

# Calibration of teachers' scores

Bruce Brown & Anthony Kuk  
Department of Statistics & Applied Probability

## 1. Introduction.

In the ranking of the teaching effectiveness of staff members through their student feedback scores, there can be a problem if it is perceived to be more difficult to obtain high scores in some courses than others. The result would be, therefore, that staff taking 'difficult' courses are disadvantaged. For example, it is often chimed that GEM and courses at the 1000 level are more difficult to score well in student feedback than courses at higher levels. Another criterion which is sometimes suggested as a classifier for 'difficulty' of courses is the class size, it being easier to obtain good student scores for smaller classes.

The problem of finding a suitable 'adjustment' to feedback scores, in order to correct for the difficulty of the courses taught; is not as easy as one might think. On the surface, based on Faculty of Science data for AY2003/04, there is not a whole lot of difference between the average feedback score of 3.934 for 1000 level modules and the average scores for other levels which range from 3.901 to 4.144. However, this direct comparison is valid only if the allocation of lecturers to modules is done at random which is certainly not the case. To see why a direct comparison is not valid, consider the extreme case where all the "best" teachers are assigned to teach 1000 level modules and the "worst" teachers to teach another level, resulting in an inflated score for 1000 level modules and a deflated score for the other level which lead to no significant difference between the two (this is similar to a situation in clinical trials where the stronger persons are assigned to the control group and the frail persons to the treatment group and as a result treatment effect cannot be established). This situation is not as far-fetched as one would think because there maybe a tendency, perhaps out of strategic considerations, for individual departments to assign the "better" teachers to teach 1000 level modules. If this is the case, and there is some evidence of this from Table 1, then the naïve calibration method of subtracting the level-average from the individual teacher scores would be misleading. Because of the selection effect, the 1000 level average would be artificially inflated relative to the scenario of random allocation of lecturers and subtracting this inflated average (over-subtraction) will not do justice to those lecturers teaching 1000 level modules. It is as a safeguard against selection bias like this that makes randomized allocation the gold standard in clinical trials. Unfortunately, randomization is not an option in the assignment of teachers to modules, and so we have to think of ways of calibrating feedback scores that remain valid under non-random assignment.

It is possible to calculate suitable corrections in an equitable manner, using some simple principles of statistical experimental design and analysis. The purpose of the present document is to outline how this can be done, using a 'matched pairs' design. Correction formulas are derived for stratification in terms of course level, but can be applied equally

well, with obvious modifications, to any other desired stratification, such as class size. To be concrete, we focus on the student response to one particular question in the teacher evaluation form, namely, the question on the overall effectiveness of the lecturer. The same procedure could be applied to the other questions as well. The focus of this document is on the calibration of student feedback scores; student comments and peer reviews are other important indicators that should be considered to give a more balanced view.

Expressed in mathematical terms, when staff member  $j$  teaches a course at level  $i$ , an average student assessment score  $\bar{y}_{ij}$  is recorded. The dot notation signifies averaging over the student responses in the class. A basic premise is that the score  $\bar{y}_{ij}$  depends in an additive way on two factors: an intrinsic, unobservable teaching ability score  $\mathbf{m}_j$  for the  $j$ th staff member, and another factor  $\mathbf{a}_i$  associated with the level of the module taught, so that

$$\bar{y}_{ij} = \mathbf{m}_j + \mathbf{a}_i + \text{error} . \quad (1)$$

The error term will be zero-mean, with an approximate normal distribution due to the central limit effects of averaging, and a variance which is proportional to the reciprocal of the class size. The goal is to estimate the  $\{\mathbf{m}_j\}$  terms for individual staff members, and to create a ranking based on these estimated values, but in order to do this it is necessary to estimate and hence eliminate the level effects  $\{\mathbf{a}_i\}$  attributable to teaching courses at different levels. Using the hat notation to denote a statistical estimate, if estimates  $\{\hat{\mathbf{a}}_i\}$  were available, then the teacher effectiveness scores  $\{\mathbf{m}_j\}$  could be estimated by  $\hat{\mathbf{m}}_j = \bar{y}_{ij} - \hat{\mathbf{a}}_i$ , and an overall estimate of  $\mathbf{m}_j$  would be the average of these  $\hat{\mathbf{m}}_j$  terms over the various courses taught by the staff member  $j$  during an academic year. But at first sight there is no obvious way to estimate the level effects  $\{\mathbf{a}_i\}$ , because of non-random allocation of modules to lecturers.

## 2. A matched-pairs design.

However, an approach which provides an effective way to estimate the level effects is made possible by virtue of the fact that there are many teachers who have taught modules at two or more different levels during the same academic year. If staff member  $j$  teaches courses at levels  $i_1, i_2$  in semesters 1, 2 respectively, then the average student assessments are modelled as

$$\bar{y}_{i_1 j} = \mathbf{m}_j + \mathbf{a}_{i_1} + \text{error} ,$$

$$\bar{y}_{i_2 j} = \mathbf{m}_j + \mathbf{a}_{i_2} + \text{error} ,$$

and the difference is

$$d_j = \bar{y}_{i_1j} - \bar{y}_{i_2j} = \mathbf{a}_{i_1} - \mathbf{a}_{i_2} + error. \quad (2)$$

Note how the lecturer effect  $\mathbf{m}_j$  is cancelled, regardless of how the teaching allocation was done, because we are taking the difference of two scores obtained by the same lecturer, so that what is left is an unbiased estimate of the level difference. We can set  $\mathbf{a}_1 = 0$ , making level 1 modules the baseline for comparisons; this will not affect the estimates of differences between the various  $\{\mathbf{a}_i\}$ . The estimate of  $\mathbf{a}_i$  for  $i > 1$  becomes an adjustment to be subtracted from each feedback average score for a level  $i$  module. Because there will be many staff members providing an observed difference  $d_j$  in any academic year, all the  $\{\mathbf{a}_i\}$  terms are estimable. Intuitively, a kind of overall repeated averaging, based upon the observe differences  $\{d_j\}$ , will provide the estimates  $\{\hat{\mathbf{a}}_i\}$ . In practice, the estimates are obtained by fitting the linear additive model (1) to the student feedback data using the method of unweighted or weighted least-squares. To obtain the calculated estimates, the data are coded in a systematic way and entered into any standard statistical computer package for analysis.

### 3. Example: analysis of Faculty of Science feedback data

The method proposed is illustrated by analysis of the Faculty of Science feedback data for AY2003/04.

First, the data was organized by deleting entries for teachers who did not teach at different levels in AY2003/04. Then the frequency of courses at different ‘levels’ was examined. There were only two cases at level 6000 and one at 8000, so these were combined with the level 5000 cases. There were still some small numbers, so the three USP groups were combined, as were the two GEM groups. After these combining steps, some teachers were teaching modules all of the one ‘level’ (for example, all USP modules, or all GEM modules), so their entries had to be deleted as well. After all deletions and grouping, there was a total of seven ‘levels’, from 1000 to 5000, plus USP and GEM, with frequencies below. Note that the omission of those lecturers who did not teach modules at different levels is for expository purpose only to highlight the fact that the level effects are estimated from the matched pairs data. In actual fact, the same estimates of level effects (relative to a baseline) would be obtained if we keep all the data in fitting the additive model (1) because the scores from lecturers who have taught at only one level will only contribute to the estimation of individual lecturer effects, but not the difference of level effects.

#### Tally for Discrete Variables: MODULE LEVEL (n = 482)

Module level	1000	2000	3000	4000	5000	GEM	USP
count	64	71	108	114	79	29	17

(i) **Unweighted analysis.** The first analysis to be discussed treats all the observed differences  $\{d_j\}$  as being of equal accuracy, i.e., the fact that feedback averages based on large class sizes are more accurate, is ignored. The matched pairs model, to eliminate all but the module effect, was then analysed. Residual plots were made for cautionary diagnostic purposes. The first plot, in Fig 1, is of residuals versus fitted values, reveals no severe deficiency and shows no apparent large outliers, though there is an apparent slight reduction of variability for higher scores, which is consistent with the compression of scores near to the upper bound of 5 units, and may suggest the application of a transformation to remedy this effect, though it is unlikely that such a transformation would lead to drastic changes in the conclusions.

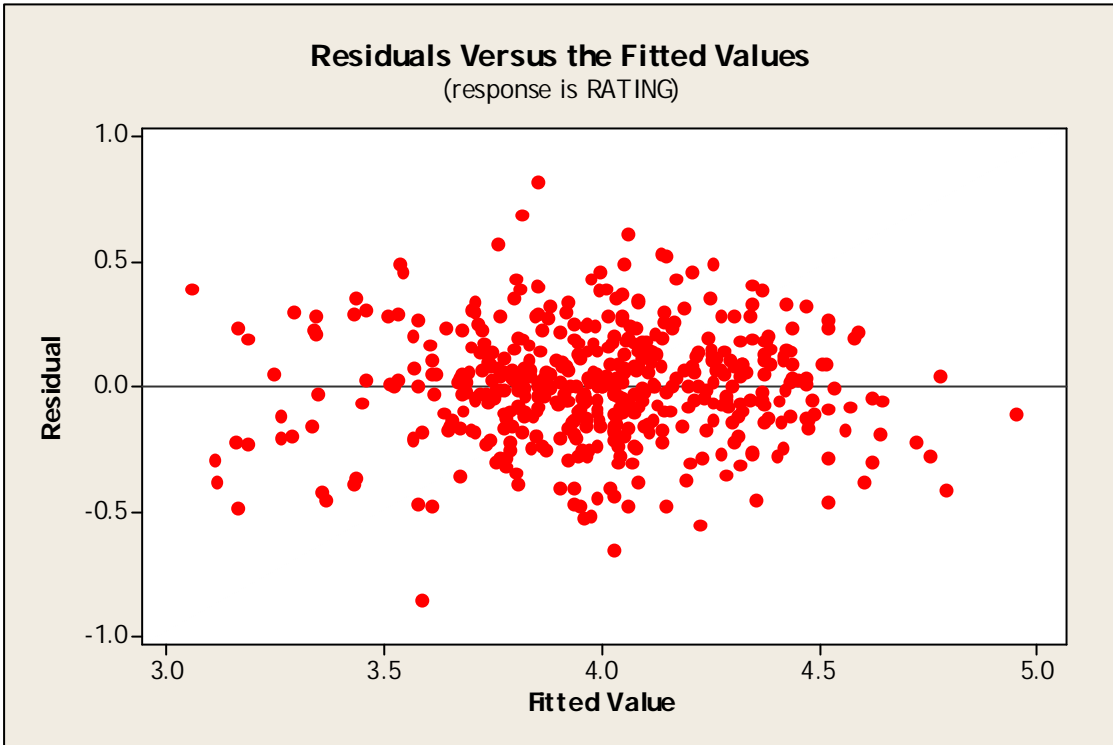
The second plot, in Fig 2, is a normal probability plot of the residuals, and is close to a straight line, providing a general confirmation of normality. There are a small number of residuals which are larger than expected, but not large or frequent enough to influence the results unduly.

An ANOVA shows that there are highly significant differences attributable to the variable *module level*. The fit indicated by R-squared is good.

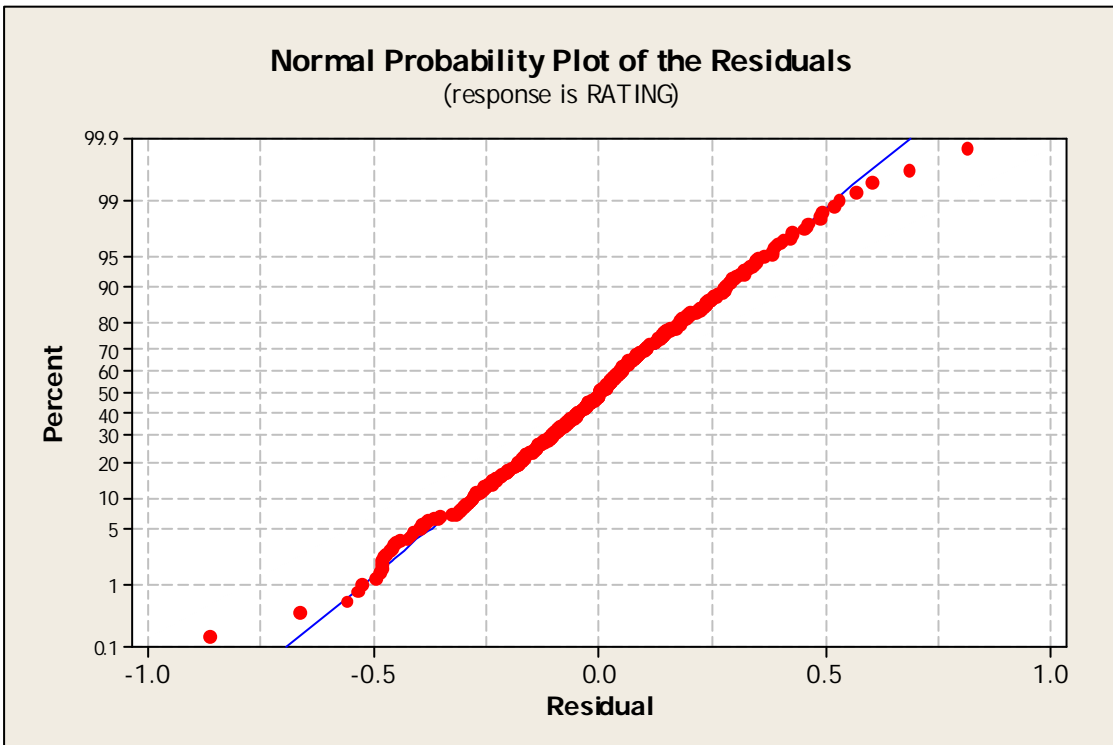
**Analysis of Variance** for RATING, using Adjusted SS for Tests

Source	DF	Seq SS	Adj SS	Adj MS	F	P
LECTURER	163	39.82645	41.41942	0.25411	3.31	0.000
MODULE LEVEL	6	6.01376	6.01376	1.00229	13.04	0.000
Error	312	23.97701	23.97701	0.07685		
Total	481	69.81722				

S = 0.277217    R-Sq = 65.66%    R-Sq(adj) = 47.06%



**Fig 1:** residuals plotted against fitted values.



**Fig 2:** normal probability plot of residuals.

**Estimated module level effects, standard errors, and adjustments**

Module level	Raw average	Estimated effect	Estimated adjustment	Standard error
1000	3.934	3.767	0	0
2000	3.901	3.858	-0.092	0.063
3000	3.947	3.979	-0.212	0.055
4000	4.025	4.039	-0.272	0.054
5000	4.144	4.214	-0.447	0.058
GEM	4.143	3.847	-0.080	0.073
USP	3.942	4.141	-0.375	0.104

**Table 1: fitted module level effects in the unweighted model.**

The estimated effects due to the different module levels are listed in Table 1, above. Several of the estimated adjustments for different module levels are significant, or nearly so. An analysis including weighting proportional to class sizes would better utilize the available information, and could be expected to yield more strongly significant results.

There is, apparently, a definite trend whereby it is more difficult to obtain high student feedback scores for lower level or GEM modules. It is easier to get higher scores where the students are at a higher level, or are more motivated, as in USP modules. There is a general trend for higher level modules to lead to higher feedback scores,

Note that the raw averages of scores for GEM and level 1000 modules are substantially higher than module level effects estimated unbiasedly using the additive model (1). This lends support to the theory that there is a tendency to allocate the “better” teachers more to teach 1000 and GEM modules than to other levels. Since the feedback scores for GEM and 1000 level modules are artificially inflated, the naïve method of adjustment based on raw level averages will under-adjust for this group of modules. The method that we proposed is able to circumvent this complication of non-random allocation of teachers, by creating matched pairs of teacher scores which differ only in module level. An alternative explanation is that model (1) allows arbitrary lecturer effects explicitly and so is general enough to cover the case of non-random assignment of teaching duties.

**(ii) Weighted analysis.** In least-squares statistical analyses, the optimal form of weighting is for weights to be chosen proportional to the reciprocal of observational variance. In the present analysis the variance of the average feedback scores for individual modules should be proportional to the reciprocal of class size if students respond independently, so the correct weights are the class sizes themselves.

Carrying out a weighted version of the analysis just described, and applying it to all the Science faculty teachers in AY 2003/2004, i.e., not just those involved in the matched-pairs part of the design, gives results described in the following output. Table 2 is an ANOVA which tests for the significance of the effects of different module levels.

**Tests of between-module effects(b)**

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
lecturer	6419.663	267	24.044	9.781	.000
module	186.351	6	31.058	12.635	.000
Error	882.475	359	2.458		
Corrected Total	7471.479	632			

a R Squared = .882 (Adjusted R Squared = .792)

b Weighted Least Squares Regression - Weighted by class size

**Table 2: ANOVA to test for the effects of different module levels.**

The adjustments, with the exception of that for USP, now have smaller standard errors, and are all significant. The estimated pattern of adjustments rises smoothly as level increases, with the exception of level 5000 modules, whose adjustment term is high, at -0.433. The difficulty of scoring highly in GEM modules is roughly the same as for level 2000 modules, and in the same respect, USP modules are roughly equivalent to level 5000 modules. This may reflect a high level of interest and motivation among students in the USP programme.

**Parameter Estimates(b)**

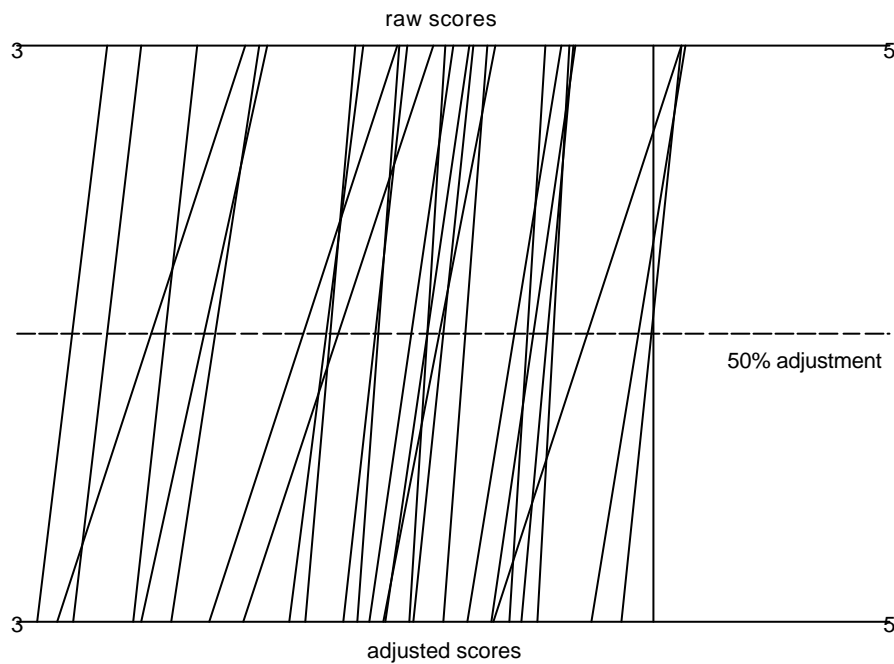
Module level	Estimated effect	Estimated adjustment	Standard error
1000	3.787	0.0	0.0
2000	3.870	-0.083	0.044
3000	3.934	-0.147	0.036
4000	3.980	-0.193	0.045
5000	4.220	-0.433	0.055
GEM	3.864	-0.077	0.038
USP	4.154	-0.367	0.115

b Weighted Least Squares Regression - Weighted by size

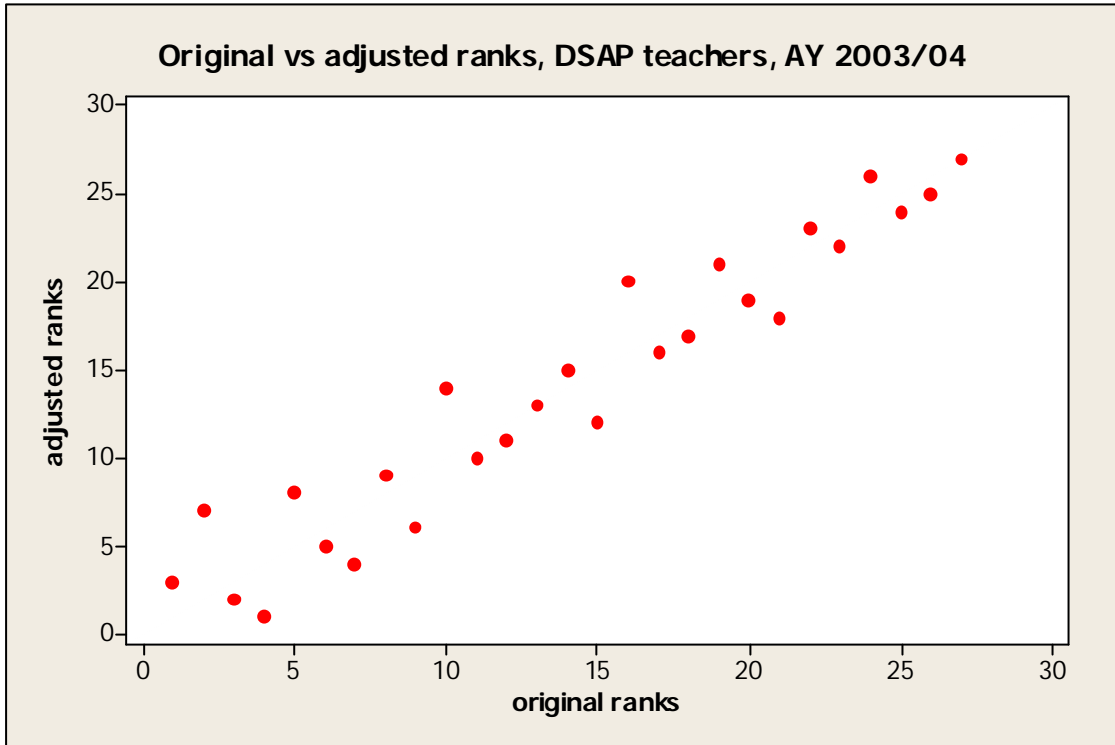
**Table 3: fitted module level effects in the weighted analysis.**

**Rankings within DSAP.** Finally, the effects upon within-Department rankings of teachers, due to adjustment of scores using the estimated module level effects, is shown in Figures 3 and 4 below, for the Department of Statistics & Applied Probability, in AY 2003/04. Note that rank = 1 is best, rank = 27 is worst. The ranking before and after adjustment are broadly similar, while displaying some distinct changes. The personnel in the top one-third (9 out of 27) remain unchanged by both rankings. These nine teachers

form two clusters. The first cluster consists of the top 4 teachers according to raw scores. Their average raw scores range from 4.458 to 4.531, a very narrow range of 0.073. After adjustment, the teacher who was ranked first dropped down to third position, because one of the modules that he taught was a 5000 level module, which resulted in a downward adjustment. The score of the teacher originally ranked second was adjusted even more because the only module that he had taught was a 5000 level module. The teacher originally ranked fourth was ranked first after adjustment because he taught only 1000 level module. While this may sound drastic, one should keep in mind that their raw scores only differ by 0.073 and so the levels of the modules taught should and are expected to have a bearing on the final rankings. The second cluster of teachers has average raw scores ranging from 4.211 to 4.277, a difference of 0.066. Again, because of this narrow range, there are interchanges of positions within this group of teachers after adjustments that took into account the levels of the modules taught.



**Fig 3: Original and adjusted scores for teaching staff in DSAP, AY 2003/2004.**



**Fig 4: Original and adjusted ranks for teaching staff in DSAP, AY 2003/2004.**

There is no doubt that making the indicated adjustments, by subtracting the estimated confounding terms for module levels from individual average feedback scores, will substantially remedy any perceived difficulty of scoring highly on lower level courses. There is a small but noticeable effect upon the resulting rank-ordering of teachers, which is shown, for the present example from DSAP, in the graph above.

The proposed adjustment method could be a bit controversial because it seems to favour lecturers teaching lower level modules and penalize those who teach high level modules. It should be stressed, however, that the required adjustments are estimated from the differences in scores received by the same lecturer and so should be credible. By taking the difference, the effects of lecturers are removed and the remaining systematic effects can be reasonably attributed to levels of modules alone. We believe that the changed rankings are made upon an equitable basis. Certainly, calibration by subtracting the level-specific average without making allowance for selection effect in the assignment of teaching duties will result in under-adjustment, whereas the adjustment proposed in this working paper appears to be somewhat drastic, particularly when it is applied to feedback scores for 5000 level modules. The truth may lie somewhere in between. One possibility is to average the raw score with the adjusted score, or equivalently to adjust by 50%, and the resulting scores can be read out from the broken line in Figure 3. Moving this line up and down corresponds to less than and more than 50% adjustment.

#### **4. More sophisticated analyses**

We have tried two more sophisticated analyses. For the first one, we took notice that the individual student responses for the high scoring modules tend to exhibit less variability. We performed a new weighted analysis that factored this into account but the conclusions were not changed by much and so will not be reported here. Our second analysis was an attempt to remove the “compression” or “ceiling” effect at both ends of the 1 to 5 scale by applying the logistic transformation to the average scores, i.e., we fit model (1) to the transformed average scores. Again, we obtained more or less the same results which will not be repeated here. These lend support to the robustness of the simple analyses that we performed in section 3.