Watching the watchmen:

Ercolani, Marco; Ercolani, Joanne

DOI:
10.1016/j.iree.2014.05.001
10.1016/j.iree.2015.02.001

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Ercolani, MG & Ercolani, JS 2014, 'Watching the watchmen: a statistical analysis of mark consistency across taught modules', International Review of Economics Education, vol. 17, pp. 17–29. https://doi.org/10.1016/j.iree.2014.05.001, https://doi.org/10.1016/j.iree.2015.02.001

International Review of Economics Education 17 (2014) 17–29

Contents lists available at ScienceDirect

International Review of Economics Education

journal homepage: www.elsevier.com/locate/iree

Watching the watchmen: A statistical analysis of mark consistency across taught modules☆

Marco G. Ercolani *, Joanne S. Ercolani

Department of Economics, Birmingham Business School, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK

ARTICLE INFO

Article history:
Received 14 February 2013
Received in revised form 12 May 2014
Accepted 14 May 2014
Available online 2 June 2014

JEL classification:
A20
A23
C12
C18

Keywords:
Assessment standard
Module grade
Quality assurance

ABSTRACT

Verifying that taught modules are marked and taught to a common standard is important but doing so by comparing mean module marks is inadequate when students' ability is not uniform across these modules. For example, a module taken by a group of students of above average ability may justifiably result in a high mean mark, without implying that inconsistent standards have been applied. We propose a modified version of the *fixed effects regression* that provides direct estimates of module mark biases while conditioning for student composition and requiring no additional, potentially confidential, information on students or staff. We describe how this modified fixed effects regression can be implemented on a set of student marks and how the results can be interpreted. Increases in student numbers and tuition fees have increased the preoccupation with, and monitoring of, marks. We show how one can generate statistics that are more informative of the biases in marking, while being explicit about their limitations.

☆ We are grateful to two anonymous referees for their thoughtful comments and suggestions. We are also grateful to Professor Peter Davies and Professor Cillian Ryan for their support and encouragement. We thank an anonymous source for the data. All errors are our own.

* Corresponding author. Tel.: +44 0121 414 7701.
E-mail addresses: m.g.ercolani@bham.ac.uk (M.G. Ercolani), j.s.ercolani@bham.ac.uk (J.S. Ercolani).

http://dx.doi.org/10.1016/j.iree.2014.05.001

1. Introduction

Within academic departments, discussions sometimes arise over the range of students' marks awarded across taught modules (Orr, 2007; Price, 2005) and whether instructors are marking and teaching to a common, consistent[1] standard. These discussions are sometimes informed by informal comparisons of modules' mean marks, as illustrated in the second column of Table 1. Sometimes the comparisons are more formal, such as calculating how each module's mean mark differs from the overall average (the *cohort mean*), as illustrated in the third column of Table 1. However, upon identifying modules with large *mean mark biases* we are still unable to, without confidential or difficult to access data, separate out if these biases are due to:

- The average ability of each student in that module.
- The level to which students' assessment has been marked.
- The standard to which students have been taught.

In this paper we show how to use *fixed effects regressions* to remove those parts of the module biases that are due to the average ability of the students, the *student fixed effects*. Additional, simple calculations can then be used to obtain the *student-conditioned module biases*. We also propose a minor modification to the fixed effects regression which estimates these biases directly.

The resulting student-conditioned module biases provide more appropriate statistics upon which *outlier modules* can be identified. For example, a module with a very high *simple mean* may in fact have a very small student-conditioned module bias because it was taken by students of high average ability. Though the student-conditioned module biases do not identify whether inconsistent marking or teaching has taken place, these biases at least reflect instructor-related issues only. Instructors may therefore be best placed to use confidential students' and colleagues' evaluations to identify marking or teaching inconsistencies. Student-conditioned module biases may substantially mitigate the incentive for instructors to mark towards the overall mean in an attempt to satisfy the monitoring, although we cannot claim that this incentive is entirely eliminated. This gradual regression towards the mean[2] of the marks distribution can be an unintended consequence of comparing simple module means.

One advantage of these fixed effects regressions is that they control for each students' average performance without requiring additional information on instructors or students. For the fixed effects methods to improve upon a simple comparison of module means, each student's overall ability must have a significant impact on the marks they achieve in every module they take. For example, a student who is good in Mathematics is also likely, though not certain, to be good in History. We feel this is a reasonable assumption and, as we will see, statistical tests strongly support this in our dataset. In the event that the student fixed effects make no significant contribution to the marks they achieve, the student-conditioned module biases will offer no benefits over the comparison of simple module averages. Although it may seem an anathema to compare marks between modules in different topics, say Mathematics and History, module marks should nonetheless be comparable because they are often used to arrive at an aggregate degree mark for each student. Another requirement for the fixed effects estimators to work is that the weaker students share at least some modules with stronger students. This method of analysis should therefore work for any cohort of students who share a large proportion of modules. The students need not be on the same programme of study and they need not all share any one module.

The concept of an outlier module necessarily requires comparison to a benchmark grade, for which there are two obvious choices. One is simply a stated average target, of say 60%, set by departments or the University. The benchmark we prefer, and therefore adopt, is the overall average mark, the *cohort mean*. The cohort mean is a good benchmark because it can be viewed as the expected mark if all students were of equal ability, exerted equal effort and if teaching staff taught and marked to the same

[1] By *consistency* we mean a social agreement among things; rather than the statistical definition of *asymptotic consistency* in large data samples or the mathematical definition of *internal consistency* within a model.

[2] This refers to the phenomenon of *regression towards the mean* as first proposed by Galton (1886).

Table 1
Mean module marks.

Module (ranked by mean)	Module mean mark	Module mark bias[a]	Observations
1	49.34	−9.00	110
2	53.87	−4.47	38
3	53.89	−4.44	19
4	54.64	−3.69	14
5	54.75	−3.59	4
6	55.49	−2.84	126
7	56.71	−1.62	52
8	57.00	−1.34	11
9	57.42	−0.91	33
10	57.44	−0.90	82
11	58.21	−0.13	43
12	58.59	0.26	54
13	60.40	2.06	5
14	60.67	2.34	143
15	61.00	2.66	11
16	61.02	2.68	102
17	61.09	2.75	11
18	62.80	4.46	10
19	63.67	5.33	12
20	63.74	5.40	106
21	64.45	6.12	11
22	65.00	6.66	16
23	67.31	8.98	16
24	68.17	9.83	6
25	72.56	14.22	9
(Overall) cohort mean	58.336		1044 marks (awarded to 170 students)

[a] Module mean minus (overall) cohort mean.

standard. In Section 3 we show how to ensure that the student-conditioned module biases are relative to the cohort mean in the conventional fixed effects regressions and, in Section 4, how our modified version estimates these biases directly. Our modified fixed effects regression offers four specific advantages over the comparison of simple module averages:

- It controls for each individual student's overall performance.
- It automatically calculates module biases relative to the overall *cohort mean*.
- It identifies if the module biases are statistically different from zero.
- It accounts for the number of students in each module when calculating this statistical difference by giving more weight to modules with more students.

The final two features could also be achieved when comparing simple module means but they would require additional statistical computations.

The dataset that we use in our empirical example is for a cohort of 170 Master's level students on various degree programmes within one academic department in one academic year. Most students took seven modules and they shared a large proportion of these modules. Table 1 lists the module mean marks for the example dataset, with modules ranked by mean mark to facilitate comparisons. As we will see in Section 4, some modules that appear to be outliers in Table 1 are not outliers when student effects are controlled for (and *vice versa*). To preserve the anonymity of the source data, no programme or module names are given.

The paper proceeds as follows. Section 2 provides further motivation for analysing the consistency of module marks. In Section 3 fixed effects regression is described and initial regression results are provided. Section 4 presents our simple modification of the fixed effects estimator to obtain the module biases directly and discusses the results for our example dataset. Section 5 deals with possible statistical refinements and analytical extensions. A final section concludes.

2. Background

In UK higher education, increases in student numbers and fees have created a need for formal procedures to help reach agreement on the range of marks awarded. In the 1970s <20% of UK school-leavers went into higher education and, with small student numbers, instructors could use informal means to reach agreement on standards (Thomas, 1976). In 1997 the Labour Government adopted a policy objective of raising higher education attendance to 50% of school-leavers. By 2010 that objective was not quite achieved but numbers increased substantially by attracting students from outside the UK. With such large numbers in higher education, informal agreement on marking standards has become harder and formal procedures more necessary. In late 2010 the Coalition Government passed legislation allowing substantial increases in higher education tuition fees. In England these fees in many cases tripled to £9000 per year. These higher fees may increase the students' demand for accountability in the marking process and hence may increase the need for formal statistical verification. Higher fees may also create more pressure from students to identify where the quality of teaching is significantly above or below the average standard in that academic institution. Despite these pressures, Blackmore (2009) suggests that instructors are shifting effort away from improving teaching and towards monitoring assessment. These challenges are also faced by higher education institutions outside the UK where student numbers or student fees can be very high.

Achieving consistent marks across different modules is of vital importance from a didactic perspective because the wish is for students to choose their optional (elective) modules based on their interests and career plans. Inconsistent marks create distorted incentives for students to select modules where achieving a higher mark is perceived to be easier. Consistent marks are also important because students' perceptions of being assessed on the basis of merit influences their learning experience, see Lizzio et al. (2002) and Nesbit and Burton (2006). Students' increased preoccupation with their marks is an inevitable consequence of increased fees and a competitive graduate job market.

In related research, the incentive for instructors to award high marks in tacit return for high student evaluations of teaching (SET) has been documented by Isley and Singh (2005), McPherson (2006), Nowell (2007) and others. Though related, the analysis in this paper is neither a substitute nor a complement for the analysis of SET. The latter is covered by a large and growing literature. For a recent empirical application of SET see Nowell et al. (2010). The potential exists for merging the analysis of SET with our analysis of module bias to formally identify how the two interact. However, if a causal relationship between the two is sought, the timing of the assessments and the SET should be carefully modelled.

As discussed in Section 1, our proposed fixed effects analysis controls for the student effects when comparing modules but it does not mitigate all of the undesirable incentives facing instructors who are scrutinised using simple module means. The use of almost any statistical procedure to monitor marks incentivises instructors to award marks that are very close to the expected cohort mean so that their module mean does not stand out. Comparison of simple module mean marks can therefore lead to a strong compression of the marks distribution. Instructors' preoccupation with marking (Hadsell and MacDermott, 2012) is already on the increase and qualitative monitoring of instructors in higher education by peer observation of teaching is now widespread and its consequences have been documented in Hammersley-Fletcher and Orsmond (2004). The proposed fixed effects regression can therefore be seen as a simple but effective refinement to existing assessments of marking and teaching. Even for instructors who simply wish to understand what standard is being applied when awarding marks, the student-conditioned biases offer a metric that is more informative than the simple use of module means. This may include instructors who are new to the teaching profession, see Nicholls (2005) and Webber (2005), or instructors who have recently arrived to a new institution with different marking practices, see Knight (2006).

3. Indirect estimation of student-conditioned module biases using standard fixed effects regressions

A fixed effects model can either be estimated using simple ordinary least squares (OLS) regression upon an equation such as (1), where the fixed effects are specified using binary variables, or using fixed

effects regression on an equation such as (4), where the data are transformed so that the fixed effects are *differenced out*. Both OLS and fixed effects estimation will provide identical parameter estimates for the remaining parameters in the model. In this section we show how to estimate models of module marks that include student fixed effects and how to derive the fixed effects estimator.

Fixed effects estimators are often available as ready-written routines within statistical software packages and only require a one-line command. The detailed description in this section is included to aid the understanding of how these fixed effects estimators are implemented in our example (see Greene, 2012, Section 11.4 for a generic discussion) and to provide a basis for understanding our modified fixed effects estimator in Section 4. Readers who are familiar with the fixed effects estimator can look at the results in Table 2 and then go straight to Section 4.

As explained in Section 1, the dataset under analysis is for a cohort of Master's level students that contains the following variables:

- Mark: This is the dependent variable in our analysis and it records the final percentage mark awarded to each student in each completed module.

Table 2
Fixed effects regressions on Marks awarded.

Estimation method:	OLS regression with student effects (Eq. (1))		Fixed effects regression (Eq. (4))		Student cond. module biases (Eqs. (5) and (6))
	Parameters	*s.e.*	Parameters	*s.e.*	
Constant (β_0)			49.97	0.82**	
Module_1					−8.36
Module_2	6.74	1.62**	6.74	1.62**	−1.62
Module_3	1.45	2.29	1.45	2.29	−6.91
Module_4	1.09	2.68	1.09	2.68	−7.27
Module_5	5.72	4.67	5.72	4.67	−2.65
Module_6	5.86	1.08**	5.86	1.08**	−2.51
Module_7	8.12	1.44**	8.12	1.44**	−0.25
Module_8	4.04	2.99	4.04	2.99	−4.33
Module_9	9.89	1.70**	9.89	1.70**	1.53
Module_10	7.61	1.23**	7.61	1.23**	−0.75
Module_11	5.49	1.77**	5.49	1.77**	−2.88
Module_12	9.39	1.39**	9.39	1.39**	1.03
Module_13	8.83	4.13*	8.83	4.13*	0.46
Module_14	11.5	1.07**	11.5	1.07**	3.14
Module_15	8.04	2.99**	8.04	2.99**	−0.33
Module_16	10.37	1.18**	10.37	1.18**	2.01
Module_17	6.83	3.00*	6.83	3.00*	−1.53
Module_18	7.99	3.10*	7.99	3.10*	−0.37
Module_19	8.56	3.00**	8.56	3.00**	0.20
Module_20	14.61	1.11**	14.61	1.11**	6.24
Module_21	11.51	2.91**	11.51	2.91**	3.15
Module_22	10.28	2.51**	10.28	2.51**	1.91
Module_23	19.75	2.44**	19.75	2.44**	11.39
Module_24	21.42	3.90**	21.42	3.90**	13.06
Module_25	14.94	3.15**	14.94	3.15**	6.58
Student effects	Included but not reported		Differenced out		
Overall-R^2	0.6336				*Cohort mean:*
Within-R^2			0.2554		$\overline{\overline{\text{Mark}}} = 58.336$
Observations	1044		1044		

Two-way ANOVA:	Effects	Partial sum of squares	*F-statistics*
(Eq. (7))	Model overall effects	96,953.2048	$F_{(193,850)} = 7.62^{**}$
	Student fixed effects	75,255.1511	$F_{(169,850)} = 6.75^{**}$
	Module fixed effects	19,230.0015	$F_{(24,850)} = 12.15^{**}$
	$R^2 = 0.6336$	Observations 1044	

* $p < 0.05$.
** $p < 0.01$.

- Module_1 to Module_*M*: These are zero-one module dummy variables identifying all of the $M=25$ modules taught by the Department to that cohort of students.
- Student_1 to Student_*N*: These are zero-one student dummy variables identifying all of the $N=170$ students in the cohort.

Each observation is uniquely identified by two indices:

- *id*: This is a student-specific identifier. Most academic institutions already assign anonymous numeric identifiers to aid in the impartial treatment of students.
- *md*: This identifies which module was taken by the student. The data panel is 'unbalanced' as students do not all take the same modules.

This is a panel dataset in a non-conventional sense insofar as the module identifier *md* replaces the usual time dimension *t*. In some statistical software packages one of the identifiers must be the time dimension and these packages can be persuaded into accepting the dataset as a panel by pretending that the *md* identifier is the time dimension.

OLS can be used to estimate Eq. (1) that includes student fixed effect dummies:

$$\text{Mark}_{id,md} = \sum_{i=2}^{M} \beta_i \text{Module\_}i_{id,md} + \sum_{j=1}^{N} \alpha_j \text{Student\_}j_{id,md} + \epsilon_{id,md} \tag{1}$$

where the α_j are the individual students' fixed effects, the β_i are the individual module effects and ε is the error term. To avoid perfect multi-collinearity, both the constant and any one of the Module_*i* dummies are excluded because the student identifiers sum to one. We have excluded the variable for Module_1, hence the first summation term starts at $i=2$. The first regression in Table 2 reports the estimates of Eq. (1). These estimates reflect how the student-conditioned module means deviate from Module 1's mean. However, as there is no constant to capture Module 1's level we cannot compute the module biases relative to the cohort mean (or any other reference level). We can see that the overall fit of the model is good with an R^2 of 0.6336 showing that over half of the variation in marks is explained by this simple model.

The standard fixed effects estimator, derived below and defined in Eq. (4), eliminates the need to estimate the numerous individual fixed effects and provides an estimate the constant that is missing from OLS estimates of Eq. (1). This makes it possible for us to calculate module mark biases with respect to any chosen reference level, such as the cohort mean. This is done using three transformations of Eq. (1). The first takes student-average values of the variables to give:

$$\bar{\text{Mark}}_{i\bar{d},md} = \sum_{i=2}^{M} \beta_i \text{Module\_}\bar{i}_{i\bar{d},md} + \sum_{j=1}^{N} \alpha_j \text{Student\_}j_{id,md} + \bar{\epsilon}_{i\bar{d},md} \tag{2}$$

where the over-lines indicate student-based averages for each variable. So, for example, for any one student, every entry for the variable $\bar{\text{Mark}}_{i\bar{d},md}$ is that student's mean mark across all of her/his own modules. Note that in Eq. (2) there are no over-lines on the student dummies because they are already equal to their own individual means ($\text{Student\_}j_{id,md} = \text{Student\_}\bar{j}_{i\bar{d},md}$).

The second transformation takes the overall cohort average value of Eq. (1):

$$\bar{\bar{\text{Mark}}}_{i\bar{d},m\bar{d}} = \sum_{i=2}^{M} \beta_i \bar{\bar{\text{Module}}}\_i_{i\bar{d},m\bar{d}} + \sum_{j=1}^{N} \alpha_j \bar{\bar{\text{Student}}}\_\bar{j}_{i\bar{d},m\bar{d}} + \bar{\bar{\epsilon}}_{i\bar{d},m\bar{d}}$$

This can be simplified further because all its elements are constants and can therefore be expressed without *id* and *md* subscripts, and because the sum of the student effects across all students and all modules is equal to a constant term denoted by β_0 and hence:

$$\bar{\bar{\text{Mark}}} = \sum_{i=2}^{M} \beta_i \bar{\bar{\text{Module}}}\_i + \beta_0 + \bar{\bar{\epsilon}} \tag{3}$$

The final transformation defines the fixed effects equation (4) by subtracting (2) and adding (3) to Eq. (1) where v is just a composite of the other error terms:

$$\text{Mark}_{id,md} - \bar{\text{Mark}}_{i\bar{d},md} + \bar{\bar{\text{Mark}}} = \beta_0 + \sum_{i=2}^{M} \beta_i (\text{Module\_}i_{id,md} - \text{Module\_}\bar{i}_{i\bar{d},md} + \bar{\text{Module}}\_\bar{i}) + v_{id,md} \quad (4)$$

The fixed effects estimator is obtained by applying OLS to Eq. (4). The paradox is that the fixed effects estimator does not actually estimate these fixed effects; it differences them out so that they do not need to be estimated, though they are controlled for. This is very convenient when there are a large number of individuals thus eliminating the need to estimate a large number of parameters.

The second regression in Table 2 reports the results from estimating Eq. (4). Notice that the same parameters as before are estimated but with the useful addition of an estimated constant β_0. Although the within-R^2 is only 0.2554 this is after the variation due to the student effects has been differenced out. This therefore indicates that this model is explaining about one quarter of the variation in marks after the variation explained by student fixed effects has been removed. An important aspect of the fixed effects estimators is that they properly account for the *degrees of freedom* in the regression by including the number of fixed effects that have been differenced out of the model, 170 in this example. This is why the standard errors in the two regressions in Table 2 are identical to one another even though 194 parameters have been estimated in Eq. (1) while only 25 have been estimated in Eq. (4).

With the estimated constant now at our disposal, we can calculate the implied student-conditioned module biases for these marks by using the overall cohort mean mark as a reference. The student-conditioned bias for the control module (Module 1) is simply the difference between the estimated constant and the cohort mean, Eq. (5), and the student-conditioned bias for each of the remaining modules is the estimated module parameter plus the bias for the control module, Eq. (6):

$$\text{Bias\_}1 = \hat{\beta}_0 - \bar{\bar{\text{Mark}}} \quad (5)$$

$$\text{Bias\_}i = \hat{\beta}_i + \hat{\beta}_0 - \bar{\bar{\text{Mark}}} \quad \text{for } i = 2 \text{ to } M \quad (6)$$

These module biases are reported in the final column of Table 2. From these we see that Module 1 has the largest negative student-conditioned module bias at −8.36 which is not very different from its simple module bias of −9.00 in Table 1. Module 25 has a positive student-conditioned module bias of 6.58 which, though large, is less than half that of its simple bias of 14.22 in Table 1. We could test whether Eqs. (5) and (6) are statistically different from 0 using simple t-tests. However this would require us to find additional information on the covariances between our estimated parameters. We prefer to specify a modified fixed effects estimator, as outlined in Section 4, that estimates these student-conditioned module biases directly and that requires just a simple degrees of freedom adjustment to the standard errors. Simple t-tests of significance on these parameter estimates, using the adjusted standard errors, will directly identify outlying modules.

A two-way analysis of variance (ANOVA), as specified by Eq. (7), provides yet another way of analysing this dataset:

$$\text{Mark}_{id,md} = \beta_0 + \sum_{i=2}^{M} \beta_i \text{Module\_}i_{id,md} + \sum_{j=2}^{N} \alpha_j \text{Student\_}j_{id,md} + \epsilon_{id,md} \quad (7)$$

A two-way ANOVA provides us with an indication as to the significance of the student effects and the module effects in predicting the marks achieved. A standard two-way ANOVA is defined by the OLS estimation of Eq. (7) and then reporting the significance of the contribution made by the modules and the students as illustrated at the base of Table 2. We see that the F-statistic for the overall model is 7.62 showing a strong overall significance for the model. The F-statistic for the student effects is 6.75 and that for the module effects is 12.15 showing that each is statistically significant, well beyond the 1% significance level, in predicting the marks obtained by students. Although Eqs. (1), (4) and (7) are statistically equivalent to one another in terms of overall fit, they provide slightly different results as

they use different reference points for the constant β_0.The equivalence in overall fit for Eqs. (1) and (7) is shown by their identical R^2 of 0.6336.

4. Direct estimation of student-conditioned module biases using a modified fixed effects regression

In this section we show how to modify Eq. (4) to define an alternative fixed effects regression that estimates the student-conditioned module biases directly, without recourse to other calculations. However, the resulting standard errors need to be corrected for the appropriate degrees of freedom.

Eq. (8) illustrates our modified fixed effects estimator. It presents two modifications upon Eq. (4). The first is that the constant β_0 has been excluded and the first module effect is therefore included alongside the others.[3] Applying only this first modification would estimate the student-conditioned module means. The second modification is *not* to add the overall cohort mean on the left-hand side variable. Applying both modifications directly estimates the student-conditioned module biases with respect to the cohort mean:

$$\text{Mark}_{id,md} - \bar{\text{Mark}}_{i\bar{d},md} = \sum_{i=1}^{M} \beta_i (\text{Module\_}i_{id,md} - \text{Module\_}\bar{i}_{i\bar{d},md} + \bar{\text{Module}}\_\bar{i}) + \nu_{id,md} \tag{8}$$

We are not aware of any software commands that will implement our modified fixed effects estimator directly. We therefore make the appropriate transformations to the data and apply OLS. The estimated student-conditioned module biases are reported in column 3 of Table 3 and we can see that, as expected, they are exactly the same as those calculated in the last column of Table 2.

The reported standard errors have been corrected for the appropriate degrees of freedom in the fixed effects estimator in Eq. (8). That is, the extra N student effects that have been differenced out. This is akin to the automatic adjustment made in the standard fixed effects estimator in Eq. (4). The corrected standard errors for the modified fixed effects estimator can be calculated by re-weighting the automatically reported erroneous OLS standard errors using the following formula:

$$s.e.(\hat{\beta}_i)_{\text{FE}} = \sqrt{\frac{Obs - M}{Obs - M - N + 1}} \times s.e.(\hat{\beta}_i)_{\text{OLS}} \quad \text{for } i = 1 \text{ to } M \tag{9}$$

This simply multiplies out the incorrect degrees of freedom and re-divides by the correct degrees of freedom. The fifth column reports the corrected t-statistics and, at the bottom of the table, the critical values needed to determine statistical significance.

To understand why the adjustment in Eq. (9) is appropriate, consider that the OLS covariance matrix is $\text{var}(\beta) = \hat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}$ where $\mathbf{X}$ is the matrix of explanatory variables. OLS estimation of Eq. (8) erroneously calculates $\hat{\sigma}^2_{\text{OLS}} = RSS/(Obs - M)$ where RSS is the residual sum of squared residuals and Obs is the total number of observations. However, under fixed effects estimation this should be $\hat{\sigma}^2_{\text{FE}} = RSS/(Obs - M - N + 1)$ so that the N student effects are also accounted for. The fact that the N student effects add up to the constant adds 1 back into the total. The correct estimated error variance can therefore be calculated as:

$$\hat{\sigma}^2_{\text{FE}} = \frac{Obs - M}{Obs - M - N + 1} \times \hat{\sigma}^2_{\text{OLS}} \tag{10}$$

The results in Table 3 show that the size, and ranking, of the module biases can be very different under the student-conditioned (fixed effects) analysis than under the simple module bias analysis.

Modules 1, 3, 4, 6 and 11 have statistically significant negative student-conditioned biases. Hence, these modules are outliers suggesting they have either been marked too harshly or that poor student learning experiences led to low assessment scores. Interestingly, outlier Module 11 has a simple bias of just −0.13 and would therefore not have been identified as an outlier if comparing simple module means. Module 2 has a small, insignificant student-conditioned bias of −1.62

[3] We are grateful to an anonymous referee for this suggestion.

Table 3
Student-conditioned module biases.

	Simple module biases (Table 1)	Student conditioned results				
		Module biases (Eq. (8))	Corrected statistics (Eq. (9))			
		β_i	*s.e.*	*t*-stats	Outliers	Obs.
Module_1	−9.00	−8.36	0.82**	−10.232	Largest *t*-stat	110
Module_2	−4.47	−1.62	1.44	−1.126		38
Module_3	−4.44	−6.91	2.09**	−3.309	Negative	19
Module_4	−3.69	−7.27	2.48**	−2.935	Negative	14
Module_5	−3.59	−2.65	4.56	−0.580		4
Module_6	−2.84	−2.51	0.75**	−3.325	Negative	126
Module_7	−1.62	−0.25	1.22	−0.204		52
Module_8	−1.34	−4.33	2.81	−1.541		11
Module_9	−0.91	1.53	1.53	1.001		33
Module_10	−0.90	−0.75	0.97	−0.777		82
Module_11	−0.13	−2.88	1.45*	−1.984	Negative	43
Module_12	0.26	1.03	1.18	0.874		54
Module_13	2.06	0.46	4.01	0.116		5
Module_14	2.34	3.14	0.69**	4.530	Positive	143
Module_15	2.66	−0.33	2.81	−0.117		11
Module_16	2.68	2.01	0.85**	2.364	Positive	102
Module_17	2.75	−1.53	2.82	−0.542		11
Module_18	4.46	−0.37	2.93	−0.128		10
Module_19	5.33	0.20	2.82	0.070		12
Module_20	5.40	6.24	0.83**	7.547	Positive	106
Module_21	6.12	3.15	2.73	1.153		11
Module_22	6.66	1.91	2.31	0.830		16
Module_23	8.98	11.39	2.26**	5.041	Positive	16
Module_24	9.83	13.06	3.77**	3.462	Largest bias	6
Module_25	14.22	6.58	2.99*	2.198	Positive	9
Within-R^2		0.2554				1044
			$t^{5\%}_{Obs-M-N+1} = \pm 1.963$			
			$t^{1\%}_{Obs-M-N+1} = \pm 2.582$			

* $p < 0.05$.
** $p < 0.01$.

despite a simple mean-based bias of −4.47. This implies Module 2 is not an outlier, with appropriate marks and students who on average also performed poorly in other modules. Under the system of comparing simple module means, the Module 2 marker may have come in for unfair criticism. Module 8 has a student-conditioned bias that, though large, is not statistically significant. This is because the weight of statistical evidence is a combination of the bias size and the sample size, and in this case the combined evidence is insufficient.

Modules 14, 16, 20, 23, 24 and 25 have statistically significant positive student-conditioned biases. These biases suggest the modules have either been marked too generously or that good student learning experiences led to high scores. Module 22 has a large simple module mean bias of 6.66 but the student-conditioned bias is small at 1.91 and statistically insignificant, suggesting the marks are consistent for this cohort. Module 22 may have been erroneously identified as one with generous marks when comparing simple means. Module 21 has a large student-conditioned bias of 3.15 but this is statistically insignificant because it was only taken by 11 students and we therefore have insufficient statistical evidence to say that its bias is significant. Modules 14 and 16 would be unlikely to come under scrutiny based on comparing simple module means, but the student-conditioned bias analysis shows them to be outliers.

Given the large number of modules, it seems unsurprising that some of them are outliers at the 5% level. A random draw from 25 unbiased modules would on average generate 1.25 biased modules ($25 \times 5\% = 1.25$). However, eleven (nearly half) of the modules appear to have statistically significant biases at the 5% level. This is important because the student-conditioned module biases suggest that issues to do with marks may need to be addressed.

It is possible to identify five broad scenarios when comparing any one module's simple mean-based bias with its student-conditioned module bias:

- Neither the simple bias nor the student-conditioned bias identify a module as an outlier and therefore no issues of marking inconsistency arise[4] (e.g. Module 12).
- Both the simple bias and the student-conditioned bias identify a module as an outlier. One may wish to explore if this bias is due to marking standards or teaching features (e.g. Module 1).
- The simple bias suggests an outlier module but the student-conditioned bias does not. The marking is probably consistent and the high (or low) marks in the module are probably due to the module being taken by students who performed better (or worse) than the student cohort as a whole (e.g. Module 2).
- The simple bias does not indicate an outlier but the student-conditioned bias does. This suggests there is a hidden source of bias, for example, the module may have been taken by students who overall performed well but the marking or teaching led to low scores (e.g. Module 11).
- A fifth, unlikely, scenario may emerge where the simple bias and the student-conditioned bias are in opposite directions but both are significant outliers. For example, a module may have been taken by students who on average were poor performers but the marking or teaching led them to achieve high scores, and this module might warrant closer analysis (though insignificant, Module 17 has biases in opposite directions).

The analysis of this dataset has shown that student-conditioned biases may identify different modules as outliers than a simple module mean comparison. Therefore, we suggest that calculating student-conditioned module biases is useful if we want to identify outliers more effectively.

5. Extensions

The fixed effects regression analysis proposed in this paper can be extended in a number of ways, some statistical and some in the modelling. The statistical extensions are relatively straightforward. The modelling extensions may enable a better understand of why the module biases arise. We are unable to implement these with our dataset due to data confidentiality restrictions and such additional data may not be easily available within many academic settings. In addition, we also mention other applications for our modified fixed effects technique.

A simple statistical extension is to produce parsimonious regression results by a step-wise removal of statistically insignificant module dummies. Typically, the removal process is stopped when the parameter on the least significant variable achieves a 5% or 1% significance level. The resulting parsimonious regression retains the significant outlier module dummies and the diagnostic statistics gain power due to the omission of the insignificant ones. Step-wise regression was first proposed by Gorman and Toman (1966) based on a suggestion by Efroymson (1960).

An alternative statistical approach is to estimate a random effects regression where the student fixed effects are assumed to have a normal distribution (see Greene, 2012, Section 11.5). This implies that student ability, or effort, is normally distributed thus imposing a structure on the otherwise freely determined fixed effects. The results of the random effects regression is reported in Table 4. We can see from the last two columns of this table that the implied module biases in the random effects model and the fixed effects model are similar to one another. Though the standard errors indicate much greater statistical significance for the parameters estimated by random effects, the Hausman tests at the bottom of Table 4 indicate that the structure implied by this random effect model is rejected and the fixed effects model is preferred. To be more precise, both Hausman tests reject the hypothesis that the always more efficient random effects estimator is in this case consistent and instead selects the always consistent, but less efficient, fixed effects estimator. In this case at least, the biases implied by the fixed effects estimators are more reliable.

[4] Subject to the caveat that these modules have not been both marked harshly and taught unusually well (or *vice versa*) so that the two effects cancel each other out.

Table 4
Random effects regression on Marks awarded.

	Random effects regression results			
	β_i	*s.e.*	R.E. module biases (Eqs. (5) and (6))	F.E. module biases (Tables 2 and 3)
Constant (β_0)	49.51	1.02**		
Module_1			−8.83	−8.36
Module_2	5.89	1.63**	−2.93	−1.62
Module_3	2.45	2.27	−6.38	−6.91
Module_4	2.38	2.63	−6.45	−7.27
Module_5	5.88	4.65	−2.94	−2.65
Module_6	5.95	1.10**	−2.88	−2.51
Module_7	8.05	1.46**	−0.77	−0.25
Module_8	5.35	2.94	−3.48	−4.33
Module_9	9.58	1.72**	0.75	1.53
Module_10	7.72	1.25**	−1.10	−0.75
Module_11	6.76	1.71**	−2.07	−2.88
Module_12	9.32	1.42**	0.50	1.03
Module_13	9.86	4.14*	1.04	0.46
Module_14	11.34	1.08**	2.52	3.14
Module_15	9.35	2.94**	0.52	−0.33
Module_16	10.68	1.19**	1.86	2.01
Module_17	8.77	2.94**	−0.05	−1.53
Module_18	9.50	3.05**	0.67	−0.37
Module_19	11.10	2.87**	2.27	0.20
Module_20	14.54	1.14**	5.72	6.24
Module_21	12.89	2.89**	4.06	3.15
Module_22	12.05	2.47**	3.23	1.91
Module_23	19.67	2.43**	10.84	11.39
Module_24	21.17	3.87**	12.35	13.06
Module_25	16.83	3.14**	8.00	6.58
Within-R^2	0.2531			0.2554
Observations	1044			1044

Hausman tests of fixed effects versus random effects

Hausman (1978) matrix-difference method: $H=71.60^{**}$
Arellano (1993) regression method: $H=71.50^{**}$
Both Hausman tests reject the hypothesis that the R.E. estimator is in this case consistent and instead select the F.E. estimator.

* $p<0.05$.
** $p<0.01$.

The first proposed modelling extension is to add individual instructor dummies to check if there is an instructor effect in the marking. One of the instructor dummies would have to be excluded in order to avoid collinearity. These instructor effects can be added to either regression equations (1), (4) or (8), and in the first case this becomes:

$$\text{Mark}_{id,md} = \sum_{s=2}^{S} \gamma_s \text{Instructor\_}S_{id,md} + \sum_{i=2}^{M} \beta_i \text{Module\_}i_{id,md} + \sum_{j=1}^{N} \alpha_j \text{Student\_}j_{id,md} + \epsilon_{id,md}$$

where the γ_s are the instructor effects. The estimation is only viable if no single instructor teaches just one module, otherwise the instructor dummy would be collinear with the module dummy β_i in which they teach.

Another extension is to analyse a dataset that spans more than one cohort of students. Again, we could achieve this by modifying regression equations (1), (4) or (8), and in the first case this would become:

$$\text{Mark}_{id,md,t} = \sum_{t=2}^{T} \tau_t \text{Year\_}t_{id,md,t} + \sum_{i=2}^{M} \beta_i \text{Module\_}i_{id,md,t} + \sum_{j=1}^{N} \alpha_j \text{Student\_}j_{id,md,t} + \epsilon_{id,md,t}$$

where the estimated parameters τ_t identify if there are any cohort effects on the marks and t indexes the year. The advantage of the multi-cohort analysis is the increase in data-points and the identification of long-term module biases.

Sometimes it is felt that students perform less well in compulsory modules. This can be modelled by including a zero-one dummy C that indicates if a particular student was compelled to take a particular module, using the regression:

$$\text{Mark}_{id,md} = \delta C_{id,md} + \sum_{i=2}^{M} \beta_i \text{Module\_}i_{id,md} + \sum_{j=1}^{N} \alpha_j \text{Student\_}j_{id,md} + \epsilon_{id,md}$$

Identifying the compulsion to take a module may not always be straightforward because some modules may be compulsory for some students but optional for others.

Developing innovative teaching methods, such as problem-solving exercises, and evaluating their impact is an important aspect of teaching development. Buckridge and Guest (2007) debate the tradeoffs between different teaching approaches. One practical contribution to the above debate would be to identify teaching methods, traditional or innovative, that provide the most 'value added'. To compare all modules where teaching is purely by traditional lectures to those using innovative methods one could simply insert identifier dummies. However, it may be that only a proportion of a module is taught using innovative methods and we should allow for the fact that mixing traditional and innovative methods may lead to learning outcomes that are greater (or lower) than the sum of their parts. A flexible modelling strategy is therefore to estimate a regression equation such as:

$$\text{Mark}_{id,md} = \rho_1 P_{id,md} + \rho_2 P^2_{id,md} + \sum_{i=2}^{M} \beta_i \text{Module\_}i_{id,md} + \sum_{j=1}^{N} \alpha_j \text{Student\_}j_{id,md} + \epsilon_{id,md}$$

where P is the proportion of the module taught using innovative teaching methods and its square P^2 is included to allow for a non-linear response. For example, a small proportion of innovative teaching may bring positive benefits but as this proportion grows the additional benefits diminish until they become detrimental. The exact form of this response is determined by the shape of the function implied by estimates of ρ_1 and ρ_2. Of course, additional teaching methods can be analysed by including additional ratios in the regression equation but variation may also be due to variations in the methods of assessment. For example, Krieg and Uyar (2001) identify how student performance is affected by multiple choice versus essay-based tests and Tian (2007) identifies how performance is affected by take-home assignments versus in-house examinations. One sometimes also comes across contradictory arguments on whether students do better on technical modules. This too can be modelled by including a variable that indicates the technicality of modules but it raises the question as to whether technical content can be captured by a zero-one dummy, an index, or a continuous variable.

The fixed effects approach in Eq. (4) or (8) may also have other applications where a metric varies along two categorical dimensions. For example, in Section 2 we mentioned the widespread use of student evaluation of teaching to monitor instructors' performance. Simply comparing average instructor scores can cause similar distortions to the comparison of mean module marks as discussed in the Introduction. If an academic institution was able to link each student's teaching evaluations to one another while still guaranteeing their anonymity from the instructors, it could control for the student fixed effects because any one student may systematically provide high or low evaluations.

6. Conclusion

Monitoring the consistency of marks is obviously important because of the incentives that inconsistent marks creates for students and because of the challenges facing instructors due to increasing student numbers and fees. Fixed effects regression, and the resulting student-conditioned module mark biases, provides a useful tool for checking the consistency of marks in each module. These biases offer a statistical analysis that is more sophisticated than the comparison of simple

module means because they use student fixed effects to accommodate individual students' overall performance. Any remaining statistically significant module bias will be instructor-related only, i.e. due to teaching standards or marking level, and instructors may then be best placed to decide what action to take. For example, a module with high marks may have been marked consistently if it was taken by students who also achieved high scores in their other modules.

In summary, we suggest estimating our modified fixed effects regression equation (8) and using the corrected standard errors in Eq. (9) to test the significance of the student-conditioned module module biases. The dataset used in this paper indicates the extent to which the student-conditioned module mark biases can identify modules as being outliers. Some modules that seemed outliers in the simple comparison of module means were not so under the student-conditioned analysis. Conversely, some modules that were not outliers in the simple analysis do appear to be outliers when student ability is controlled for. This highlights that in practice it is important to account for student performance when evaluating the consistency of module marks.

References

Arellano, M., 1993. On the testing of correlated effects with panel data. J. Econom. 59 (1–2) 87–97.
Blackmore, J., 2009. Academic pedagogies, quality logics and performative universities: evaluating teaching and what students want. Stud. Higher Educ. 34 (8) 857–872.
Buckridge, M., Guest, R., 2007. A conversation about pedagogical responses to increased diversity in university classrooms. Higher Educ. Res. Dev. 26 (2) 133–146.
Efroymson, M.A., 1960. Multiple regression analysis. In: Ralston, A., Wilf, H.S. (Eds.), Mathematical Methods for Digital Computers. Wiley, New York, pp. 191–203.
Galton, F., 1886. Regression towards mediocrity in hereditary stature. J. Anthropol. Inst. Great Britain Ireland 15, 246–263.
Gorman, J.W., Toman, R.J., 1966. Selection of variables for fitting equations to data. Technometrics 8 (1) 27–51.
Greene, W.H., 2012. Econometric Analysis, 7th ed. Pearson, New York.
Hadsell, L., MacDermott, R., 2012. Faculty perceptions of grades: results from a national survey of economics faculty. Int. Rev. Econom. Educ. 11 (1) 16–35.
Hammersley-Fletcher, L., Orsmond, P., 2004. Evaluating our peers: is peer observation a meaningful process? Stud. Higher Educ. 29 (4) 489–503.
Hausman, J., 1978. Specification tests in econometrics. Econometrica 46, 1219–1240.
Isley, P., Singh, H., 2005. Do higher grades lead to favorable student evaluations? J. Econom. Educ. 36, 29–42.
Knight, P., 2006. The local practices of assessment. Assess. Eval. Higher Educ. 31 (4) 435–452.
Krieg, R.G., Uyar, B., 2001. Student performance in business and economics statistics: does exam structure matter? J. Econom. Finan. 25 (2) 229–241.
Lizzio, A., Wilson, K., Simons, R., 2002. University Students' Perceptions of the Learning Environment and Academic Outcomes: implications for theory and practice. Stud. Higher Educ. 27 (1) 27–52.
McPherson, M.A., 2006. Determinants of how students evaluate teachers. J. Econom. Educ. 37, 3–20.
Nesbit, P.L., Burton, S., 2006. Student justice perceptions following assignment feedback. Assess. Eval. Higher Educ. 31 (6) 655–670.
Nicholls, G., 2005. New lecturers' constructions of learning, teaching and research in higher education. Stud. Higher Educ. 30 (5) 611–625.
Nowell, C., 2007. The impact of relative grade expectations on student evaluation of teaching. Int. Rev. Econom. Educ. 6, 42–56.
Nowell, C., Lewis, R.G., Handley, B., 2010. Assessing faculty performance using student evaluations of teaching in an uncontrolled setting. Assess. Eval. Higher Educ. 35 (4) 463–475.
Orr, S., 2007. Assessment moderation: constructing the marks and constructing the students. Assess. Eval. Higher Educ. 32 (6) 645–656.
Price, M., 2005. Assessment standards: the role of communities of practice and the scholarship of assessment. Assess. Eval. Higher Educ. 30 (3) 215–230.
Tian, X., 2007. Do assessment methods matter? A sensitivity test. Assess. Eval. Higher Educ. 32 (4) 387–401.
Thomas, R.H., 1976. The necessity of examinations – and their reform. Stud. Higher Educ. 1 (1) 23–29.
Webber, D.J., 2005. Reflections on curriculum development, pedagogy and assessment by a new academic. Int. Rev. Econom. Educ. 4, 58–73.