Int. J. Engng Ed. Vol. 20, No. 5, pp. 820–833, 2004
Printed in Great Britain.

0949-149X/91 $3.00+0.00
© 2004 TEMPUS Publications.

A Robust Grade Adjustment Procedure*

D. A. BARRY, D.-S. JENG,** R. B. WARDLAW, M. CRAPPER and S. D. SMITH
School of Engineering and Electronics, The University of Edinburgh, Edinburgh EH9 3JL, UK.
E-mail: d.a.barry@ed.ac.uk

Assurance of equity in student outcomes includes ensuring that module results are comparable within reasonable bounds. We present a procedure that first compares (and adjusts if necessary) module variances. Then, student results are individually normalised to standard variates. A simple sign test is used to identify modules with disproportionately good or poor results. Module results are offset so as to adjust averages that do deviate. A detailed examination of an artificial data set shows that the proposed very simple procedure yields results that agree with a sophisticated statistical analysis.

INTRODUCTION

STUDENT EXPECTATIONS of fair and robust grading schemes are mirrored by more general community expectations: each group uses university results for a variety of purposes. Employers and admissions personnel at the student's home university and at other educational institutions examine grades routinely. In countries like the United Kingdom and Australia, engineering graduates are ranked according to the class of honours attained; this honours degree outcome can affect a graduate's career prospects for several years. Results of examinations should, therefore, be subject to processes to ensure fair and equitable outcomes. Such outcomes are also necessary because student assessment of their own performance is often at odds with the assessment of tutors and lecturers [1].

Not surprisingly then, at least annually, academics expend much time and effort in reaching agreement on student grades. For example, a scheme was developed for comparing consistency of marking of undergraduate theses [2]. It is not at all unusual for engineering academics to attend lengthy meetings at which detailed discussions occur regarding differences between the performance of individual students in different subjects/modules (we use 'module' to refer to a subject or similar discrete component in a given year of a degree programme), and in the overall performance displayed in those modules. In particular, summary statistics such as averages and standard deviations (or variances) for various modules are compared, discussed and, indeed, debated. Perhaps module *X* has a very low average, whereas modules *Y* and *Z* have averages 20% or more higher. Explanations abound concerning the difficulty of the material in *X* relative to *Y* and *Z* or, conversely, the lack of difficulty of *Y* and *Z* relative to *X*. Or, perhaps *X* has been taught by a junior or otherwise inexperienced academic [3, 4], there was an unforeseen timetable clash, formative assessment items were not returned in a timely fashion, the examiners placed different emphasises on different aspects of assessed work [5], or library/laboratory facilities were inadequate in the relevant area. In addition, these discussions are exacerbated by the fact that usually students have a good deal of choice in the range of modules studied, in which case module *X* might have been taken by a group of interested students, while those taking *Y* and *Z* were highly motivated. How can examination boards (the term 'examination board' refers to the committee with formal responsibility for awarding student grades) be assured that the excellent grades in *Y* and *Z* are not simply a manifestation of generous marking?

Another explanation for disparity in module averages is that academics, being individuals, will produce exams of varying difficulty, or will vary in the results awarded in marking student work of identical quality. Perhaps on sound pedagogical grounds, a new teaching method was employed, with unforeseen consequences.

Clearly, procedures to quantify whether differences between module outcomes are justified on reasonable statistical grounds can provide very useful timesaving guidance to assist examination boards in identifying the need for grade-moderation discussions. On the other hand, statistical measures and comparisons can likewise help identify when detailed discussions are not needed.

As mentioned already, given that student results can have a major effect on the prospects of graduating students [6], it is important that consistent and fair grades are given. More generally, quality assurance procedures should be robust and ensure equity in student outcomes. That is, although a brief perusal of student/module results might not suggest any obvious outcomes necessitating discussion, a

* Accepted 5 September 2003.

** School of Engineering, Griffith University, Gold Coast, PMB 50 Gold Coast Mail Centre, 9726 Australia mature quality assurance procedure would include a standard set of quantitative checks to alert the examination board to possibly biased or otherwise unusual results. Quality assurance involves identifying where moderation is justified, as well as where it is not.

A perusal of the statistical and educational assessment literature reveals various statistical methods that could be applied to identify the need grade moderation [7]. Unfortunately, the level of statistical sophistication needed is likely not readily available in many examination boards, particularly given the time pressure under which boards typically operate. Even with available expertise, for many sophisticated procedures there is frequently a substantial requirement for data preparation and manipulation, checking and interpretation. Statistical methodologies that cannot be readily programmed within a spreadsheet are not likely to be of widespread practical use. Furthermore, interpretation of results from more advanced procedures might rely on statistical training; individuals with such training would, likewise, not generally be available for most examination boards.

To make the process of arriving at final student outcomes more concrete, we provide a conceptual outline of the steps to be taken in Fig. 1. Grades are collected and summarised, usually in a spreadsheet, and summary statistics calculated. Our purpose is to present a simple yet robust statistical procedure for analysing results of a given cohort of university students. It is envisaged that the procedure would be applied to the results of all students in a given year of an engineering degree program. We proceed with this context in mind. The part of Fig. 1 that will be the main focus of this paper is the procedure to identify the need to adjust grades. We present, in addition, a simple grade-adjustment procedure.

STATISTICAL PROCEDURE

We wish to identify amongst a group of modules, modules for which student results are too high or too low. Following such identification, it is expected that a grade adjustment procedure would be implemented to adjust grades up or down, as necessary to remove anomalies. Following adjustment, the procedure would be re-applied to ensure that equitable results were obtained.

Consider a cohort of students taking N modules. Some degree programmes are based on pass-by-year system, whereas others work on a pass-by-module system. Either case is accommodated as we are aiming to compare consistency of module outcomes. It is, however, worth recalling that in many circumstances the overall average grade is important (say, for progression to the next academic year, for determination of honours classifications or for allocation of academic prizes). The overall grade will be the weighted sum of several module results. As noted previously [8], it 'is a common misconception that the nominal weights correspond to the relative weights of the variables in the composite'. Put another way, the variance computed for a student's overall average (the composite) is weighted not according to the assigned (nominal) weight of each module grade, but according to the relative weight computed for the variance of that sum. This effect is exacerbated

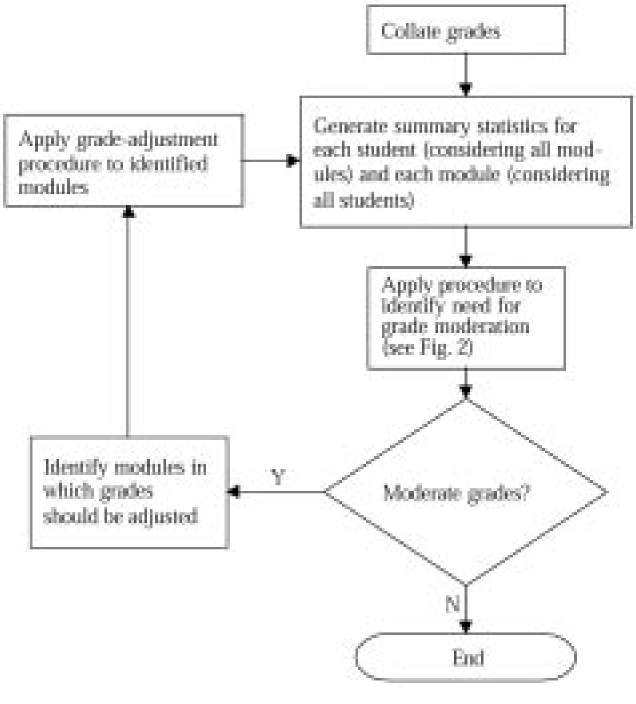

Collate grades
Generate summary statistics for each student (considering all modules) and each module (considering all students)
Apply procedure to identify need for grade moderation (see Fig. 2)
Moderate grades?
Y
N
End
Identify modules in which grades should be adjusted
Apply grade-adjustment procedure to identified modules

Fig. 1. Flow chart of the grade moderation process.

in circumstances where each module grade comes from distributions with different variances. The variability of each summed component should be approximately equal to ensure that the variability of the weighted average fairly reflects the weighting assigned to each component module [9]. Even where an average year grade is not computed, it is still good practice to have approximately equal variances in each module. In practical terms, modules with variances that are too large (relative to a given norm) simultaneously benefit some students (with relatively higher grades) and disadvantage others (with relatively lower grades). Conversely, modules with a relatively smaller variance cluster all grades towards the mean. Thus, students with relatively good performances in the module are disadvantaged while those who performed less well benefit. Below, we shall use a simple check to ensure module variances are comparable.

With this in mind, a possible next step in the procedure is to compare averages of the individual modules. Rather than compare averages directly, in our procedure we adopt a student-centred approach and look at overall student performance. To do this, the procedure identifies outliers in each student's performance, and aggregates these for each module. This procedure will be detailed below.

In terms of comparing averages directly, we note that student results in each module give the averages $\overline{x}_i$, $i = 1, \ldots, N$. Each average is an estimate of the population mean, μ_i, assuming the population consists of all students eligible to take module i. Then, it is useful to test the hypothesis $H_0\colon \mu_1 = \mu_2 = \ldots \mu_N$, in which case the one-way ANOVA (analysis of variance) test would be applied. Alternatively, one could check means in a pairwise fashion using a variety of tests [10], although this approach would engender considerable effort—a total of $N(N-1)/2$ comparisons.

Elsewhere, it was concluded that an ANOVA was the most reliable way to check comparability of secondary subjects in national exams carried out in the United Kingdom [11]. But, in order to have confidence in the ANOVA results, several checks would need to be undertaken. Since the standard ANOVA relies on the assumption of equality of variances (which is desirable in any case), the estimated variances to be used in the ANOVA should be checked statistically for equality. Likewise, the ANOVA procedure assumes that normally distributed populations are being sampled; again, a check should be carried out.

In a standard one-way ANOVA, the experimental subjects (in this case, students) should be drawn as independent samples. Sample outcomes used in the ANOVA should be uncorrelated. Since the students being assessed generally take several modules in a given academic year, it is highly doubtful that the results in different modules would be uncorrelated. The one-way repeated measures ANOVA [12] is applicable for correlated samples [13]. Non-parametric approaches are available [14], however the implementation of these is not straightforward and can involve data manipulation such as ranking. As mentioned in the introduction, given time constraints and perhaps lack of significant statistical expertise, generally it is not practically feasible to carry out these procedures and, in the case of parametric tests, associated assumption checking. In that case, we proceed to a simplified approach, described below.

In comparing student performance across many modules, the key question we wish to answer is whether the outcomes are comparable. This question immediately focuses attention on the average grade awarded in each module. However, even if all modules have comparable averages, as already mentioned a fair comparison would be based on the condition that the results in each module also exhibit comparable variability. To evaluate comparability of module outcomes, we aim to discern cases where a student's grade in a given module is markedly different from their year average. Modules with a high proportion of better-than-average or worse-than-average performances are identified for grade moderation. Details of an algorithm (Grade Adjustment Procedure, GAP) that achieves these steps is presented in Fig. 2. The various steps in this figure are discussed below.

GAP: details of steps undertaken in Fig. 2

For convenience, we assume all module grades are given as percentages.

In Step 1, individual module variances are checked for equality. There are several methods for comparing variances [15]. A simple (although parametric) procedure begins with calculation of the average variance over all the modules. Then, we compare individual module variances with the average. That is, compute:

$$\overline{s}^2 = \frac{1}{N}\sum_{i=1}^{N} s_i^2 \tag{1}$$

where N is the number of modules, s_i^2 is the estimated variance for module i and $\overline{s}^2$ is the average module variance. For each module, we accept the hypothesis that $\sigma_i^2 = \overline{\sigma}^2$ at the 0.1 (0.05, 0.01) significance level if:

$$\frac{|s_i - \overline{s}|}{\overline{s}}\sqrt{2n_i} < 1.64(1.96, 2.58) \tag{2}$$

where n_i is the number of students taking module i. The check in (2) is an approximation to the appropriate χ^2-based statistic [13]:

$$\frac{(n_i - 1)s_i^2}{\overline{s}^2} < \chi^2_{n_i-1,\alpha} \tag{3}$$

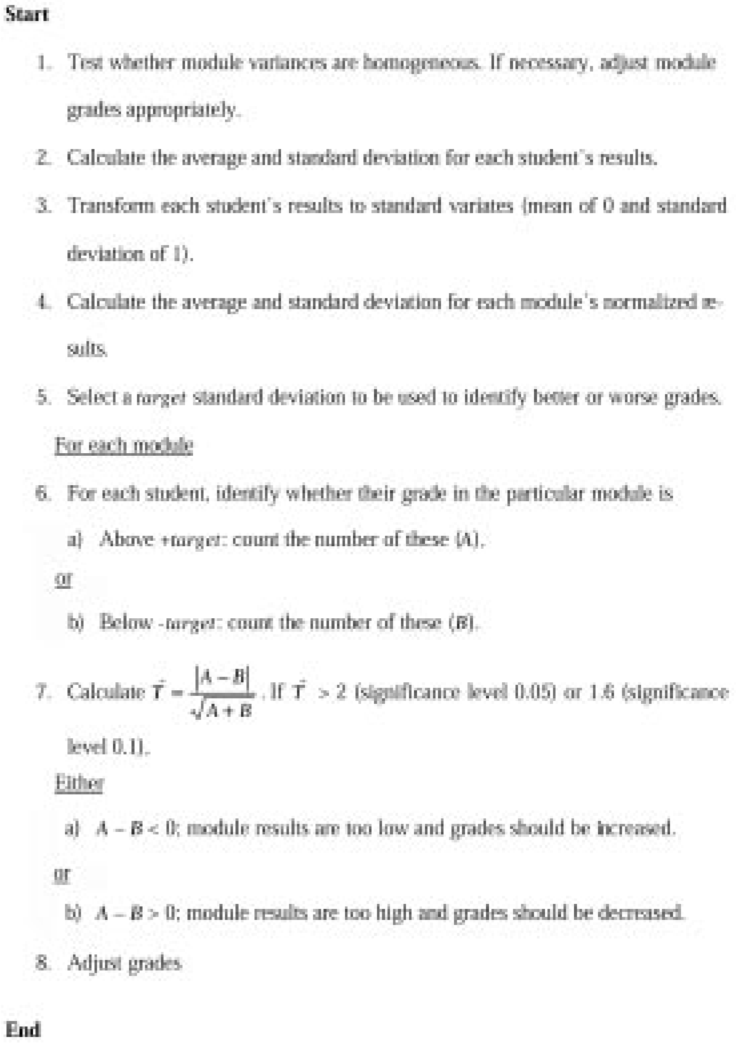
where α is the significance level. The tests in (2) and (3) both assume that $\overline{s}$ is independent of s_i, which is clearly not the case. However, it would be the case if the average in (1) were modified such that it was computed without considering s_i. We do not follow this approach here since it adds another small complication to the GAP, and has only a small effect. With or without this modification, the check in (2) is conceptually easy to apply and might be more appropriate than (3) since normality in the examination marks data set is not checked. If (2) is not satisfied for any given module, then either a more sophisticated test such as (3) should be applied, or the module grades should be adjusted (adjustments are discussed subsequently). On the other hand, there could be grounds for leaving the module results untouched and accepting that the variance of grades in that module is larger or smaller than in other modules.

In Step 2, the average and standard deviation is computed for each student's results. These are used in Step 3 to compute, for each student, normalised results. This step is simply to allow for easy searching of each student's (possibly) better-than-expected and worse-than-expected results.

Next, in Step 4, the average and standard deviation (or variance) of each module's normalised results are calculated. These statistics are not used directly; rather they are computed to give an overall view of the variability between the outcomes of particular modules. For example, it highlights which modules have high and low averages. The variances of each module should cluster around unity since variance compatibility has been checked in Step 1. If module grades were adjusted subsequently, then the change in these summary statistics would confirm the action taken.

The *target* standard deviation selected in Step 5 should be large enough to assist in detection of outlier results. A typical choice would be select a *target* of 1. If the module results were normally distributed with mean 0 and standard deviation of 1, then there would be about 16% of the grades above $+target$ and below $-target$. The numbers of results outside $\pm target$ are counted in Step 6.

In Step 7 the statistic $\hat{T}$ is used in the modified sign test [16, 17], which tests whether $A = B$. By carrying out this test, we are checking whether the proportion of students who did well in the module balances the proportion that did not. If a skewed distribution was expected, then the test could be



```
Start
  1. Test whether module variances are homogeneous. If necessary, adjust module
     grades appropriately.
  2. Calculate the average and standard deviation for each student's results.
  3. Transform each student's results to standard variates (mean of 0 and standard
     deviation of 1).
  4. Calculate the average and standard deviation for each module's normalized re-
     sults.
  5. Select a target standard deviation to be used to identify better or worse grades.
   For each module
  6. For each student, identify whether their grade in the particular module is
     a) Above +target: count the number of these (A).
   or
     b) Below -target: count the number of these (B).
  7. Calculate $\hat{T} = \frac{|A - B|}{\sqrt{A + B}}$. If $\hat{T}$ > 2 (significance level 0.05) or 1.6 (significance
     level 0.1).
   Either
     a) $A - B < 0$; module results are too low and grades should be increased.
   or
     b) $A - B > 0$; module results are too high and grades should be decreased.
  8. Adjust grades
End
```

Fig. 2. Details of a Grade Adjustment Procedure (GAP) to identify modules where moderation of grades is warranted.

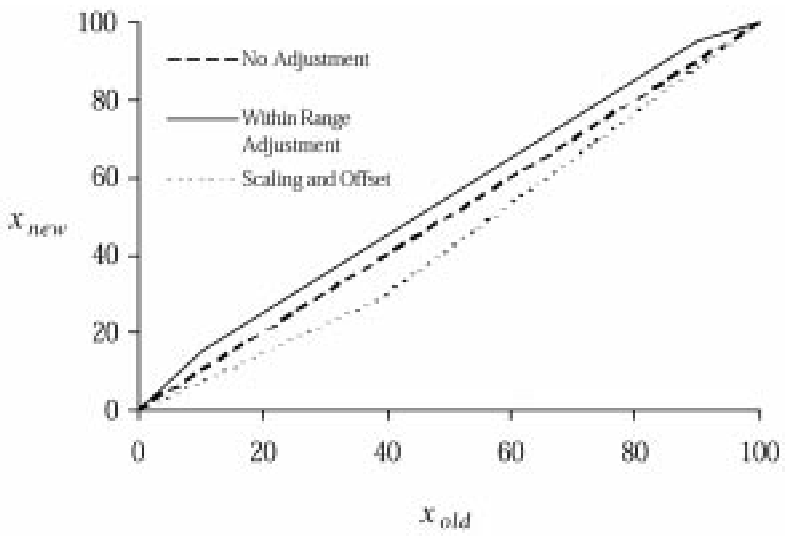


Fig. 3. Various grade adjustments (x_{old} is the original grade and x_{new} is the adjusted grade). No adjustment is given by the 1:1 line (dashes). The solid line shows a nearly variance-preserving offset with small adjustments at the end of the range to remove the possibility of outliers. The dotted line shows a combination of straight-line adjustments, with changes to both the module variance and mean.

adjusted to account for this. The test does not rely on distributional assumptions, although it is assumed that the outcomes are independent, which is a reasonable assumption for grades of individual students. The test $\hat{T} > 2$ is significant at the 0.05 level, and the hypothesis is rejected. It is significant at the 0.1 level for $\hat{T} > 1.6$.

At this juncture it is worth mentioning that some modules might be expected to have better outcomes than others. Typically, modules that rely mainly on coursework material such as assignments have outcomes that are much higher than modules that are graded wholly on supervised examinations [18]. The difference in each mode of assessment can be estimated on current or historical data. Our experience is that modules relying on coursework for assessment have individual student outcomes around 0.5 of a standard deviation above supervised examinations. Assuming that (on average) higher grades are acceptable for such modules, this effect is easily included in Step 7. For each module, the proportion of the grade allocated to coursework material is denoted as p. Then, in Step 7a, replace *target* by $target + 0.5p$ and in Step 7b replace $-target$ by $-target + 0.5p$. On the other hand, if the examination board deems it unacceptable that coursework material should have higher average grades, the test in Step 7 will identify clearly any such modules.

Grade adjustment

Once a decision has been taken to adjust module grades, the question of the amount of adjustment naturally arises. This is a policy decision that should be decided prior to taking action. Below, we adopt the criterion that grades should be adjusted minimally to achieve the goal of satisfying the statistical test imposed.

There are two opportunities to adjust grades in Fig. 2, at Steps 1 and 8. At Step 1, the goal is to adjust the variance of a particular module (or modules). If student grades in a module are denoted as x, then a change in variance is achieved using:

$$x_{new} = a\, x_{old}. \tag{4}$$

where a is the adjustment factor. Clearly, the transformation in (4) will change the module average by the factor a. This is not important since below we discuss how the module grades will be offset to adjust the average. Alternatively, we can adjust the variance while maintaining the calculated module average by altering (4) to:

$$x_{new} = a\, x_{old} + (1 - a)\overline{x}_{old}. \tag{5}$$

For both (4) and (5), the relationship between the estimated variances (Var) is:

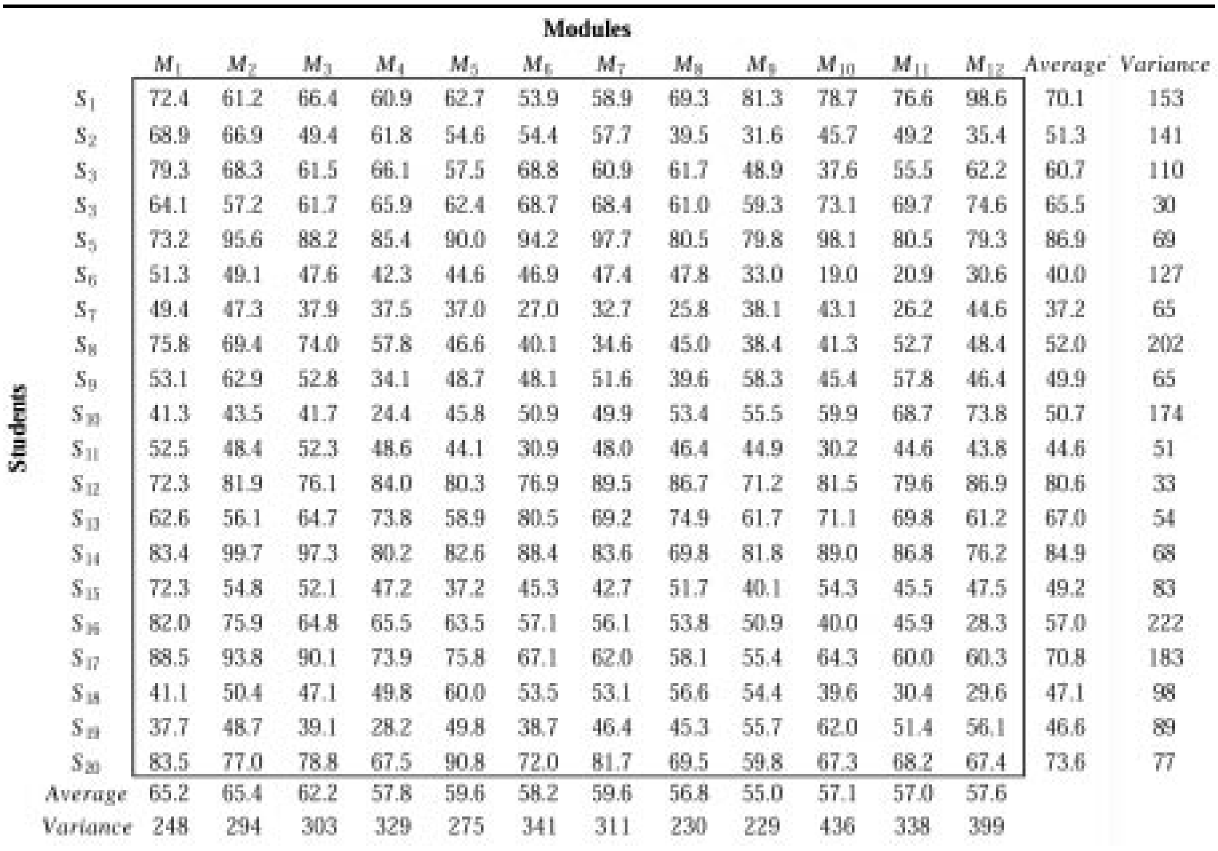
$$\mathrm{Var}(x_{new}) = a^2\, \mathrm{Var}(x_{old}). \tag{6}$$

Since the amount of the adjustment is known, an approximate value for a is easily calculated to satisfy the condition in (2). Note that (6) is not exact in terms of computing a since the average variance ($\overline{s}^2$) computed in (1), and used in (2), will change when any module grades are adjusted.

The other place where adjustment is suggested is in Step 8. At this stage, it is expected that only an offset is needed, i.e., $x_{new} = +x_{old}\, b$. If more than one module is identified as needing adjustment, only the module with the largest $\hat{T}$ is adjusted by a fixed, small amount (say 0.5%). Adjustment continues until the imposed statistical test is satisfied.

Following adjustment, the adjusted grades are examined in case any lie outside the range 0–100%.

Table 1. Synthesised data set of student results and modules: *M* denotes a set of module results and *S* denotes a set of student results



		Modules													
		M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	*Average*	*Variance*
Students	S_1	72.4	61.2	66.4	60.9	62.7	53.9	58.9	69.3	81.3	78.7	76.6	98.6	70.1	153
	S_2	68.9	66.9	49.4	61.8	54.6	54.4	57.7	39.5	31.6	45.7	49.2	35.4	51.3	141
	S_3	79.3	68.3	61.5	66.1	57.5	68.8	60.9	61.7	48.9	37.6	55.5	62.2	60.7	110
	S_3	64.1	57.2	61.7	65.9	62.4	68.7	68.4	61.0	59.3	73.1	69.7	74.6	65.5	30
	S_5	73.2	95.6	88.2	85.4	90.0	94.2	97.7	80.5	79.8	98.1	80.5	79.3	86.9	69
	S_6	51.3	49.1	47.6	42.3	44.6	46.9	47.4	47.8	33.0	19.0	20.9	30.6	40.0	127
	S_7	49.4	47.3	37.9	37.5	37.0	27.0	32.7	25.8	38.1	43.1	26.2	44.6	37.2	65
	S_8	75.8	69.4	74.0	57.8	46.6	40.1	34.6	45.0	38.4	41.3	52.7	48.4	52.0	202
	S_9	53.1	62.9	52.8	34.1	48.7	48.1	51.6	39.6	58.3	45.4	57.8	46.4	49.9	65
	S_{10}	41.3	43.5	41.7	24.4	45.8	50.9	49.9	53.4	55.5	59.9	68.7	73.8	50.7	174
	S_{11}	52.5	48.4	52.3	48.6	44.1	30.9	48.0	46.4	44.9	30.2	44.6	43.8	44.6	51
	S_{12}	72.3	81.9	76.1	84.0	80.3	76.9	89.5	86.7	71.2	81.5	79.6	86.9	80.6	33
	S_{13}	62.6	56.1	64.7	73.8	58.9	80.5	69.2	74.9	61.7	71.1	69.8	61.2	67.0	54
	S_{14}	83.4	99.7	97.3	80.2	82.6	88.4	83.6	69.8	81.8	89.0	86.8	76.2	84.9	68
	S_{15}	72.3	54.8	52.1	47.2	37.2	45.3	42.7	51.7	40.1	54.3	45.5	47.5	49.2	83
	S_{16}	82.0	75.9	64.8	65.5	63.5	57.1	56.1	53.8	50.9	40.0	45.9	28.3	57.0	222
	S_{17}	88.5	93.8	90.1	73.9	75.8	67.1	62.0	58.1	55.4	64.3	60.0	60.3	70.8	183
	S_{18}	41.1	50.4	47.1	49.8	60.0	53.5	53.1	56.6	54.4	39.6	30.4	29.6	47.1	98
	S_{19}	37.7	48.7	39.1	28.2	49.8	38.7	46.4	45.3	55.7	62.0	51.4	56.1	46.6	89
	S_{20}	83.5	77.0	78.8	67.5	90.8	72.0	81.7	69.5	59.8	67.3	68.2	67.4	73.6	77
	Average	65.2	65.4	62.2	57.8	59.6	58.2	59.6	56.8	55.0	57.1	57.0	57.6		
	Variance	248	294	303	329	275	341	311	230	229	436	338	399		

If values outside this range are detected, then either they are moved to the appropriate boundary or a different correction is needed. In order to preserve the variance, it is suggested that a modified offset adjustment is used. An example is given in Fig. 3. This figure shows a uniform offset over most the range (in this case 5–95%), with small adjustments made close to the boundaries (in practice only one end, 0 or 100%, would need this limitation imposed). Note, also, that the Step 1 variance adjustment described by (6) might lead to an out-of-range adjustment. It is perfectly acceptable to let this situation continue until the completion of Step 8, as the Step 8 adjustment would tend to bring the outliers back into range. Alternatively, the variance adjustment in Step 1 can be changed to include an offset and scaling, as given by the dotted line in Fig. 3. This type of adjustment might in any case be more applicable to modules where student grades are strongly skewed. Because the 1:1 mapping has been altered to 2 straight lines, the variance of the transformed data will depend on the location of the majority of the scores in the adjusted module. For example, the dotted line in Fig. 3 will tend to increase the overall variance (relative to the initial module variance) if most of the results lie to the right of the slope discontinuity.

APPLICATION

The procedure outlined above is demonstrated on an artificial data set (Table 1). Two different methods of analysis are used: (1) the GAP described above and (2) an ANOVA-based analysis. For (1), the approach taken was exactly as described above. For (2), several steps were taken; these were: (i) data were checked for normality, (ii) variances were checked using equation (3) and (iii) a one-way repeated-measures ANOVA was performed.

The construction of the data set followed several steps, with the aim of making the data distribution somewhat non-normal, so as to emulate our experience the type of data sets that are the outcome of academic examination procedures in engineering degree programmes, but not so non-normal that use of the ANOVA was precluded. All calculations were carried out in Microsoft EXCEL.

Data were generated using the four steps:

1. A set of standard normal variates, $z_{i,j}$, $i = 1, \ldots, 20$; $j = 1, \ldots, 12$; was generated.
2. These were transformed into correlated variates (i.e., correlated in 'j', but independent in 'i'), $c_{i,j}$, using the formula [19]: $c_{i,j+1} = r\,c_{i,j} + z_{i,j}(1—r^2)^{1/2}$,

$i = 1, \ldots, 20$; $j = 1, \ldots, 11$; $c_{i,1} = z_{i,1}$. The 20 independent sequences of 12 correlated variates were computed; these were used below to simulate grades for 20 students. A correlation coefficient, r, of 0.95, was used.

3. Another two (20×12) sets of random variates were generated; these were uniformly distributed over (0,1). Call these Set u and Set v. Based on the sequences generated in 2) above, synthetic student grades were then calculated using: $g_{i,j} = c_{i,j}(11 + 20u_{i,j}) + 56 + 10v_{i,j}$, $i = 1, \ldots, 20$; $j = 1, \ldots, 12$.
4. Various entries were randomly removed.

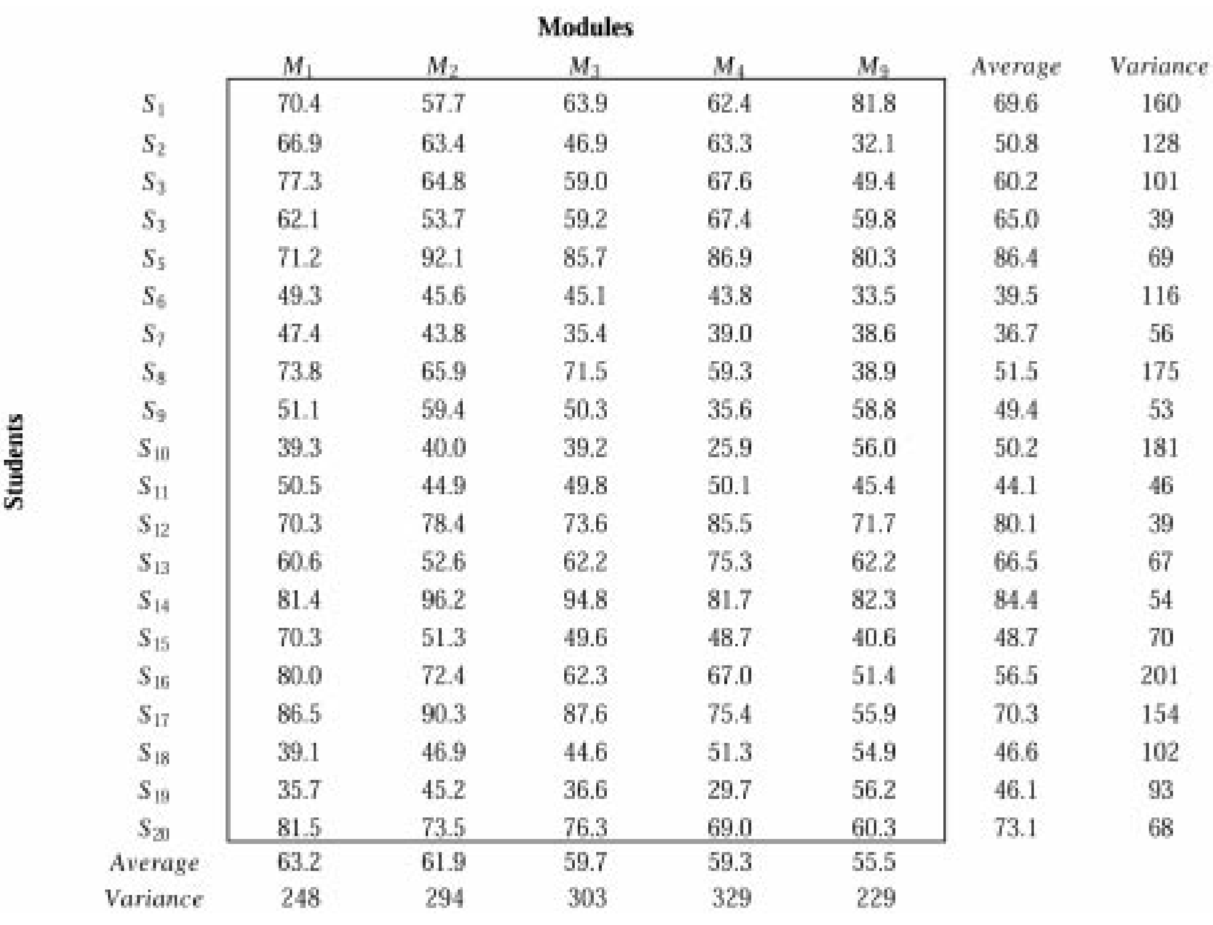
One realisation of the sequence of steps (1– 3) used in the subsequent analysis is given in Table 1. There are 12 modules ($M_1, \ldots, M_{12}$) taken by 20 students ($S_1, \ldots, S_{20}$). A cursory examination of the module outcomes does not reveal any markedly untoward results, particularly with regard to the module averages. The module variances vary by a factor of 2 approximately (compare M_8 and M_{10}). Although the GAP is not reliant on normality of the underlying data, normality is a requirement for application of the ANOVA procedure. The module data shown in Table 1 were checked for normality using the χ^2 goodness-of-fit test. Each module's results satisfied the test at the 0.1 significance level.

The simplified procedure described above was carried out. In the analysis we took in all steps the $\alpha = 0.1$ level of significance. The check in (2) showed that no variance adjustments were needed. This result was confirmed by applying the χ^2 test in (3). The largest statistic was computed for M_{10}, with a value of 26.6. This is less than the critical value of $\chi^2_{19,0.1} = 27.2$. Next, Step 7 in Fig. 2 revealed that modules M_1, M_2, M_3, M_4 and M_9 should be adjusted. The relevant $\bar{T}$ values were within range when these modules were offset by −2, −3.5, −2.5, 1.5 and 0.5%, respectively. The modified set of grades is shown in Table 2.

Clearly, the adjustments made between Table 1 and Table 2 are modest. We can examine these changes in more detail by applying a one-way repeated-measures ANOVA to the data in each table. Results for Table 1 are given in Table 3, while those for Table 2 are given in Table 4.

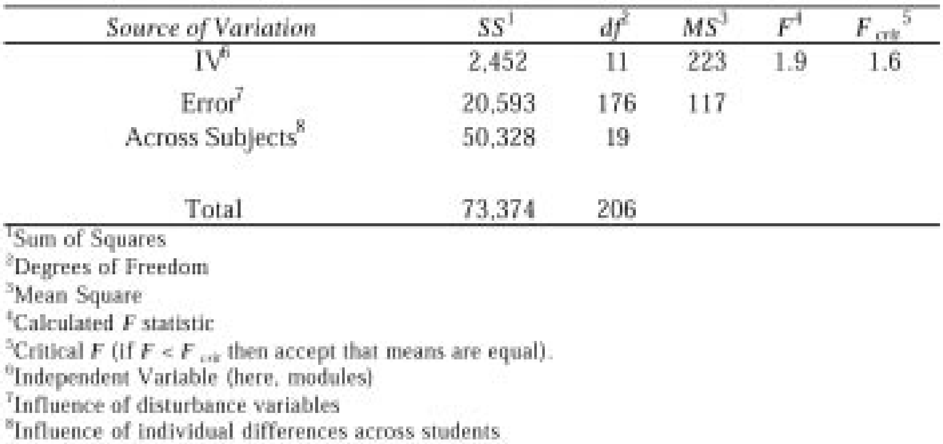
After this preliminary analysis, the full set of grades in Table 1 was modified by randomly removing results. This step was to obtain a set of results that more closely replicates academic outcomes. The modified set of results and

Table 2. Grades from Table 1 after adjustment (M_1, M_2, M_3, M_4 and M_9 modified, other results as in Table 1)



		Modules						
		M_1	M_2	M_3	M_4	M_9	*Average*	*Variance*
Students	S_1	70.4	57.7	63.9	62.4	81.8	69.6	160
	S_2	66.9	63.4	46.9	63.3	32.1	50.8	128
	S_3	77.3	64.8	59.0	67.6	49.4	60.2	101
	S_3	62.1	53.7	59.2	67.4	59.8	65.0	39
	S_5	71.2	92.1	85.7	86.9	80.3	86.4	69
	S_6	49.3	45.6	45.1	43.8	33.5	39.5	116
	S_7	47.4	43.8	35.4	39.0	38.6	36.7	56
	S_8	73.8	65.9	71.5	59.3	38.9	51.5	175
	S_9	51.1	59.4	50.3	35.6	58.8	49.4	53
	S_{10}	39.3	40.0	39.2	25.9	56.0	50.2	181
	S_{11}	50.5	44.9	49.8	50.1	45.4	44.1	46
	S_{12}	70.3	78.4	73.6	85.5	71.7	80.1	39
	S_{13}	60.6	52.6	62.2	75.3	62.2	66.5	67
	S_{14}	81.4	96.2	94.8	81.7	82.3	84.4	54
	S_{15}	70.3	51.3	49.6	48.7	40.6	48.7	70
	S_{16}	80.0	72.4	62.3	67.0	51.4	56.5	201
	S_{17}	86.5	90.3	87.6	75.4	55.9	70.3	154
	S_{18}	39.1	46.9	44.6	51.3	54.9	46.6	102
	S_{19}	35.7	45.2	36.6	29.7	56.2	46.1	93
	S_{20}	81.5	73.5	76.3	69.0	60.3	73.1	68
	Average	63.2	61.9	59.7	59.3	55.5		
	Variance	248	294	303	329	229		

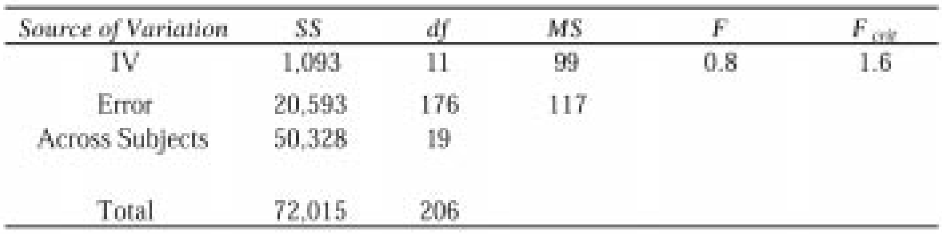
Table 3. One-way repeated-measures ANOVA results for grades in Table 1 (0.1 significance level): notation for this and other ANOVA tables follows that of Jaccard and Becker [12]



Source of Variation	*SS*[1]	*df*[2]	*MS*[3]	*F*[4]	*F*$_{crit}$[5]
IV[6]	2,452	11	223	1.9	1.6
Error[7]	20,593	176	117		
Across Subjects[8]	50,328	19			
Total	73,374	206			

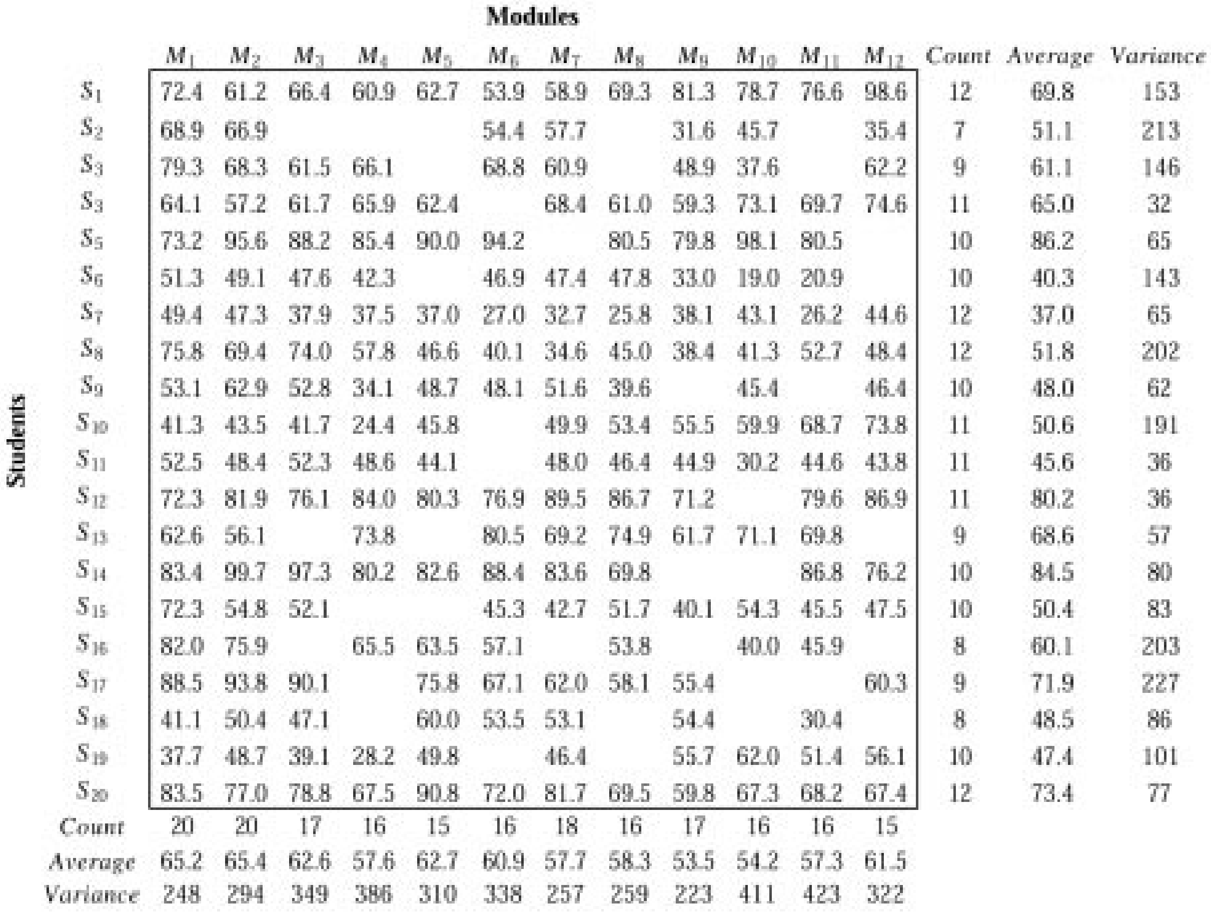
[1] Sum of Squares
[2] Degrees of Freedom
[3] Mean Square
[4] Calculated *F* statistic
[5] Critical *F* (if $F < F_{crit}$ then accept that means are equal).
[6] Independent Variable (here, modules)
[7] Influence of disturbance variables
[8] Influence of individual differences across students

Table 4. One-way repeated-measures ANOVA results for grades in Table 2 (0.1 significance level)



Source of Variation	*SS*	*df*	*MS*	*F*	*F*$_{crit}$
IV	1,093	11	99	0.8	1.6
Error	20,593	176	117		
Across Subjects	50,328	19			
Total	72,015	206			

Table 5. New set of student grades generated by randomly removing entries from Table 1



		Modules													
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}	*Count*	*Average*	*Variance*
Students S_1	72.4	61.2	66.4	60.9	62.7	53.9	58.9	69.3	81.3	78.7	76.6	98.6	12	69.8	153
S_2	68.9	66.9				54.4	57.7		31.6	45.7		35.4	7	51.1	213
S_3	79.3	68.3	61.5	66.1		68.8	60.9		48.9	37.6		62.2	9	61.1	146
S_3	64.1	57.2	61.7	65.9	62.4		68.4	61.0	59.3	73.1	69.7	74.6	11	65.0	32
S_5	73.2	95.6	88.2	85.4	90.0	94.2		80.5	79.8	98.1	80.5		10	86.2	65
S_6	51.3	49.1	47.6	42.3		46.9	47.4	47.8	33.0	19.0	20.9		10	40.3	143
S_7	49.4	47.3	37.9	37.5	37.0	27.0	32.7	25.8	38.1	43.1	26.2	44.6	12	37.0	65
S_8	75.8	69.4	74.0	57.8	46.6	40.1	34.6	45.0	38.4	41.3	52.7	48.4	12	51.8	202
S_9	53.1	62.9	52.8	34.1	48.7	48.1	51.6	39.6		45.4		46.4	10	48.0	62
S_{10}	41.3	43.5	41.7	24.4	45.8		49.9	53.4	55.5	59.9	68.7	73.8	11	50.6	191
S_{11}	52.5	48.4	52.3	48.6	44.1		48.0	46.4	44.9	30.2	44.6	43.8	11	45.6	36
S_{12}	72.3	81.9	76.1	84.0	80.3	76.9	89.5	86.7	71.2		79.6	86.9	11	80.2	36
S_{13}	62.6	56.1		73.8		80.5	69.2	74.9	61.7	71.1	69.8		9	68.6	57
S_{14}	83.4	99.7	97.3	80.2	82.6	88.4	83.6	69.8			86.8	76.2	10	84.5	80
S_{15}	72.3	54.8	52.1			45.3	42.7	51.7	40.1	54.3	45.5	47.5	10	50.4	83
S_{16}	82.0	75.9		65.5	63.5	57.1		53.8		40.0	45.9		8	60.1	203
S_{17}	88.5	93.8	90.1		75.8	67.1	62.0	58.1	55.4			60.3	9	71.9	227
S_{18}	41.1	50.4	47.1		60.0	53.5	53.1		54.4		30.4		8	48.5	86
S_{19}	37.7	48.7	39.1	28.2	49.8		46.4		55.7	62.0	51.4	56.1	10	47.4	101
S_{20}	83.5	77.0	78.8	67.5	90.8	72.0	81.7	69.5	59.8	67.3	68.2	67.4	12	73.4	77
Count	20	20	17	16	15	16	18	16	17	16	16	15			
Average	65.2	65.4	62.6	57.6	62.7	60.9	57.7	58.3	53.5	54.2	57.3	61.5			
Variance	248	294	349	386	310	338	257	259	223	411	423	322			

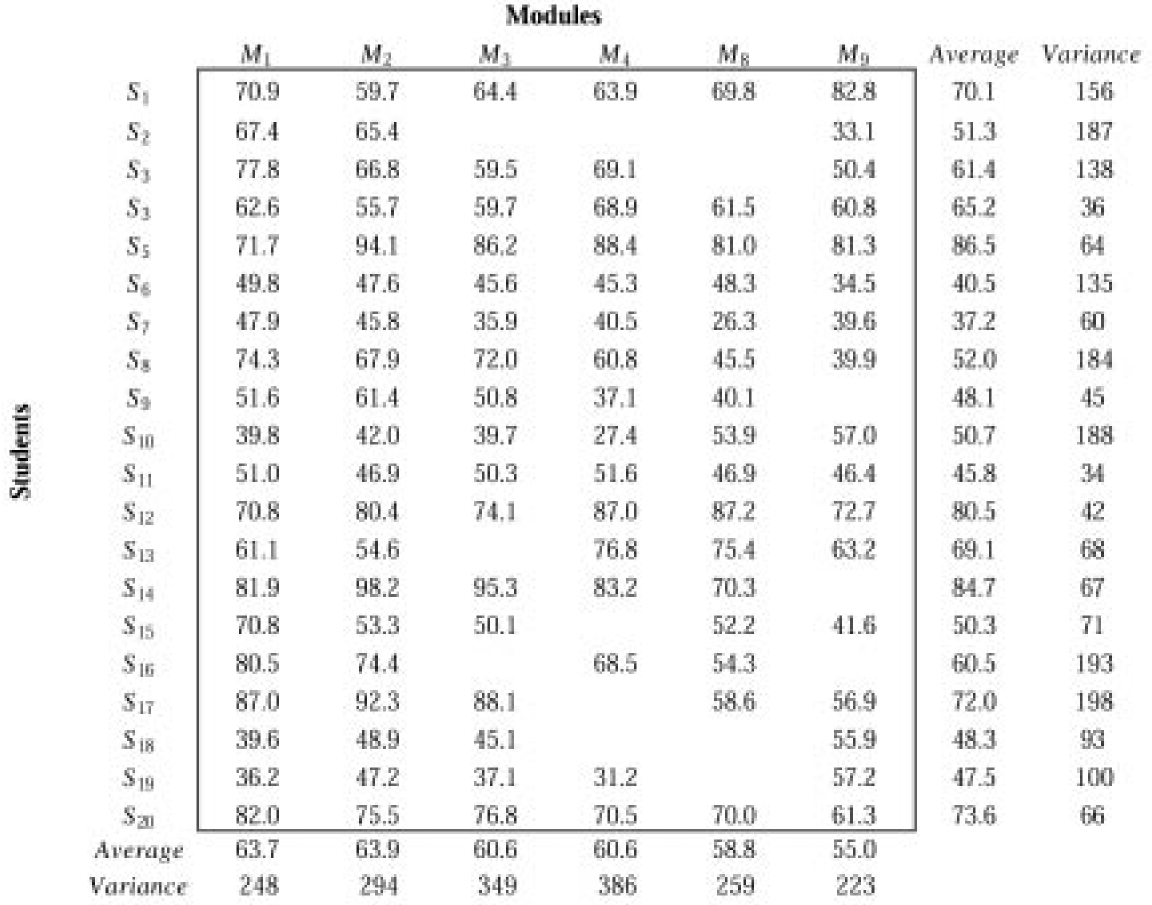
summary statistics are shown in Table 5. All students take two compulsory modules, M_1 and M_2, with the other modules taken by a portion of the cohort. In addition, several students took all available modules. The averages of the results in the compulsory modules are higher than the other modules. The averages of several other modules have increased, relative to Table 1. Also relative to Table 1, variances in Table 5 have changed slightly, but overall there is little difference.

The procedure outlined above is implemented again. Again, the variance check, (2), indicated that no modules required adjustment. Similarly, the χ^2 test in (3) was satisfied for all modules. In contrast to the data in Table 1, where M_{10} produced the largest statistic, for Table 5 M_{11} produced the closest statistic (19.9) to the critical χ^2 value (22.3). Again, the $\hat{T}$ statistic was not satisfied for modules M_1, M_2, M_3, M_4 and M_9. Also, it was not satisfied for M_8. The $\hat{T}$ -test was satisfied by adjusting each module grade as follows: M_1 and M_2 by −1.5%, M_3 by −2%, M_4 by 3%, M_8 by 0.5% and M_9 by 1.5%. The modified grades are shown in Table 6.

The results in Table 5 and Table 6 were subjected to a one-way repeated-measures ANOVA. Unlike the previous case, where the entire grade matrix was filled, the one-way repeated-measures ANOVA cannot be applied where the matrix has missing entries, as is the case in Table 5 and Table 6. Thus, the missing data has to be replaced in order to carry out the analysis. Here, since we have the missing data (Table 1), we could simply replace it. However, in practice the data would not be known so a fair comparison of the approach presented and the results of the one-way repeated-measures ANOVA should entail replacing the missing data following a standard approach, before the ANOVA is performed.

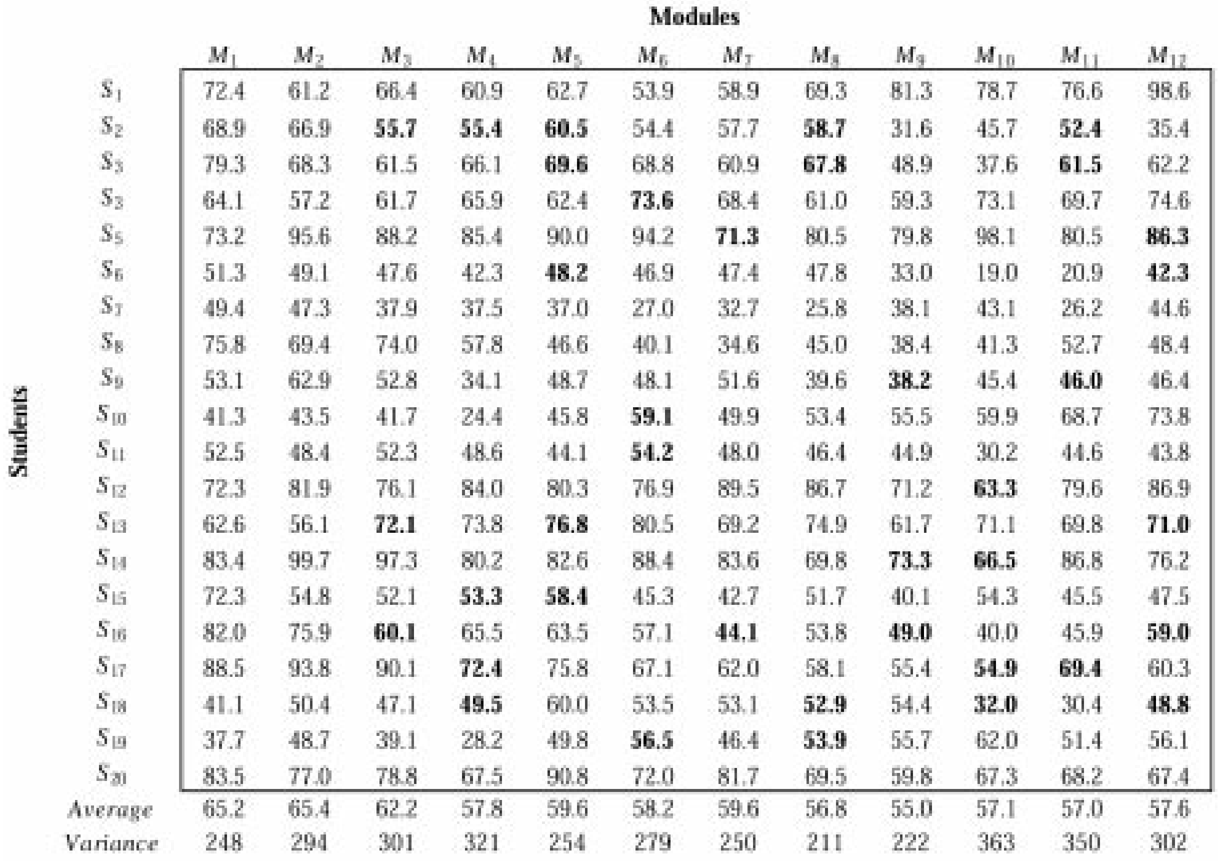
Kirk [20] recommends replacing the data such that the error sum-of-squares is minimised (while maintaining the module averages). For the data matrix shown in Table 5, the minimisation was performed, with results as given in Table 7. Similarly, the missing data from Table 6 were replaced, with results as given in Table 8. The one-way repeated-measures ANOVA was applied, in turn, to the grades in Table 7 and Table 8. In the ANOVA, degrees of freedom were reduced to account for the replaced data. This test is sensitive to departures of circularity [20]. Where circularity is in doubt, an approximate F statistic should be used, with (further) reduced degrees of freedom. Because missing grades were added, we used an

Table 6. Results from Table 5 after modification due to application of the GAP; modules not included have not been altered



	Modules							
Students	M_1	M_2	M_3	M_4	M_8	M_9	*Average*	*Variance*
S_1	70.9	59.7	64.4	63.9	69.8	82.8	70.1	156
S_2	67.4	65.4				33.1	51.3	187
S_3	77.8	66.8	59.5	69.1		50.4	61.4	138
S_3	62.6	55.7	59.7	68.9	61.5	60.8	65.2	36
S_5	71.7	94.1	86.2	88.4	81.0	81.3	86.5	64
S_6	49.8	47.6	45.6	45.3	48.3	34.5	40.5	135
S_7	47.9	45.8	35.9	40.5	26.3	39.6	37.2	60
S_8	74.3	67.9	72.0	60.8	45.5	39.9	52.0	184
S_9	51.6	61.4	50.8	37.1	40.1		48.1	45
S_{10}	39.8	42.0	39.7	27.4	53.9	57.0	50.7	188
S_{11}	51.0	46.9	50.3	51.6	46.9	46.4	45.8	34
S_{12}	70.8	80.4	74.1	87.0	87.2	72.7	80.5	42
S_{13}	61.1	54.6		76.8	75.4	63.2	69.1	68
S_{14}	81.9	98.2	95.3	83.2	70.3		84.7	67
S_{15}	70.8	53.3	50.1		52.2	41.6	50.3	71
S_{16}	80.5	74.4		68.5	54.3		60.5	193
S_{17}	87.0	92.3	88.1		58.6	56.9	72.0	198
S_{18}	39.6	48.9	45.1			55.9	48.3	93
S_{19}	36.2	47.2	37.1	31.2		57.2	47.5	100
S_{20}	82.0	75.5	76.8	70.5	70.0	61.3	73.6	66
Average	63.7	63.9	60.6	60.6	58.8	55.0		
Variance	248	294	349	386	259	223		

Table 7. Grades from Table 5 after filling in missing data (bold face indicates filled-in data)



		Modules											
		M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
Students	S_1	72.4	61.2	66.4	60.9	62.7	53.9	58.9	69.3	81.3	78.7	76.6	98.6
	S_2	68.9	66.9	**55.7**	**55.4**	**60.5**	54.4	57.7	**58.7**	31.6	45.7	**52.4**	35.4
	S_3	79.3	68.3	61.5	66.1	**69.6**	68.8	60.9	**67.8**	48.9	37.6	**61.5**	62.2
	S_4	64.1	57.2	61.7	65.9	62.4	**73.6**	68.4	61.0	59.3	73.1	69.7	74.6
	S_5	73.2	95.6	88.2	85.4	90.0	94.2	**71.3**	80.5	79.8	98.1	80.5	**86.3**
	S_6	51.3	49.1	47.6	42.3	**48.2**	46.9	47.4	47.8	33.0	19.0	20.9	**42.3**
	S_7	49.4	47.3	37.9	37.5	37.0	27.0	32.7	25.8	38.1	43.1	26.2	44.6
	S_8	75.8	69.4	74.0	57.8	46.6	40.1	34.6	45.0	38.4	41.3	52.7	48.4
	S_9	53.1	62.9	52.8	34.1	48.7	48.1	51.6	39.6	**38.2**	45.4	**46.0**	46.4
	S_{10}	41.3	43.5	41.7	24.4	45.8	**59.1**	49.9	53.4	55.5	59.9	68.7	73.8
	S_{11}	52.5	48.4	52.3	48.6	44.1	**54.2**	48.0	46.4	44.9	30.2	44.6	43.8
	S_{12}	72.3	81.9	76.1	84.0	80.3	76.9	89.5	86.7	71.2	**63.3**	79.6	86.9
	S_{13}	62.6	56.1	**72.1**	73.8	**76.8**	80.5	69.2	74.9	61.7	71.1	69.8	**71.0**
	S_{14}	83.4	99.7	97.3	80.2	82.6	88.4	83.6	69.8	**73.3**	**66.5**	86.8	76.2
	S_{15}	72.3	54.8	52.1	**53.3**	**58.4**	45.3	42.7	51.7	40.1	54.3	45.5	47.5
	S_{16}	82.0	75.9	**60.1**	65.5	63.5	57.1	**44.1**	53.8	**49.0**	40.0	45.9	**59.0**
	S_{17}	88.5	93.8	90.1	**72.4**	75.8	67.1	62.0	58.1	55.4	**54.9**	**69.4**	60.3
	S_{18}	41.1	50.4	47.1	**49.5**	60.0	53.5	53.1	**52.9**	54.4	**32.0**	30.4	**48.8**
	S_{19}	37.7	48.7	39.1	28.2	49.8	**56.5**	46.4	**53.9**	55.7	62.0	51.4	56.1
	S_{20}	83.5	77.0	78.8	67.5	90.8	72.0	81.7	69.5	59.8	67.3	68.2	67.4
	Average	65.2	65.4	62.2	57.8	59.6	58.2	59.6	56.8	55.0	57.1	57.0	57.6
	Variance	248	294	301	321	254	279	250	211	222	363	350	302

existing approach [20, Table 6.4–2] to compute the reduced degrees of freedom factor $\hat{\theta}$. Results making use of this factor in the one-way repeated-measures ANOVA test of the grades in Table 7 and Table 8 are shown in Table 9 and Table 10, respectively.

DISCUSSION

We consider first the analysis of the full set of results, as given in Table 1. The GAP outlined in Fig. 2 identified that no module variances were in need of adjustment. This was confirmed by the test based on the χ^2 distribution. If the χ^2 test is available, as it is in Microsoft EXCEL, the simplified test in (2) could be replaced by (3). However, strictly speaking the data should be checked for normality before applying the test. Our experience is that the test in (2) yields results that are acceptable. This discussion also applies to the variance adjustment check carried out on the modified data in Table 5, so the variance adjustment procedure is not mentioned further.

Application of the $\hat{T}$ statistic test in the GAP indicated the need to offset several sets of module grades. We note that the average before adjustment was 59.3, after adjustment it was 58.8. All the adjustments were in the direction of the average. Two changes are worthy of note as they bring out features of the GAP.

First, the average for M_4 increased from 57.8 to 59.3, i.e., above the final average. The process of adjustment involved changing module grades in increments of 0.5%, re-evaluating $\hat{T}$ for each module, with the sequence of adjustments based on the largest $\hat{T}$. This stepwise adjustment procedure is simple, but yields adjusted grades that are not unique in the sense that other combinations of changes are feasible. For example, the average of M_4 in Table 5 could be adjusted back to its original (starting) value and still satisfy the $\hat{T}$ statistic test. In our applications of the GAP, an overarching principle is to change grades as little as possible, so in practice the result for M_4 would stay at its original value. Because the GAP yields results that are (possibly) non-unique, after adjusting grades it is recommended that module offsets be perturbed in the direction of the original average to check whether a smaller offset would suffice.

The second feature exhibited by the results in Table 5 is that the averages of M_1 and M_2 are further away from the overall average than that of M_3, yet the average of the latter was adjusted downwards by 2.5%, which is nearly as much as the adjustments to the former (downward adjustments of 2% and 3.5%, respectively). This can occur since the GAP is not directly concerned

Table 8. Grades from Table 6 after filling in missing data (bold face indicates filled-in data)

		Modules											
		M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}	M_{12}
Students	S_1	70.9	59.7	64.4	63.9	62.7	53.9	58.9	69.8	82.8	78.7	76.6	98.6
	S_2	67.4	65.4	**53.7**	**58.4**	**60.5**	54.4	57.7	**59.2**	33.1	45.7	**52.4**	35.4
	S_3	77.8	66.8	59.5	69.1	**69.6**	68.8	60.9	**68.3**	50.4	37.6	**61.5**	62.2
	S_3	62.6	55.7	59.7	68.9	62.4	**73.6**	68.4	61.5	60.8	73.1	69.7	74.6
	S_5	71.7	94.1	86.2	88.4	90.0	94.2	**71.3**	81.0	81.3	98.1	80.5	**86.3**
	S_6	49.8	47.6	45.6	45.3	**48.2**	46.9	47.4	48.3	34.5	19.0	20.9	**42.3**
	S_7	47.9	45.8	35.9	40.5	37.0	27.0	32.7	26.3	39.6	43.1	26.2	44.6
	S_8	74.3	67.9	72.0	60.8	46.6	40.1	34.6	45.5	39.9	41.3	52.7	48.4
	S_9	51.6	61.4	50.8	37.1	48.7	48.1	51.6	40.1	**39.7**	45.4	**46.0**	46.4
	S_{10}	39.8	42.0	39.7	27.4	45.8	**59.1**	49.9	53.9	57.0	59.9	68.7	73.8
	S_{11}	51.0	46.9	50.3	51.6	44.1	**54.2**	48.0	46.9	46.4	30.2	44.6	43.8
	S_{12}	70.8	80.4	74.1	87.0	80.3	76.9	89.5	87.2	72.7	**63.3**	79.6	86.9
	S_{13}	61.1	54.6	**70.1**	76.8	**76.8**	80.5	69.2	75.4	63.2	71.1	69.8	**71.0**
	S_{14}	81.9	98.2	95.3	83.2	82.6	88.4	83.6	70.3	**74.8**	**66.5**	86.8	76.2
	S_{15}	70.8	53.3	50.1	**56.3**	**58.4**	45.3	42.7	52.2	41.6	54.3	45.5	47.5
	S_{16}	80.5	74.4	**58.1**	68.5	63.5	57.1	**44.1**	54.3	**50.4**	40.0	45.9	**59.0**
	S_{17}	87.0	92.3	88.1	**75.4**	75.8	67.1	62.0	58.6	56.9	**54.9**	**69.4**	60.3
	S_{18}	39.6	48.9	45.1	**52.5**	60.0	53.5	53.1	**53.4**	55.9	**32.0**	30.4	**48.8**
	S_{19}	36.2	47.2	37.1	31.2	49.8	**56.5**	46.4	**54.4**	57.2	62.0	51.4	56.1
	S_{20}	82.0	75.5	76.8	70.5	90.8	72.0	81.7	70.0	61.3	67.3	68.2	67.4
	Average	63.7	63.9	60.6	60.6	62.7	60.9	57.7	58.8	55.0	54.2	57.3	61.5
	Variance	248	294	301	321	254	279	250	211	222	363	350	302

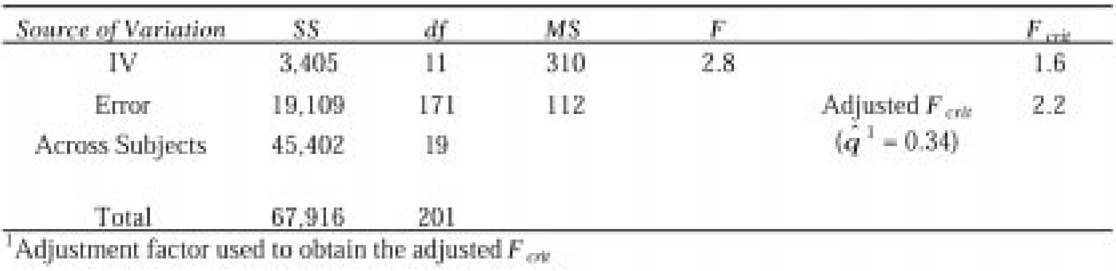
with adjusting module averages. Rather, because it is student-centred the GAP is aimed at altering imbalances in overall student performance, where the latter is calculated on an individual basis. It is the aggregate of the better-than-expected and worse-than-expected individual performances in each module that is checked in applying the $\hat{T}$ statistic. Clearly, adjusting the average of a module will adjust this aggregate, which is why module results are offset. Indeed, it is the ability of the $\hat{T}$ statistic to uncover overall aggregate performance that makes it a useful tool in identifying modules for which very high or very low averages are acceptable. In this context, simple examination of averages alone, in the absence of aggregate performance, would not be a suitable way of identifying acceptably high or low averages. In the case of M_3, it could be argued that since all the students took that module, there should be less adjustment to it and more to M_1 and M_2. However, that assertion relies on the notion that student performance in each module should be somehow identical. Our starting point is that student performance in individual modules should not be identical; rather, that it should be expected to vary within reasonable limits.

Because the data in Table 1 satisfy the normality assumption, we can apply the one-way repeated-measures ANOVA to test the hypothesis of equality of module means. The results in Table 3 give the F statistic of 1.9 that exceeds the critical F value, F_{crit}, of 1.6. That is, the ANOVA outcome is that we would reject the hypothesis of equality of module means in Table 1, in agreement with the GAP

Next, the one-way repeated-measures ANOVA was used to check equality of means for the adjusted grades in Table 2, with results presented in Table 4. In this case, the outcome is that the null hypothesis is not rejected, and so module means can be accepted as being equal. This outcome confirms that the GAP adjustment is reasonable, and that it has achieved its aim of moderating the module outcomes such that they are comparable.

For a slightly more realistic examination of the GAP, grades were randomly removed from Table 1 to create the data set in Table 5. Application of the GAP yielded several adjusted modules (Table 6). Before the one-way repeated-measures ANOVA could be applied, however, the missing data were replaced such that the Error Sum-of-Squares was minimised. The filled-in grades corresponding to Table 5 and Table 6 are presented in Table 7 and Table 8, respectively. This step was taken to allow a more realistic test of the GAP than simply filling in the missing grades with the original data. Following Kirk [20], the value of F_{crit} was modified using the correction of Box [21, 22]. Keppel [23]

Table 9. One-way repeated-measures ANOVA results for grades in Table 7 ($\alpha = 0.1$ significance level)



Source of Variation	*SS*	*df*	*MS*	*F*		F_{crit}
IV	3,405	11	310	2.8		1.6
Error	19,109	171	112		Adjusted F_{crit}	2.2
Across Subjects	45,402	19			($\hat{q}^1 = 0.34$)	
Total	67,916	201				

^{1}Adjustment factor used to obtain the adjusted F_{crit}

recommends using this correction where there is any doubt regarding the underlying assumptions for the one-way repeated-measures ANOVA. Because the repeated-measures ANOVA relies on in-filling of the data, it was concluded that the F_{crit} modification was necessary.

In Table 9 we present the ANOVA analysis of the data in Table 7. The calculated F statistic (2.8) is well above the adjusted F_{crit} (2.2), in which case the hypothesis of equal module means is rejected. This conclusion was also reached by the GAP. Next, the ANOVA analysis was repeated on the GAP-adjusted data presented in Table 8, with results given in Table 10. In this case, the F statistic (1.8) is less than the adjusted F_{crit}, and so we do not reject the hypothesis of equality of module means. Again, this is the outcome that was desired as a result of applying the GAP.

In practical terms, we have found that the variance adjustment in the GAP is often necessary, unlike the synthetic cases examined here. Within EXCEL, the variance adjustment can be set up and solved as an optimisation problem (using Solver in EXCEL). We have found it convenient to use a penalty function approach [24] to ensure minimal grade adjustment while satisfying (2) for each module. On the other hand, EXCEL's Solver is less useful for satisfying the $\hat{T}$ statistic test, as this test is not in the form of a continuous function. However, 'manual' adjustments as described above can be carried out very rapidly.

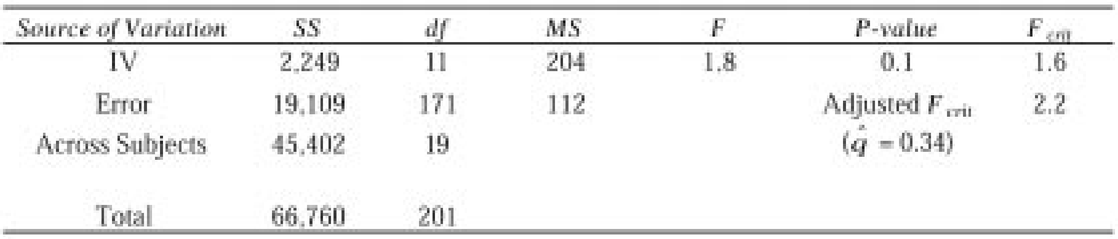
We now turn to discussing overall features of the GAP:

- The GAP process is designed to bring module results to within a pre-determined range. This range is controlled by the significance level used, with a smaller significance leading to a broader allowable range.
- During the GAP iterations, it can occur that modules are identified for moderation that were not identified previously. This merely indicates that that module is near the limits of the allowable range.
- We have suggested that, at completion of the GAP, the module offsets are adjusted back towards zero in order to check sensitivity.
- If, after adjustment, the examination board decides that overall results are too high or too low, albeit within the allowable range, then all results can be adjusted up or down simply by offsetting each grade uniformly.
- The variance adjustment procedure is very simple, but may not be applicable to modules with strongly skewed distributions. For such distributions, there are grounds for applying the 'scaling and offset' adjustment shown in Fig. 3.
- Properly applied, this type of adjustment will reduce the skew of the module grade distribution, while simultaneously increasing or decreasing the overall variance of the module grades. It should be remembered, however, that in using this type of adjustment scheme, the effect on the module average should be ignored; rather the focus should be wholly on the module variance. The module average will be accounted for subsequently in Step 8 of the GAP.

CONCLUSIONS

Our aim was to present and evaluate a simple, easily applied grade-adjustment procedure (GAP) to help analyze and moderate grades in engineering degree programs. The GAP is guided by the desire to permit different performances and academic assessments to stand, i.e., we recognise that it is not desirable to simply scale results so that a predefined distribution is obtained for all modules. Rather, while recognising that differences between

Table 10. One-way repeated-measures ANOVA results for grades in Table 8 ($\alpha = 0.1$ significance level)



Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	F_{crit}
IV	2,249	11	204	1.8	0.1	1.6
Error	19,109	171	112		Adjusted F_{crit}	2.2
Across Subjects	45,402	19			($\hat{q} = 0.34$)	
Total	66,760	201				

modules are expected, we wanted to constrain variability between results of different modules. This approach permits different performances and academic assessments to stand, while maintaining equity and fairness across all module outcomes.

Clearly, the approach taken in the GAP will produce sets of module results that are internally consistent. The question of normative scaling then naturally arises in the following form: How do we accommodate external norms that should be applied to the results of a given cohort? Again, we reiterate that scaling all grades to a single normed distribution would assume that all teaching is identical, student opportunities and circumstances are not markedly different and that teaching quality is invariant with time. The system described here aims to allow all these (and other) variables to operate while identifying and adjusting results modules that appear statistically to be outliers relative to overall performance of the cohort. Thus, scaling to a norm is possible simply by using the norm variance (actually, standard deviation) in the test of variances, see (2). If the norm distribution were symmetric, then application of the $\hat{T}$ statistic in the GAP would proceed as given. Otherwise, it would be adjusted (specifically, either A or B) to account for the asymmetry of the norm.

The GAP has been shown to be consistent with a more sophisticated ANOVA approach in the detailed analysis of an artificial data set. While we recognise that the data set used is possibly more 'well behaved' than real student outcomes, the artificial set was used as it was mildly non-normal, and could be reasonably tested in an ANOVA for comparison with the GAP. We have used the GAP and variants of it over the past few years and find that it produces outcomes are acceptable to our academic colleagues in that the moderated grades obtained are agreed as representing fair and justifiable outcomes for students.

REFERENCES

1. J. Penny and C. Grover, An analysis of student grade expectations and marker consistency, *Assess. & Eval. Higher Educat.*, **21**, 1996, pp. 173–183.
2. K. L. Chan, Statistical analysis of final year project marks in the computer engineering undergraduate program, *IEEE Trans. Educat.*, **44**, 2001, pp. 258–261.
3. G. Allen, Risk and uncertainty in assessment: exploring the contribution of economics to identifying and analysing the social dynamic in grading, *Assess. & Eval. Higher Educat.*, **23**, 1998, pp. 241–258.
4. V. N. Tariq, L. A. J. Stefani, A. C. Butcher and D. J. A Heylings, Developing a new approach to the assessment of project work, *Assess. & Eval. Higher Educat.*, **23**, 1998, pp. 221–238.
5. S. E. Gilliatt, and N. F. Hayward, A testing time: The role of subjective practices in making sense of student performance, *Assess. & Eval. Higher Educat.*, **21**, 1996, pp. 161–171.
6. R. V. J. Alberts, Equating exams as a prerequisite for maintaining standards: experience with Dutch centralised secondary examinations, *Assess. Educat.*, **8**, 2001. pp. 353–367.
7. N. S. Petersen, M. J. Kolen and H. D. Hoover, Scaling, norming, and equating, in *Educational Measurement, 3rd edition*, R. L. Linn (editor), American Council on Education, Series on Higher Education, Oryx Press, Phoenix, Arizona (1993) pp. 221–262.
8. M. W. Wang, and J. C. Stanley, Differential weighting: A review of methods and empirical studies. *Rev. Educat. Res.*, **40**, 1970, pp. 663–705.
9. T. Kubiszyn and G. Borich, *Educational Testing and Measurement*, John Wiley and Sons, Inc. New York (1999).
10. P. A. Games, H. J. Keselman and J. C. Rogan, A review of simultaneous pairwise multiple comparisons, *Statistica Neerlandica*, **37**, 1983, pp. 53–58.
11. D. L. Nuttall, J. K. Backhouse and A. S. Willmott, *Comparability of Standards Between Subjects*. Evans/Methuen, London (1974).
12. J. Jaccard and M. A. Becker, *Statistics for the Behavioural Sciences*, Brooks/Cole Publ. Co., Pacific Grove, Calif. (1997).
13. L. Sachs, *Applied Statistics: A Handbook of Techniques, 2nd edition* (Z. Reynarowych, translator). Springer-Verlag, New York (1984).
14. D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, CRC Press, Boca Raton, Florida (1997).
15. P. A. Games, H. B. Winkler, and D. A. Probert, Robust tests for homogeneity of variance, *Educat. Psychol. Msmt.*, **32**, 1972, pp. 887–909.
16. W. J. Dixon, and A. M. Mood, The statistical sign test. *J. Amer. Statist. Assoc.*, **41**, 1946, pp. 557–566.
17. W. E. Duckworth and J. K. Wyatt, Rapid statistical techniques for operations research workers. *Oper. Res. Quart.*, **9**, 1958, pp. 218–233.
18. P. Bridges, A. Cooper, P. Evanson, C. Haines, D. Jenkins, D. Scurry, H. Woolf and M. Yorke, Coursework marks high, examination marks low: Discuss, *Assess. & Eval. Higher Educat.*, **27**, 2002, pp. 35–48.
19. H. A. Thomas and M. B. Fiering, Mathematical synthesis of streamflow sequences for the analysis of river basins by simulation, in *Design of Water-Resource Systems*, A. Maass, M. M. Hufschmidt, R. Dorfman, H. A. Thomas, S. A. Marglin and G. M. Fair (eds) Harvard University Press. Cambridge, Massachusetts (1962).
20. R. E. Kirk, *Experimental Design: Procedures for the Behavioral Sciences, 2nd Edition*, Brooks/Cole Publ. Co., Belmont, Calif. (1982).

21. G. E. P. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems, I: Effect of inequality of variance in the one-way classification, *Ann. Math. Stat.*, **25**, 1954, pp. 290–302.
22. G. E. P. Box, Some theorems on quadratic forms applied in the study of analysis of variance problems, II: Effect of inequality of variance and correlation between errors in the two-way classification, *Ann. Math. Stat.*, **25**, 1954, pp. 484–498.
23. G. Keppel, *Design and Analysis: A Researcher's Handbook*, *3rd Edition*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey (1991).
24. K. Bajracharya and D. A. Barry, MCMFIT: Efficient optimal fitting of a generalised nonlinear advection-dispersion model to experimental data, *Comput. Geosci.*, **21**, 1995, pp. 61–76.

D. A. Barry is Professor of Environmental Engineering in the School of Engineering and Electronics at the University of Edinburgh. His work spans the discipline of engineering hydrology; he has published extensively in the international literature on biogeochemically reactive contaminant transport, stochastic transport equations, effect of heterogeneity on model predictions, computational schemes for reactive transport modelling, groundwater and vadose zone flow modelling, centrifugal scaling of flow and transport, multiphase flow and constitutive relationships, density-dependent flow, distributed catchment modelling and behaviour of coastal aquifers.

D.-S. Jeng is Senior Lecturer in Water Resources Engineering in the School of Engineering at Griffith University with research interests in hydraulics, coastal engineering and geotechnics. He has published extensively in the international literature on fluid-soil-structure interaction, ocean wave theories and groundwater hydraulics.

R. B. Wardlaw is a Senior Lecturer in Hydraulics, Hydrology and Water Resources Engineering at the University of Edinburgh, with research interests in water resources planning and management. He is a Fellow of the Institution of Civil Engineers and worked for 15 years with a firm of consulting engineers, prior to taking up his appointment at Edinburgh University in 1993, where he currently serves as Head of Civil and Environmental Engineering.

M. Crapper is a chartered engineer with professional interests in civil engineering and water and environmental management. He started his career as a consulting engineer working on a wide range of infrastructure projects, and developed interests in computer applications for hydraulic engineering and coastal sediment transport. In 1994 he joined the University of Edinburgh as a lecturer in 1994 and has since acquired a wide experience in various aspects of course administration and examination procedures. His research areas include hydraulics, sustainability and engineering education.

S. D. Smith is Lecturer in Project and Construction Management in the School of Engineering and Electronics at the University of Edinburgh. With a background as a construction engineer for a large multi-national civil engineering contractor, his research interests fall in to two main areas. First, he has undertaken extensive investigation of cyclic construction processes—effectively modelling the random nature of construction resources. Second, and more recently, he has undertaken research into construction health and safety, particularly hazard identification, assessment and management. He is the author of more than 20 refereed international journal and conference papers.x