EXAMINATION MARKING AND ECONOMIC MODELS*

Derek Leslie

Abstract

Examination marking is often inaccurate. This inaccuracy is shown to be structurally different from the classical approach to errors in variables. Many economic problems can be analyzed within this generic " teacher-student" framework where grading errors are the central feature. Examples are submission to a peer reviewed journal, the job market, loans, crime, the market for lemons and matching problems such as marriage. Students decide whether to have their examinations graded based on a rational assessment of the costs and expected benefits. This decision takes into account the teacher's ability to grade accurately in addition to student assessment of the true grade. (JEL C5, J15)

Key words; response errors, examination marking, signal extraction, selection

* Department of Economics, Manchester Metropolitan University, Manchester M156BG, United Kingdom. E-mail d.leslie@mmu.ac.uk

Many economic problems resemble examination marking on close inspection. Individuals decide whether it is worthwhile to undertake some particular activity (equivalent to entering an examination) and someone else (the examination grader) judges that activity. Passing the examination implies a reward, but failure involves a cost. Section V gives six examples; submitting an article for peer review: applying for a job: crime: loans: the market for lemons and marriage. All of these problems can be thought of as a form of examination marking.

The particular focus is the uncertainty surrounding the student's true grade, which is typical in realistic applications. It is not always possible to mark examinations with complete accuracy. If two competent teachers grade an essay or dissertation independently, in general, they award different marks. Both give an honest assessment of what they think the correct grade is. Examination marking is, therefore, an excellent paradigm to study economic problems that involve measurement errors. The key argument is that the classical approach is not the right way to think about errors in examination marking and *a fortiori,* it is not the right way to think about errors in a wide class of economic problems.

Econometric textbooks are dominated by the classical approach to measurement error, and it is not surprising that many researchers believe that this is the only plausible treatment.[1] John Bound *et al.* (2001, p.3709) comment "Researchers virtually always rely on the assumption that measurement error is classical, usually with no justification at all". The assumption here is that examination marking errors are independent of reported values rather than the true values as in the classical approach. What this means in practice is that those who grade do not make systematic mistakes in their marking. Dean R. Hyslop and Guido

W. Imbens (2001) refer to this as the optimal prediction error (OPE) model. William Fuller (1987, p.79) also briefly discusses the OPE model and the idea is to champion this less usual approach. Whereas these focus on the econometric implications, the purpose here is to show that the OPE approach to measurement errors should be given more prominence in economic models, especially where rational expectations are a feature.

Students often do not have a choice whether to have their examinations graded. Once entered, there is no possibility of deciding afterwards not to let the teacher see the script. However, from Section II onwards, it is assumed that students do have this opportunity. The reason for this strange assumption is that the type of economic problem envisaged often involves a self-selection rule, for example whether or not to apply for a particular job. There may be an advantage in deciding not to be graded, because there is a saving on the marking fee, which could include non-pecuniary costs such as, among others, the humiliation of failure. The point of interest is the selection rule. Assuming that decisions are made rationally, how do the numbers of students who find it worthwhile to have the examination graded by the teacher change as uncertainty varies? Do more select to be graded as teachers become more competent? Section II shows that the answer to this question is not straightforward and unusual outcomes are possible.

Errors are the key driving feature. Both the student and the teacher have an imperfect idea about the true examination mark, so there is two sided uncertainty. Both do their best with the objective to be fair and unbiased markers, but both are aware that the student and the teacher grades are subject to error. Students take account of the fact that the teacher is an imperfect marker in deciding whether to be graded, as well as their own assessment of how well they have done. This twist is important in practical situations. As an application, think of the potential criminal (student). The decision to undertake a criminal activity (have the exam graded) is influenced by how well the criminal thinks the authorities (the teacher) can accurately detect (assess the grade) the criminal activity.

I. The Examination Marking Paradigm

A typical structure for the classical error model is

(1) $q = z + v$,

where q is the observed variable, $z \sim N(\bar{z}, \sigma_z^2)$ is the true value and $v \sim N(0, \sigma_v^2)$ is the error term. The key point of the classical approach is that q and v are positively correlated. Assuming z and v are uncorrelated, then the covariance between q and v is σ_v^2. It is well known that this error generating mechanism leads to biases in the estimated coefficients in a regression model. Alan B. Krueger and Mikael Lindahl (2001) provide an example (there are probably thousands in the literature) where noisy data are used as an explanation for the poor performance of many human capital models.

Is the classical approach a useful starting point in thinking about examination marking errors? Think of q in (1) as the mark awarded by the teacher. It is composed of two components. The first is z, which is the unobservable true mark and v represents the teacher's unavoidable error. As a model of examination marking, this has a fatal drawback. The problem is that if q is the observed distribution of marks, then a person awarded a higher than average grade ($q > \bar{z}$) tends to receive too generous a mark and contrariwise a person with a below average grade ($q < \bar{z}$) tends to receive too low a mark. If a teacher grades according to this statistical model, he or she quickly realizes that a high q is partly the result of generous marking, not a high z.

Why should a teacher systematically over-mark and under-mark at the opposite ends of the grading scale? It is implausible to believe that people with high marks are treated with more generosity than they typically deserve and those with low grades are systematically hard done by. The classical measurement error model is inappropriate because in this instance a competent teacher has personal insight into the fact that grades can never be awarded with complete accuracy. The errors are not mechanically generated.

The competent teacher would reasonably claim that his or her marks are unbiased along all points of the marking scale, not just at the mean value, which is what a naïve interpretation of (1) implies. Nobody can deny this is what a fair-minded teacher aspires to achieve. Believers in rational expectations would also reject the classical error structure for the teacher's marking errors.

The examination marking model has the property that the expected value of the error for any given awarded mark is zero, i.e. the teacher is not systematically biased in the way that the classical approach implies. The "insightful" teacher recognizes that the classical approach does not meet this objective and implicitly weights q to eliminate the bias. Here a linear weighting scheme is considered. Let the reported marks be

(2) $\tilde{z} = a + bq$,

where a and b are weights. The distribution of marking errors is then

(3) $\varepsilon = z - \tilde{z} \Rightarrow z = a + bq + \varepsilon$.

Being unbiased at all points along the marking scale requires that $E\varepsilon \mid \tilde{z} = 0$, i.e. $\mathrm{cov}(\tilde{z}, \varepsilon) = 0$. It is easily seen that $\mathrm{cov}(\tilde{z}, \varepsilon) = (1-b)b\sigma_z^2 - b^2\sigma_v^2$ and the value of b that ensures a zero covariance is calculated as $b = \dfrac{\sigma_z^2}{\sigma_z^2 + \sigma_v^2}$ and $a = \bar{z}(1-b)$ ensures that $E(\tilde{z}) = \bar{z}$. This also utilizes the information from the examination scripts in the most efficient way, because the weighting scheme minimizes σ_{ε}^{2}.

In the classical model, *q* is the distribution of the observed marks and *v* is the distribution of errors. In the examination marking model $\tilde{z}$ is the distribution of observed marks and ε is the distribution of errors. No complicated statistical knowledge is required to generate $\tilde{z}$ in the same way as an expert pool player can pot balls without any understanding of Mechanics. It is the idea of being fair, whilst recognizing that complete accuracy is impossible.

This is a signal extraction model. In a typical signal extraction problem, *q* is the observed (noisy) variable. For example, in the famous Robert E. Lucas (1977) model, *q* refers to an observed price rise, which is composed of an unobserved general price rise (*z*) and an unobserved relative price rise *(v)*. The best guess of *z* is $\tilde{z}$. In the examination marking model $\tilde{z}$ is reported directly and *q* is unobserved.

The examination marking model effectively reduces to the OPE model with $z = \tilde{z} + \varepsilon$, and $\operatorname{cov}(\tilde{z}, \varepsilon) = 0$. Writing it in the form of (2) and (3) is useful, where *v* is derived from *z* and ε ($v = [(z - \bar{z})(1 - b) - \varepsilon]/b$). It shows the connection with the idea of signal extraction using noisy scripts to extract information about true grades. Although (2) and (3) appear more elaborate than is strictly necessary, they give a more tractable analysis in the end. The *b* parameter ($0 \le b \le 1$) measures the accuracy of grading, with *b=1* representing perfect marking.

The OPE approach to examination marking has much to commend it. The variance of the reported marks is $b\sigma_{z}^{2}$, which is less than the variance of the true grades as long as the marking is less than accurate.[2] An incompetent teacher, who is unable to extract any useful information about the true grades from the scripts, simply assigns each candidate the average score $\bar{z}$ in this statistical model of examination marking. By contrast, the classical errors in variables model typically predicts the opposite. In practice, a smaller variance is more likely. Teachers are frequently observed to be reluctant to use all points of the marking scale. Grading the range of papers that an economist normally takes illustrates this. Mathematics and econometrics papers (where *b* is likely to be close to 1) generally show a bigger spread of grades than `essay' type examinations.

Table (1) illustrates the tendency of regression towards the mean in awarding grades. It shows the results of the 2003 round of the British Civil Service Selection Boards (CSSB) for the Fast Stream. This is a major annual competition that recruits elite Civil Servants, who are expected to fill the most senior posts as their career progresses. Around 18,000 apply, but only 1398 made it to the final selection boards in 2003. The boards involve a series of tests spread over three days, where much of the assessment is imprecise and judgmental. Decisiveness, lucidity, robustness, impartiality, ability to collaborate, interpersonal sensitivity and adaptability are each assessed on a marking scale, which are then (again based on assessors' judgment) mapped into an overall point score. Those graded four or better are successful, but grades better than four mean candidates are picked out for the most desirable jobs. Similarly, near miss candidates might have an opportunity to re-apply – so the scale does matter and CSSB assessors spend much time deciding on precise grades. Table (1) shows that only 3.3 percent of candidates are found in three out of the seven categories and none in the top category. Clearly, it is implausible to believe that nobody in 2003 was in the top grade one – after all the competition aims to find the very best graduates across all British universities. Nevertheless, the signal extraction view of grading suggests that the assessors (who themselves are an elite of top civil servants and nobody's fool) are nevertheless reaching correct (unbiased) judgments.

They are implicitly recognizing that the competencies assessed to be a top civil servant can only be imprecisely measured. All this is despite the best efforts of the *Cabinet Office,* who explicitly advise assessors "the full range of the scale is meant to be used".

There is also a lot of evidence that errors in survey data are not classical which is reviewed in Bound *et al.* (2001).[3] However, an important point to make is that the OPE model is different from the so-termed mean reverting error model, which is sometimes referred to as a non-classical error. John Bound and Alan B. Krueger (1991) discuss this case where they explore direct evidence with response data containing measurement errors matched with observations on the true value found from independent sources. The approach is to posit (1) as the error generating mechanism, but with a covariance term between *z* and *v.* Let this covariance be $\rho\sigma_v\sigma_z$. They describe this as a non-classical mean reversion model. In this case

$$(4) \qquad \begin{bmatrix} v \\ q \end{bmatrix} \sim \begin{bmatrix} 0 \\ \bar{z} \end{bmatrix}, \begin{bmatrix} \sigma_v^2 & \sigma_v^2 + \rho\sigma_v\sigma_z \\ \sigma_v^2 + \rho\sigma_v\sigma_z & \sigma_z^2 + \sigma_v^2 + 2\rho\sigma_v\sigma_z \end{bmatrix}.$$

The question is whether the examination marking model is a special case of this. If *q* is the reported distribution, setting the covariance between *q* and *v* to zero (as in the examination marking model) requires

$$(5) \qquad \rho = -\frac{\sigma_v}{\sigma_z}.$$

Interestingly, Bound and Krueger (1991) find evidence of a fairly strong negative covariance or mean reversion effect, which is highly suggestive that the examination marking model may be a better starting point than the classical errors model.

For (5) to hold requires that the variance of the errors is less than the variance of the true values, because ρ cannot be less than -1. The examination marking model is not so restricted. It is somewhat disingenuous to claim that only some teachers are capable of being unbiased. The examination marking model looks at the problem in a different way, giving a good reason for a negative covariance, compared with an *ad hoc* "covariance corrected" classical errors model.

II. Examination Marking and Self-selection

A model of two sided uncertainty is set out, but in order to draw out some interesting issues the initial focus is on a special one sided uncertainty case. The special case supposes that students are able to grade with complete accuracy, but the teacher is unable to match this precision. Despite this, students are never asked to grade themselves because such a process lacks credibility. Students have every incentive to cheat in reporting their grades. Now suppose that students who pass the examination are awarded a prize, with opting to be graded being voluntary but subject to a marking fee. Students self-select into those for whom the gamble is worthwhile. Self-selection is partly determined by a knowledge about how accurately the teacher grades. It turns out that there is sometimes a specific degree of teacher marking inaccuracy that makes students self-select into only those with a passing grade or better. Logically all these students should pass and furthermore the teacher knows this. However, because the teacher marks inaccurately many of these students are failed. There appears to be a classic time inconsistency problem. The short-run best solution is to pass everyone without grading the papers, but then this alters the self-selection behavior. Next time round everyone elects for assessment, because students know papers are not actually graded. Rejection of some papers that are known to be of passing quality is necessary to ensure the integrity of the self-selection process. It is shown how credibility can be restored by altering the way in which scripts are marked and varying the entrance fee in a very simple way.

The assumption that students can grade perfectly is far from ridiculous. First, making extreme assumptions gives useful insights about more realistic situations. Secondly, in many applications the student may be the better grader compared with the teacher. Think of the accused (student) facing a jury (teacher), where the student knows the truth and the jury is unsure. Alternatively, think of the person selling a used car. The seller knows the truth about mechanical condition, but the purchaser is unsure. Thirdly, an examination may be the means to evaluate a student's understanding of a course rather than merely the content of an examination script. Students may have a much better insight as to the amount of deep learning that has occurred than the teacher.

Both the teacher and students follow the examination marking model when grading. The statistical model replicates the discussion of eqs (2) and (3). For the teacher

$$\begin{aligned} &z = \tilde{r} + \varepsilon = \alpha + \beta(z + \varphi) + \varepsilon \qquad \operatorname{cov}(\tilde{r}, \varepsilon) = 0 \\ &\alpha = (1-\beta)\bar{z} \quad \beta = \frac{\sigma_z^2}{\sigma_z^2 + \sigma_\varphi^2} \end{aligned} \quad . \tag{6}$$

$\tilde{r}$ is the distribution of awarded grades, $z \sim N(\bar{z}, \sigma_z^2)$ is the distribution of true marks, ε is the distribution of marking errors and $\varphi = [(z - \bar{z})(1-\beta) - \varepsilon]/\beta \sim N(0, \sigma_\varphi^2)$. The parameter β $(0 \le \beta \le 1)$ is a measure of the accuracy of the teacher, with higher values indicating greater accuracy and β = 1 is when the teacher is perfect marker. A similar set-up prevails for students' evaluation of their own performance,

$$(7)\qquad \begin{aligned} &z = \tilde{q} + \zeta = \gamma + \pi(z+\delta) + \zeta \qquad \operatorname{cov}(\tilde{q},\zeta) = 0 \\ &\gamma = (1-\pi)\bar{z} \quad \pi = \frac{\sigma_z^2}{\sigma_z^2 + \sigma_\delta^2} \end{aligned} \quad .$$

$\tilde{q}$ is the distribution of student grades. The parameter π measures student accuracy, which typically differs from the teacher.

The breakeven probability that makes it worthwhile to have the examination graded is set at p^*. This is the fee/prize ratio in the risk neutral case, but p^* can exceed this if there is risk aversion and in practical cases the "fee" and "prize" are monetary equivalents reflecting a variety of costs and rewards. The same p^* is assumed for all. If students assess the probability of passing as $\geq p^*$, they elect to have the examination graded. Students use $\tilde{q}$ to calculate this probability and know β and π when making this calculation. The decision to enter is rationally based.

If the teacher grades perfectly, then the expected pass rate is $P(z \geq z_T)$, where z_T is the pass mark standard. However, the teacher can only set a target pass mark for the *observed* grades. Denoting this pass mark as $\tilde{r}_T$, then the expected pass rate is $P(\tilde{r} \geq \tilde{r}_T)$. If the teacher's objective is to fix z_T and make $P(\tilde{r} \geq \tilde{r}_T) = P(z \geq z_T)$, then $\tilde{r}_T$ must vary with β. In actual fact, nothing substantive changes if $\tilde{r}_T$ is fixed at z_T, but the former assumption is a more appealing idea and leads to a selection equation which is symmetric in β and π.

The rational student knows his or her $\tilde{q}$ value and calculates $P(\tilde{r} \geq \tilde{r}_T \mid \tilde{q})$, where the parameters of (6) and (7) are known. This means that the accuracy or otherwise of the teacher in his or her ability to grade examinations has an influence on the student's decision whether or not to be graded. A higher $\tilde{q}$ makes it more likely to pass (excepting the extreme case where $\pi = 0$). A breakeven value of $\tilde{q}$ can therefore be calculated above which it is worthwhile to enter. The appendix shows that the critical value is $\breve{q}$, where

(8) $$\breve{q} = \frac{\breve{z}}{(\beta\pi)^{0.5}} - \breve{p}\left[\frac{1-\beta\pi}{\beta\pi}\right]^{0.5} .$$

$\breve{q}$ is the standardized critical value of $\tilde{q}$, thus $\breve{q} = 0$ means exactly 50 percent choose to be graded and a higher value means fewer elect to be graded. As well as β and π, this depends on $\breve{z}$ and $\breve{p}$. $\breve{z}$ is the standardized value of z_T, i.e. $\breve{z} = (z_T - \bar{z}) / \sigma_Z$. Thus $\breve{z} = 0$ means the pass mark is $\bar{z}$, and higher values mean a higher z_T. $\breve{p}$ is the standardized value of p^*, thus $\breve{p} = 0$ means the breakeven probability is 50 percent. Students who assess their chances of passing at or better than 50 percent enter in this case. Higher values of $\breve{p}$ mean a fall in the breakeven probability, meaning that more elect to be graded *ceteris paribus.*

This selection equation has a very simple structure.[4] One sided uncertainty problems are just special cases of this general selection rule. Derek Leslie (2004) explores one such case without showing how it fits within the general framework explored here. The problem concerns the decision whether to submit a paper for peer review to an academic journal. The person submitting is the student, who inaccurately assesses the paper's worth and the teacher is the referee, who is assumed to have β=1.

The special case considered here is the complementary problem when students are able to grade themselves perfectly with π = 1, but the teacher is inaccurate. With perfect student grading, $\tilde{q} = z$. This case is interesting because it highlights the issue of credibility. Students use $\tilde{q}$ to determine those who elect to be graded and the teacher uses $\tilde{r}$ to determine the grades. The teacher cannot use information from $\tilde{q}$ directly. To see this consider those values of β when $\breve{q} = \breve{z}$. This means that only those with a passing grade self-select to be graded by the teacher. Despite this, students that enter know that there is a chance (depending on the value of β) that the teacher fails them. There are two such possible values. The trivial case is when β =1. In this baseline case there is no teacher uncertainty; only those who pass elect to be graded and the omniscient teacher duly passes them. However, the more interesting case occurs when

$$(9) \qquad \left(\frac{1+\beta^{0.5}}{1-\beta^{0.5}}\right)^{0.5} = \frac{\breve{z}}{\breve{p}}.$$

This square root exists when either $\breve{p}$ and $\breve{z}$ are positive or when they are both negative. In addition

$$(10) \qquad \beta^{0.5} = \frac{\breve{z}^2 - \breve{p}^2}{\breve{z}^2 + \breve{p}^2}.$$

This square root exists when $\breve{z}^2 > \breve{p}^2$. The four cases (with $\pi = 1$) show how self-selection varies as β goes from 1 to 0.

1. When $\breve{z}^2 > \breve{p}^2$ and $\breve{p}$ and $\breve{z}$ are both positive, the number of students who elect to be graded at first increases and later steadily declines to zero. When (10) holds the numbers who enter exactly equals those with $z \geq z_T$.
2. When $\breve{z}^2 > \breve{p}^2$ and $\breve{p}$ and $\breve{z}$ are both negative, the number of students who elect to be graded at first decreases and later steadily increases towards 100%. As before, there is a crossover point when the numbers who enter are exactly those with $z \geq z_T$.

3. In other cases the number of students who elect to be graded either steadily declines or steadily increases.

As an example, if the pass mark is set at the highest 30 percent of the z distribution and the fee is such that only those with a probability of passing at or above 40 percent think it worthwhile to enter, then it turns out that $\beta = 0.386$ ensures that exactly the top 30 percent of candidates decide to be graded. However, $\beta = 0.386$ means that only 58.8 percent of those graded are predicted to pass the exam.

(10) illustrates the credibility issue, where some information cannot be directly exploited. In this special case, the preferred option is to ignore all the information from $\tilde{r}$. The problem is that if this is done, students exploit this knowledge and report misleadingly high grades. Passing everyone who elects to be graded is equivalent to this because, if everyone is passed, everyone opts for grading. This is not a satisfactory state of affairs. In this extreme case it appears necessary to fail some students (even though the teacher knows they should all pass) to enforce the credibility of the selection rule. Is this unfair? At first blush, this seems to be the case; after all, it goes against common sense to think that some students are failed, even though it is known that they all should pass. At a deeper level, perhaps it is not unfair to failed students. The selection rule is nothing more than a sophisticated gamble and students enter if the bet is "in the money" when $p \geq p^*$. Students take on the bet with their eyes open. They know in advance that there is a risk of failure, even though they know they have a passing grade when they decide to be graded. The objective odds dictate they should enter, and like all bets that are in the money there should be no regrets if the wrong horse wins. However, the logic is somewhat uncomfortable. As an example think of the job market and the case where the

rejected candidates come from a minority group. It is hard to justify the morality of a position where candidates with a passing grade are rejected.

Can the teachers exploit the selection information when (10) holds? The teacher's objectives are to increase the pass rate and also to reduce the marking load. Can this be credibly achieved? The answer is yes. Suppose the teacher announces that a randomly selected fraction of submitted scripts will be passed without any attempt at grading, but with the rest graded as usual. If a student had previously assessed his or her probability of passing at *p,* the new probability is

(11) $$p(m) = m + (1 - m)p\,,$$

where *m* is the fraction automatically passed. Clearly for any given $\tilde{q}$, the probability of passing has increased, which causes more students to wish to be graded. However, raising the marking fee, which raises the breakeven probability that makes it worthwhile to be graded, discourages this process. Consequently, if the fee is suitably adjusted, it is possible to achieve the same (all passing) proportion of students who opt for grading. Figure 1 illustrates this idea. The lower line (typically non-linear), labeled $f(\tilde{q})$, plots the relationship between the probability of being passed by the teacher and students' normalized $\tilde{q}$, when $m = 0$. The critical value $\breve{q}$ associated with $p*$ is shown. The upper curve shows the relationship when the teacher automatically passes the fraction *m*. This is just $m + (1 - m) f(\tilde{q})$. It can be seen that raising $p*$ to $p** = m + (1 - m)p*$ ensures the same critical value $\breve{q}$.

So what does this adjustment process achieve? The teacher gains, because there are fewer scripts to mark and they collect more in fees. Students gain because a larger fraction of scripts is passed overall. They pay for this because the prize for passing remains the same, but they must pay a higher fee in return for a higher pass rate. It can be seen from the figure, that there is no real upper limit on *m* as long as it lies below 1. So in the limit as $m \to 1$, fewer scripts are graded and more students are passed. It appears that the teacher can exploit $\tilde{q}$ without requiring students to reveal their grades directly (when they have an incentive to lie).

The same idea can be applied in other circumstances. Suppose case 3 prevails and that those who elect to be graded steadily decline as β falls. With $\beta < 1$, some students with a passing grade deem it not worthwhile to enter. The same principle of a random audit can be applied, but now the objective is to change p^* to ensure that all those with a passing grade choose to enter. Similarly, if the numbers who enter steadily increases, with too many now entering when $\beta < 1$, the two instruments of a random audit and changes in p^* can ensure that only those with a passing grade have their examinations graded. In the limit, the same condition holds in all circumstances. Set the fee such that it is only worthwhile to be graded if there is a very small but finite chance of being audited. Then set *m* close to one. Credibility requires that students believe there is a finite probability of being audited, otherwise all failing students enter.

III. Two Sided Uncertainty

With two sided uncertainty there are four cases among those that are graded. There is a probability p_1 that the individual is passed ($\tilde{r} \geq r_T$) and is of a passing standard ($z \geq z_T$). There is a probability p_2 that the individual is failed but is of a passing standard. There is a probability p_3 that the individual is passed but is of a failing standard. Finally, there is a probability p_4 that the individual is failed and is of a failing standard. For example, with $\beta = 0.4$ and $\pi = 0.9$, then 29.6 percent enter, p_1 = 53.0 percent, p_2 = 32.7 percent, p_3 = 4.8 percent and p_4 = 9.5 percent. Because of two sided uncertainty, mistakes are inevitable. A smaller p_2 and p_3 is the more desirable outcome.

The same random audit procedure can be applied to manipulate these probabilities. Let m denote the fraction automatically passed as before. If the breakeven probability of entry is maintained at the existing level, by manipulating the fee/reward structure, then the new probabilities are

$$\begin{aligned} p_1(m) &= p_1 + mp_2 \\ p_2(m) &= (1-m)p_2 \\ p_3(m) &= p_3 + mp_4 \\ p_4(m) &= (1-m)p_4 \, . \end{aligned} \tag{12}$$

Raising m increases $p_1(m)$ and reduces $p_2(m)$. Since $p_1(m)$ are the true passes and $p_2(m)$ are the incorrect failures, this is a desirable outcome. However, this comes at the expense of an undesirable increase in false passes $p_3(m)$.[5] It is now a matter of preference as to whether the random audit procedure is applied and by how much. The higher $p_2 - p_4$ then the bigger is the reduction in $p_2(m)$ relative to any increase in $p_3(m)$. The pay-off to a random audit increases directly with the size of $p_2 - p_4$. Effectively if students do a good job at self-evaluation (π is high) then p_4 tends to be low. Similarly, if the teacher is not so good at grading (β is low) then p_2 tends to be high. Hence the random audit is a better option when π is high and β is low. In the special case when $\pi = 1$, then p_4 is zero and the random audit is always desirable because there is no downside to the trade-off (other than a higher entrance fee).

IV. Knowing the Truth

If students always reveal the truth about $\tilde{q}$, student evaluations could be used as a second marker, just as in actual examinations when double marking is used to improve accuracy. The teacher and student grades are then combined to give the best estimate of the true mark. Let the combined grades be[6]

(13) $\tilde{c} = \lambda\tilde{r} + (1-\lambda)\tilde{q}$.

The weights are chosen to minimize the variance of the errors of the combined grades (in general, simply averaging the two marks is not the best option in blind double marking). The selection rule is now based on $P(\tilde{c} \geq c_T \mid \tilde{q})$, i.e. students take account of the fact that their own (honest) evaluations are combined with the teacher grade when deciding whether to enter (c_T is the implied passing grade). Hence the probability of entry differs for any given π and β compared with the previous situation. As an example, suppose π = 1. In this case students know that $\tilde{c} = z$ irrespective of the value of β. Only those who know they will pass, therefore, enter. This is 30 percent if the pass mark is set at the highest 30 percent of the z distribution. In the previous case the selection probability changes.

In the absence of any additional information, the combined grade gives the maximum degree of accuracy for the true examination mark. There is accordingly little incentive for the random audit procedure, given that the information from $\tilde{q}$ has been utilized efficiently. With the same parameter values as before, p_2 falls to 0.9 percent and p_4 falls to 1.4 percent.

V. Some Applications

Examination marking is applicable to situations where an individual decides whether to undertake some activity and where, if the activity is undertaken, someone else judges it. Both sides are uncertain as to the exact value of the activity. The examination marking model does not mirror every detail of the following examples, but it should be clear how examination marking describes the basic structure. In other words, it may be a good reference point and it is always useful to recognize that different problems have a common parent model.

A. Submitting to an Academic Journal.

Students are those thinking of submitting to an academic journal, where z is the distribution of true quality of academic papers and $\tilde{q}$ is the distribution of authors' opinions of their work. The teacher is the journal's refereeing process, with $\tilde{r}$ representing the ability of the referee to identify z. The journal aims to publish papers with $z \geq z_T$. Leslie (2004) explores this application.

B. Applying for a Job

Students are those thinking of applying for a job, where z is the distribution of productivity and $\tilde{q}$ the distribution of potential applicants' opinion of their productivity. The teacher is the interviewing panel, with $\tilde{r}$ representing the distribution of the evaluation of productivity of potential applicants. The aim is to hire those with $z \geq z_T$. Edmund S. Phelps (1972) explores a special case of this type model.

C. Crime

Students are those thinking of undertaking some action, where z is a distribution measuring the "morality" of the action. Only those actions with $z \geq z_T$ are considered moral, otherwise they are considered immoral and subject to sanction. The problem is that both the potential criminal and the authorities are uncertain as to what the true value of z is. The selection equation in this case tells us which actions are undertaken. The teacher represents the judgment on those actions, where $\tilde{r}$ represents the ability or otherwise of authority to correctly identify immoral actions. The examination marking model recognizes a key point about crime, namely that potential criminals are influenced by β. This is a measure of the degree to which criminals think they can “get away with it”.

D. Loans

Students are those thinking of borrowing money to finance a project, where z measures the returns to the project. Only projects that involve $z \geq z_T$ are financially viable. The teacher is the potential lender where $\tilde{r}$ represents the lender’s ability to evaluate z. Joseph E. Stiglitz (1987) analyses this.

E. The market for lemons.

Students are those thinking of selling something, where z is a distribution of quality. Only those with $z \geq z_T$ are not substandard (known as lemons in the USA). The teacher is the potential purchaser, and $\tilde{r}$ represents the purchaser’s evaluation of z. George A. Akerlof (1970) explores this model.

F. Marriage.

M and F are in the marriage market.[7] M is therefore both the teacher and student of F and likewise F is both the student and teacher of M. Courtship (M marks F and F marks M) only occurs if both M and F decide it is worthwhile to enter. This is not straightforward, because the entry decision depends on not just whether M will be passed by F, but also on M's view that F will be acceptable to M (and vice versa for F's decision). Marriage takes place when both M and F are mutually passed. Divorce occurs when either M and F or both eventually realize that the true grade falls short of the awarded (passing grade) mark. Unlike most examination marking processes, courtship sometimes lasts several years and this is where problems may occur. M and F are most probably in love, so the danger is that M allows F (the student) to grade herself and likewise F allows M to grade himself. The examination marking model warns that this is likely to lead to misreporting. Against this, true love may lead to the co-operative solution of Section V, which gives the most accurate grades. All one can say is that if M and F have a proper understanding of examination marking, divorce is less likely (aim for Section V examination marking) and if divorce happens recrimination is minimized. Even with Section V joint marking, it sometimes turns out that $\tilde{c} \geq c_T$ but $z < z_T$ so no complaints. Gary S Becker (1973; 1974) analyses the marriage market.

VI. Concluding Comments

Applying the OPE framework to examination marking makes sense if it is believed that rational agents do not make systematic mistakes in grading. It is a far better anchor point than an *ad hoc* adjustment to the classical approach. An inherent characteristic of this model is its relevance for the analysis of a range of problems. The basic structure is deliberately simple and the two equations (6) and (7) are a good start point for a class of economic problems that involve a selection rule. The model can be further developed. For example, though the pass mark is exogenous, actual applications could endogenise this. As an example, the familiar job search model of Stephen A. Lippman and John J. McCall (1976) establishes the optimal value of the reservation wage, which maximizes the expected present value of job search. The reservation wage is the pass mark in an examination marking interpretation of the job search model. The fixed reward for a passing grade is useful in highlighting the credibility issue, but may be too sharp for some applications. Rather than a fixed reward, this could be linked to the awarded grade.

APPENDIX

A. Deriving the Selection Rule

To calculate $P(\tilde{r} \geq r_T \mid \tilde{q})$, consider the following linear relationship between $\tilde{r}$ and $\tilde{q}$,

(A1) $$\tilde{r} = f + g\tilde{q} + e \quad .$$

$P(\tilde{r} \geq r_T \mid \tilde{q}) = P(e \geq r_T - f - g\tilde{q})$. This depends on the values of *f* and *g* as well as the distribution of *e.* It can be seen that

(A2) $$\begin{bmatrix}\tilde{r}\\ \tilde{q}\end{bmatrix} \sim \begin{bmatrix}\bar{z}\\ \bar{z}\end{bmatrix} \begin{bmatrix}\beta\sigma_z^2 & \pi\beta\sigma_z^2\\ \pi\beta\sigma_z^2 & \pi\sigma_z^2\end{bmatrix}.$$

Hence

(A3) $$g = \frac{\sigma_{\tilde{r},\tilde{q}}}{\sigma_{\tilde{q}}^2} = \beta \quad f = (1-\beta)\bar{z}$$

and

(A4) $$\sigma_e^2 = \sigma_{\tilde{r}}^2 - \frac{[\sigma_{\tilde{r},\tilde{q}}]^2}{\sigma_{\tilde{q}}^2} = \beta^2(\frac{1}{\beta} - \pi)\sigma_z^2$$

Next derive the critical value of $\tilde{q}$ (denoted as $\tilde{q}\,*$) associated with $p*$, by calculating $P(e \geq r_T - a - b\tilde{q})$ for the breakeven probability $p*$. Students with $\tilde{q} \geq \tilde{q}\,*$ self-select to have their examinations graded by the teacher.

(A5) $$\sigma_e \breve{p} = r_T - (1-\beta)\bar{z} - \beta\tilde{q}\,*,$$

where $\breve{p}$ is the critical value of the standard normal distribution associated with $p*$.

Let $\breve{z}$ be the critical value of the standard normal distribution associated with z_T, i.e. $z_T = \bar{z} + \breve{z}\sigma_Z$. As noted r_T is set so that $P(\tilde{r} \geq r_T)$ equals $P(z \geq z_T)$. Hence

(A6) $$r_T = \bar{z} + (z_T - \bar{z})\beta^{0.5} = \bar{z} + \breve{z}\sigma_z\beta^{0.5}.$$

Substituting into (A5)

(A7) $$\tilde{q}* = \bar{z} + \frac{\breve{z}\sigma_z}{\beta^{0.5}} - \frac{\breve{p}\sigma_e}{\beta}.$$

Letting $\breve{q}$ be the critical value of the standard normal distribution associated with $\tilde{q}*$, and noting that $\tilde{q} \sim N(\bar{z}, \pi\sigma_z^2)$, it follows that

(A8) $$\breve{q} = \frac{\breve{z}}{(\beta\pi)^{0.5}} - \breve{p}\left[\frac{1-\beta\pi}{\beta\pi}\right]^{0.5}.$$

B. Selection using the combined grades

The combined grades are $\tilde{c} = \lambda\tilde{r} + (1-\lambda)\tilde{q}$. The weights are chosen to minimize

(A9) $$E(z-\tilde{c})^2 = \lambda^2 E(z-\tilde{r})^2 + (1-\lambda)^2 E(z-\tilde{q})^2 + 2\lambda(1-\lambda)E(z-\tilde{q})(z-\tilde{r})$$

The optimal value of λ is [8]

(A10) $$\lambda^* = \frac{\beta(1-\pi)}{\beta+\pi-2\beta\pi}$$

The selection rule uses $P(\tilde{c} \geq c_T \mid \tilde{q})$, where as before c_T varies to ensure that $P(\tilde{c} \geq c_T) = P(z \geq z_T)$. To calculate $P(\tilde{c} \geq c_T \mid \tilde{q})$, consider the following linear relationship between $\tilde{c}$ and $\tilde{q}$.

(A11) $$\tilde{c} = f + g\tilde{q} + e \quad .$$

Going through as before, the equivalent of (A8) is calculated as

(A12) $$\breve{q} = \frac{\breve{z}\sigma_{\tilde{c}}}{g\pi^{0.5}\sigma_z} - \frac{\breve{p}\sigma_e}{g\pi^{0.5}\sigma_z}$$

where $g = \sigma_{\tilde{c},\tilde{q}} / \sigma_{\tilde{q}}^2 = \lambda\beta + 1 - \lambda \quad f = \lambda(1-\beta)\bar{z}$

$\sigma_{\tilde{c}} = [\lambda^{*2}\,\sigma_{\tilde{r}}^2 \ + \ (1-\lambda^{*2})\sigma_{\tilde{q}}^2 \ + \ 2\lambda^*(1-\lambda^*)\sigma_{\tilde{r},\tilde{q}}]^{0.5}$ and

$\sigma_\varepsilon = [\sigma_{\tilde{c}}^2 - (\sigma_{\tilde{c},\tilde{q}})^2 / \sigma_{\tilde{q}}^2]^{0.5}$

Although (A12) looks complicated, ultimately it is just a function of β, π, $\breve{z}$ and $\breve{p}$.

REFERENCES

Akerlof, George A. “The Market for `Lemons’: Quality Uncertainty and the Market Mechanism.” *Quarterly Journal of Economics,* August 1970, 84(3), pp. 488-500.

Becker, Gary S. “A Theory of Marriage: Part I.” *Journal of Political Economy,* July-Aug. 1973, 81(4), pp. 813-846.

---------------------. “A Theory of Marriage: Part II.” *Journal of Political Economy,* March - April 1974, 82(2), pp. S11-S26.

Block, Francis and Ryder, Harl. “Two-sided Search, Marriages, and Matchmakers.” *International Economic Review,* February 200, 41(1), pp. 93-155.

Bound, John and Krueger, Alan B. “The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs make a Right?” *Journal of Labor Economics,* January 1991, 12(1), pp. 345-368.

Bound, John, Brown, Charles and Mathiowetz, Nancy “Measurement Error in Survey Data.” In Heckman, James J. and Leamer, Edward, *Handbook of Econometrics Volume 5.* Amsterdam: Elsevier, 2001.

Clotfelter, Charles T. “Tax Evasion and Tax Rates: An Analysis of Individual Returns.” *Review of Economics and Statistics,* August 2003, 65(3) pp. 263-73.

Dougherty, Christopher. *Introduction to Econometrics.* Oxford: Oxford University Press, 2nd ed., 2002.

Fuller, William A. *Measurement Error Models.* New York: John Wiley, 1987.

Greene, William H. *Econometric Analysis.* New Jersey: Prentice Hall, 5th ed., 2003.

Gujarati, Damodar N. *Basic Econometrics.* New York: McGraw Hill, 4th ed., 2003.

Hill, Carter R., Griffiths, William E. and Judge, George G. *Undergraduate Econometrics.* New York: John Wiley, 2nd ed., 2001.

Hyslop, Dean R. and Imbens, Guido W. "Bias from Classical and Other Forms of Measurement Error." *Journal of Business and Economic Statistics*, October 2001, 19(4), pp. 475-81.

Kmenta, Jan Elements *of Econometrics.* Michigan: University of Michigan Press, 2nd ed., 1997.

Krueger, Alan B. and Lindahl, Mikael "Education for Growth: Why and for Whom." *Journal of Economic Literature,* December 2001, 39(4), pp. 1101-36.

Leslie, Derek "Applying the Rational Expectations Hypothesis to the Case of Blind Double Marking off Examinations." *Discussion Paper, Manchester* Metropolitan University, 2003.

_________ . "Are Delays in Academic Publishing Necessary?" *Discussion Paper,* Manchester Metropolitan University, 2004 (forthcoming *American Economic Review).*

Lippman, Steven A. and McCall, John J. "The Economics of Job Search: A Survey." *Economic Inquiry,* September 1976, 14(3)Pp. 347-68.

Lucas, Robert E. "Understanding Business Cycles." *Journal of Monetary Economics*, Supplementary series 1977, *5*(0), pp. 7-29.

Maddala, G. S. *Introduction to Econometrics.* Chichester, UK: Wiley, 4th ed., 2001.

Phelps, Edmund S. "The Statistical Theory of Racism and Sexism." *American Economic Review,* September 1972, *62*(4), pp. 659-61.

Pissarides, Christopher. *Equilibrium Unemployment Theory,* Oxford, UK: Basil Blackwell 1990.

Stiglitz, Joseph E. "The Causes and Consequences of the Dependence of Quality on Price." *Journal of Economic Literature,* March 1987, 25, pp. 1- 48.

Thomas, R. L. *Modern Econometrics.* Harlow, UK: Addison Wesley, 1997.

FIGURE 1. HOW PROBABILITIES CHANGE IF A

FRACTION *m* IS AUTOMATICALLY PASSED

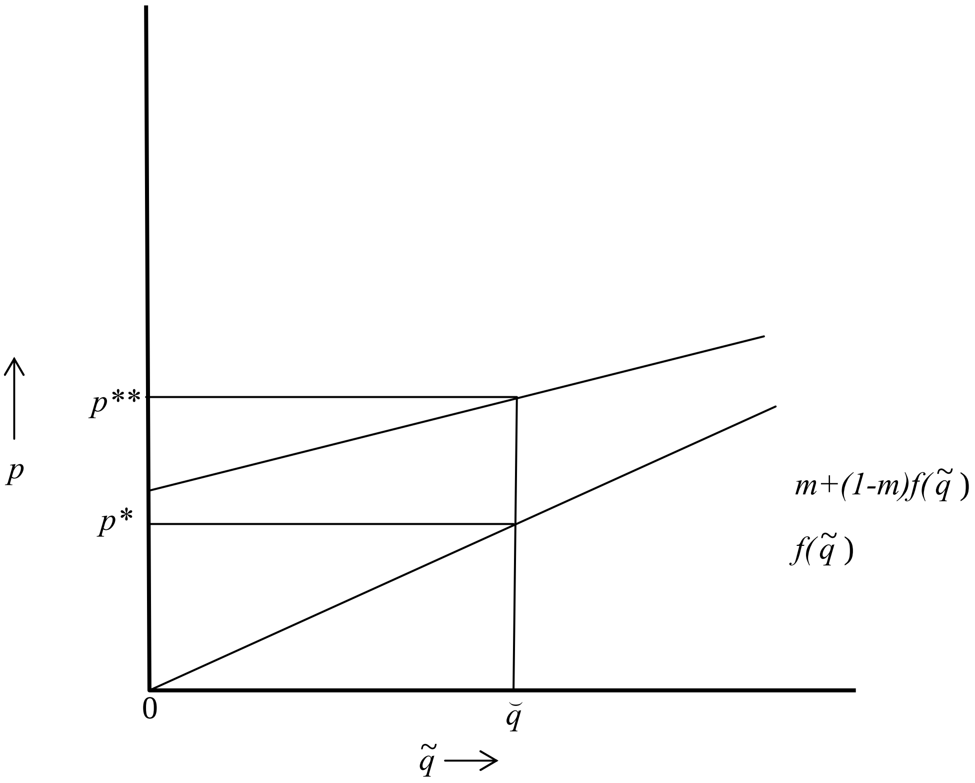

Table 1

RESULTS OF 2003 (NOV. 02-JUL. 03) CSSB COMPETITION FINAL ROUND

Final Board Mark	Numbers
1 (Highest grade)	0 (0%)
2	32 (2.3%)
3	211 (15.2%)
4	167 (12.0%)
5 (Failing grade)	708 (51.1%)
6	254 (18.3%)
7 (Lowest grade)	14 (1.0%)

ENDNOTES

[1] Examples are: Christopher Dougherty (2002), William H. Greene (2003), Damondar N. Gujarati (2003), Carter R. Hill *et al.* (2001), Jan Kmenta (1997), G. S. Maddala (2001) and R. L. Thomas (1997).

[2] This is true even if $\text{cov}(z, v) \neq 0$. Derek Leslie (2003) explores a more general model of double blind marking.

[3] In some cases there will be an unwillingness to admit to socially undesirable behavior. See Charles T. Clotfelter (1983) on tax evasion as a motive to misreport income. Alcohol, cigarette consumption, sexual behavior are other examples where respondents are known to lie.

[4] The selection equation can be `individualized' by simply subscripting π and $\breve{p}$. So the critical value varies according to each student's own π and $\breve{p}$ values.

[5] This is not the whole population of passing scripts, because selection means that some with a passing grade will elect not to be graded. In the example 6.6 percent of those who are not graded have $z \geq z_T$.

[6] The appendix gives a more detailed derivation.

[7] "Marriage" is sometimes used as a metaphor for general matching problems. See Christopher Pissarides (1990) and Francis Bloch and Harl Ryder (2000).

[8] Leslie (2003) analyses the blind double marking problem. (A10) assumes that the students and teachers grade independently.