to set norms in a particular year due to the presence of special cause variation

such as an occasionally good batch of students being present, while allowing for the common cause variations over a period of time. The charting procedure presented in this paper is found to work well and can be easily modified to other similar situations.

4 Mathematical Appendix

For the one-sided control chart procedure with a single warning limit, we define

X = control statistic, the sum of the two chi-squares

CL = control limit

WL = warning limit (< CL)

$$p_0 = \Pr\left(0 < X \le WL\right)$$

$$p_1 = \Pr\left(WL < X \le CL\right)$$

$$p_2 = \Pr\left(X > CL\right)$$

The two signal rules (i) a point beyond CL, and (ii) two consecutive points between the warning and control limits lead to the following expression for average run length (ARL):

$$ARL = \frac{(1 + p_1)}{(1 - p_0 - p_0 p_1)}$$

(see Wetherill and Brown (1991) for a proof). For a fixed $p_2 = 0.01$, and a desired ARL of about 20 (i.e. a false alarm probability α of 5%), p_0 and p_1 can be numerically solved to be 0.77 and 0.22 (two digit accuracy) respectively. These values achieve an ARL of 20.132 instead of 20. $p_2 = \Pr(X > CL) = 0.01$ which for a χ^2 distribution with 2 d.f, gives $CL = 9.21$. $p_0 = \Pr(0 < X \le CL) = 0.77$ gives $WL = 2.94$.

For the one-sided control chart procedure with two warning limits, we define

CL = control limit

WL_1 = upper warning limit (< CL)

WL_2 = lower warning limit (< $WL1$)

$$p_0 = \Pr\left(0 < X \le WL_2\right)$$

$$p_1 = \Pr\left(WL_2 < X \le WL_1\right)$$

$$p_2 = \Pr\left(WL_1 < X \le CL\right)$$

$$p_3 = \Pr\left(X > CL\right)$$

In order to obtain the *ARL* for the control procedure with signal rules (i) a point beyond *CL*, (ii) three consecutive points between the warning limits and (iii) two consecutive points between the upper warning and control limits, we shall follow the Markov Chain approach of Brook and Evans (1972). Also see Wetherill and Brown (1991). We define the states

$$S_{00}:\left(0 < X_{i-1} \le WL_2 \;\&\; 0 < X_i \le WL_2\right),$$
$$S_{01}:\left(0 < X_{i-1} \le WL_2 \;\&\; WL_2 < X_i \le WL_1\right),$$
$$S_{02}:\left(0 < X_{i-1} \le WL_2 \;\&\; WL_1 < X_i \le CL\right)$$
$$S_{10}:\left(WL_2 < X_{i-1} \le WL_1 \;\&\; 0 < X_i \le WL_2\right),$$
$$S_{11}:\left(WL_2 < X_{i-1} \le WL_1 \;\&\; WL_2 < X_i \le WL_1\right),$$
$$S_{12}:\left(WL_2 < X_{i-1} \le WL_1 \;\&\; WL_1 < X_i \le CL\right)$$
$$S_{20}:\left(WL_1 < X_{i-1} \le CL \;\&\; 0 < X_i \le WL_2\right),$$
$$S_{21}:\left(WL_1 < X_{i-1} \le CL \;\&\; WL_2 < X_i \le WL_1\right), \text{ and}$$
$$S: \text{the absorbing state representing all signal events},$$

where X_{i-1} and X_i (i = 1,2,3….) are the two consecutive control statistic values. For the above states, consider the following transition matrix P:

$$\begin{array}{c} \\ 00 \\ 01 \\ 02 \\ 10 \\ 11 \\ 12 \\ 20 \\ 21 \\ S \end{array}\!\!\begin{array}{c} \begin{array}{cccccccc} 00 & 01 & 02 & 10 & 11 & 12 & 20 & 21 \end{array} \\ \left[\begin{array}{cccccccc|c} p_0 & p_1 & p_2 & - & - & - & - & - & 1-p_0-p_1-p_2 \\ - & - & - & p_0 & p_1 & p_2 & - & - & 1-p_0-p_1-p_2 \\ - & - & - & - & - & - & p_0 & p_1 & 1-p_0-p_1 \\ p_0 & p_1 & p_2 & - & - & - & - & - & 1-p_0-p_1-p_2 \\ - & - & - & p_0 & - & p_2 & - & - & 1-p_0-p_2 \\ - & - & - & - & - & - & p_0 & p_1 & 1-p_0-p_1 \\ p_0 & p_1 & p_2 & - & - & - & - & - & 1-p_0-p_1-p_2 \\ - & - & - & p_0 & p_1 & p_2 & - & - & 1-p_0-p_1-p_2 \\ \hline - & - & - & - & - & - & - & - & 1 \end{array}\right] \end{array} = \begin{bmatrix} \mathbf{R} & \mathbf{p} \\ \mathbf{0}' & \mathbf{1} \end{bmatrix}$$

The vector of average run lengths **ARL** is written as

$$\mathbf{ARL} = \left(\mathbf{I} - \mathbf{R}\right)^{-1}\mathbf{1}$$

The first element of the **ARL** vector will be the usual average run length depending on the formulation of the initial state. In our case, it is

$$ARL = \frac{1 + p_1^2 + p_2 p_1^2 + p_1 + p_1 p_2 + p_2 + 1}{1 - (p_1^2 p_0 + p_1^2 p_2 p_0 + p_2 p_1^2 + p_0 p_1 + p_1 p_0 p_2 + p_1 p_2 + p_0 p_2 + p_0)}$$

For a fixed $p_3 = 0.01$, and a desired *ARL* of about 20 (i.e. a false alarm probability α of 5%), p_0, p_1 and p_2 can be numerically solved to be 0.59, 0.20 and 0.20 (two digit accuracy) respectively. These values achieve an *ARL* of 20.09 instead of 20. Using these values, we obtain $CL = 9.21$, $WL_1 = 3.28$, and $WL_2 = 1.83$ and corresponding to tail areas 0.01, 0.20 and 0.40 respectively for a χ^2 distribution with 2 d.f.

Acknowledgement

The authors wish to thank Eugene Lai for his assistance in extracting data from extensive student record databases and summarizing them for analysis.

References

Brook, D. and Evans, D.A. (1972). An Approach to the Probability Distribution of Cusum Run Length. *Biometrika,* 59:539-549.

Freund R. A. (1985). Definitions and Basic Quality Concepts. *Journal of Quality Technology*, 17(1):50-56.

Marcucci, M. (1985). Monitoring Multinomial Processes. *Journal of Quality Technology*, 17(2):86-91.

Ryan, T.P. (1997). A Discussion on Statistically-Based Process Monitoring and Control. *Journal of Quality Technology*, 29(2):148-156.

Ryan, T.P. (1989). *Statistical Methods for Quality Improvement*. John Wiley & Sons, New York.

Wetherill, G.B. and Brown, D.W. (1991). *Statistical Process Control*. Chapman and Hall, London.