TEST SCORE COMPENSATION, MORAL HAZARD, AND STUDENT REACTION TO HIGH-SCORER DISINCENTIVES IN ECONOMICS PRINCIPLES

Johnnie B. Linn III, Concord University

ABSTRACT

At Concord University, principles of economics students have mean exam scores consistently lower than the percentage rate that traditionally represents a "C" average. Through a linear transformation, the scores are "curved" so that the mean score fits a C and the best score is not greater than 100%. The highest scorers have an incentive to expend less effort than they would if the curve compensated all students equally. The data for macroeconomics suggest that students anticipate the disincentives affecting the highest scorers and change their own effort accordingly. Evidence for disincentives in microeconomics is lacking. The difference might be accountable to microeconomics courses having a greater proportion of business majors, who are likely to invest more effort in the course.

INTRODUCTION

The author is in his third year of teaching economics principles courses at Concord University and has amassed enough data to make meaningful analyses of student performance in these courses. This paper seeks to ascertain to what extent the "curve" used for exam scores generates disincentives among the better scorers and to what extent these incentives are anticipated by the students in general.

At the beginning of the course, the rule for the "curve" transformation is explained in the syllabus. The transformation in done in two steps:

1. If the mean score in terms of percentage of questions answered correctly is less than 77%, the percentage score of all test takers is increased by the difference of 77% and the mean score percentage.
2. If Step 1 results in a percentage value of the highest score greater than 100%, a linear transformation of the raw scores is made so that the mean is 77% and the maximum is 100%.

For example, if the mean score on an exam with 25 questions is 18 (72%) and the highest score is 23 (92%), Step 1 would result in a mean of 77% and a highest score of 97%. Step 2 would not be used. If the mean score on an exam with 25 questions is 15 (60%) and the highest score is 23 (92%), Step 1 results in mean of 77% and a highest score of 109%. Step 2 would be used. If we let the linear transformation be represented as

$$y = mx + b \quad , \qquad (1)$$

where *y* is the curved grade and *x* is the raw score, the value of *m* and *b* are given by

$$m = \frac{k-77}{h-\bar{x}}, \qquad b = \frac{77h - k\bar{x}}{h-\bar{x}}. \qquad \text{(2a, b)}$$

where h is the highest score, *k* is the value to which the highest score is curved, and $\bar{x}$ is the mean. In this case, *m* has the value of 2.875 and *b* has the value of 33.875.

For a given $\bar{x}$, the values of *m* and *b* are sensitive to the value to which the highest score is curved. The value of *m* is seen to be positively correlated with *k* and the value of *b* is negatively correlated with *k*. Thus, when the data are evaluated and a correction to the transformation function is called for, such correction can be made by altering the value of *k*.

THE MODEL

The transformation for each exam is based on the performance of the students in that exam, so the question arises, what curve will be "selected" by the students? In what ways will the students' performance be affected by the curve that they expect to be given? To answer these questions, we assume the following model of behavior.

Test takers maximize utility functions whose arguments are test score and effort. Rank is a predictor of test score but is not an explicit argument of the utility function. The number of students taking the exam (class size) is sufficiently large that effort on the margin by any one student has negligible effect on the mean score. The items on the exam are arranged in increasing order of difficulty so that the marginal product of effort is monotonically decreasing. We assume for the moment that the effort on the margin by any one student has negligible impact on the expected score transformation. We then have an objective function that can be expressed as

$$U = U[y(x(e)), e], \qquad (3)$$

where *e* is effort, *x* is the raw score, and *y* is the transformed score. We assume for the moment that guessing makes no contribution to the value of *x*. We establish the following elasticized parameters for the utility function:

$$\frac{y}{U}\frac{\partial U}{\partial y} = \alpha, \qquad\qquad (4a,b)$$
$$\frac{e}{U}\frac{\partial U}{\partial e} = -\beta.$$

All students have a positive benefit, α, from higher test scores. Students for whom leisure is a normal good have positive values of β; students who enjoy the challenge of taking test questions will have negative values of β.

The first-order condition is

$$\frac{\alpha y' x'(e) U}{e} - \frac{\beta U}{e} = 0, \qquad (5)$$

or

$$\alpha y' x'(e) = \beta. \qquad (6)$$

We can make a number of general conclusions from this equation. For a linear score transformation, the value of y' for all students is *m*. If all students have the same utility function, they will sort themselves out such that they all show a marginal product of effort determined by the local ratio of β to α, as shown in Figure 1.

{FIGURE 1 Here}

If the students differ widely in the degree of curvature of their production functions, a change in *m* could alter their rankings, and could cause a change in leadership. If we assume that the test items are of a reasonably varying degree of difficulty*, i.e*, *x(e)* is reasonably bow-shaped, we can disregard the possibility of a change in leadership.

If all students have the same production function of effort, they will sort themselves out in order of preference for leisure; those with higher values of β will have higher values of x' and lower values of *x*, as shown in Figure 2. A reasonably small change in *m* will not alter the students' rankings. An increase in curvature in the production function lessens the impact of preference for leisure on scores.

{FIGURE 2 Here}

If only Step 1 is used in the transformation process for an exam with 25 questions, the value of *m* is 4. Step 2 always results in value of *m* less than 4. A reduction of *m* results in a substitution of leisure for effort. Figure 3 shows the effect of an "income-compensated" curve, *i.e.*, a transformation that allows the student to select the same combination of leisure and test score as before:

{FIGURE 3 Here}

Without the curve, the student selects T_1, with the income-compensated curve, T_2. A mid-C-level student at T_2 whose score before the curve was 77 could expect to receive that same score after the curving process is complete, so the student would expect an over-compensation of income--an increase in the constant term of the transformation from Ob_2 to Ob_3--to raise the selected score to 77 at T_3. For a positive β this income effect reinforces the substitution effect. Curving the exam scores generates moral hazard for all students excepting the few that have values of β sufficiently negative to cancel the substitution effect. Furthermore, the greater variance in β, the greater variance in raw scores will result from the curve.

We now turn to another possible cause of moral hazard. The mid-C-level student selects from a family of (m, b) combinations any of which produces a grade of 77. The particular combination selected will depend upon what the score of the highest-ranking member of the class is expected to be. From Equations (2a) and (2b) we see that the value of h, the highest expected score, will force the values of m and b once the mean score has been selected. But for the highest-ranking student we must abandon the assumption made above that performance on the margin has a negligible effect on the transformation function. When Step 2 is performed, the value of y' for the highest-ranking student is zero for values of x greater than the expected raw score of the second-highest ranking student. The transformation function for the highest-ranking student will appear as shown in Figure 4.

{FIGURE 4 Here}

The three utility curves represent a positive β, a zero β, or a negative β for the highest-ranking student. For a relatively small class size, the likelihood of there being a student with a negative β is small, but for a larger class size, the likelihood of negative-β students is higher. But also as class size increases, the expected score of the second-highest-ranking student increases, so the scope of disincentive for the highest-ranking student decreases. In the limit, for infinite class size, we can expect that the second-highest-ranking student will have a perfect score and disincentive for the highest-ranking student will be completely eliminated, regardless of the value of β for that student.

By Equations (2a) and (2b), high-scorer disincentive will raise the value of m and reduce the value of b, but the latter will be affected only by a negligible amount under our assumption that with sufficient class size, any one score, even the highest score, has a negligible effect on the mean. The effect on m will be non-negligible, though, because only the highest score, and no other, is determinant of the value of m.

There is one additional issue to be addressed. What is the effect on student performance if guessing can make a positive contribution to the score? Let us suppose that the exam consists of multiple choice questions with the average likelihood c of a choice being correct. The expected raw percentage score of someone who turns in a completely randomized answer sheet would be c. The expected opportunity cost of attempting to answer a question is c times the value of the question, so the net marginal product of attempting a question is ($1 - c$) times the value of the question. If in Equation (3) we replace effort, e, with number of questions attempted, n, we obtain

$$U = U\left[y(x(n) + c(N - n)), n\right], \tag{7}$$

where N is the number of items on the exam. The equivalent first-order condition is

$$\alpha y'(x'(n) - c) = \beta. \tag{8}$$

We see that for a given number of questions attempted, an increase in c can be compensated for by an increase in y', that is, m. Thus allowing guessing and curving the exam are substitutes. A smaller curve should therefore be expected for multiple-choice exams than for non-multiple choice exams of the same difficulty.

Solutions of Equation (8) are possible for zero and negative values of β when x' is sufficiently small. The production function will have the form as seen in Figure 5:

{FIGURE 5 Here}

If guessing is not allowed, the student will select tangency point T on arc OAB. The marginal product of attempting the questions falls below the value of guessing, if guessing is allowed, at point A. The slope of AC is *c*. A student who initiates guessing at point A would receive a raw score of C, which when translated into utility space (assuming there is no disutility in the minimal amount of time needed to fill in the answers for the guessed questions) is equivalent to point D. Students are not constrained to initiate guessing at point A; those with positive β will initiate guessing before reaching point A. The production function with guessing is arc FDB, with a maximum score at point D, but the optimum score is at point E, with is compatible with initiation of guessing at point A'. Students with negative β who operate on the margin will have an optimum on arc DB and they initiate guessing after passing point A. They forego a higher score because they enjoy the challenge of attempting questions. The highest-ability students may have production functions without guessing whose slope never falls below *c*, in which case they will never guess.

The curve for a multiple-choice exam will reduce the slopes of OAB, FDB, and AC in the same proportion. The length of DC will be unaffected. Students with positive β will substitute leisure for effort, but students with negative β will substitute effort for leisure. Students with zero β will experience an income effect only. The transformation function for the highest-ranking student will have a flat-topped region in the vicinity of point D, bounded by the highest-scoring student on each flank. In the limit of class size, as in the case of non-multiple choice exams, the disincentive among the highest-ranking students is reduced to zero.

The way that we will test for student reaction to disincentive among the high scorers is to apply the envelope theorem to the students' utility functions. If students pay no attention to what *h* will be when they select *m* and *b*, the observed values of *m* and *b* will conform to what is expected when the envelope theorem is applied to a utility function whose arguments are only *m* and *b*. If *m* and *b* do not so conform, the hypothesis that students account for disincentive among the high scorers will not be rejected.

The envelope theorem as applied to the utility function, in either equation (3) or in equation (7), is

$$\frac{\partial U}{\partial m}dm + \frac{\partial U}{\partial b}db = 0\,, \qquad (9)$$

or

$$\frac{\partial y}{\partial m}dm + \frac{\partial y}{\partial b}db = 0\,. \qquad (10)$$

By applying Equation (10) to Equation (1), we find that

$$\frac{db}{dm} = -x\,. \qquad (11)$$

or, equivalently, in terms of *y*,

$$\frac{db}{y-b} = \frac{-dm}{m}\,. \qquad (12)$$

Integrating both sides, we obtain

$$-\ln(y-b) = -\ln(m) + C \qquad (13)$$

It will be noted that this result is independent of the values of *k*, α, and β. If students do not take account of disincentive among the highest scorers, the plot of the two sides of Equation (13) against each other should be perfectly linear, or a plot of the ratio of the logarithm arguments should have no slope.

If students do take account of disincentive among the highest scorers, we need to determine the sign of correlation to look for in the one-tailed test. The ratio of the two logarithm arguments in Equation (13) should have a positive slope, according to the following reasoning.

Let us suppose that students perceive the curve as follows:

$$y = m(1 - ph)x + b \,, \tag{14}$$

where *m* is unobservable and *p* is a small positive number that indicates to what extent the highest scorer is expected to reduce the marginal return to effort. When the envelope theorem is applied, the utility function has three arguments instead of two, with the following result

$$xd((1 - ph)m) + db - mpxdh = 0 \,, \tag{15}$$

or

$$\frac{db}{d((1 - ph)m)} = -x + mpx\frac{dh}{d((1 - ph)m)} \,. \tag{16}$$

The quantity (1 - *ph*)*m* is observable and is equivalent to *m* in Equation (11). How do we interpret the new term in the right-hand side? The students know that a higher value of *h* results in a lower value of *m* for a given value of *x*, from Equation (2a), so both terms on the right-hand side of Equation (16) are negative. Thus we have a negative enhancement of the correlation of changes in the observed values of *b* and of *m*. Approximating the new term as a constant multiple of *x*, and returning to our original Equation (11), we obtain

$$\frac{db}{dm} = -\lambda x \,, \tag{17}$$

where λ is greater than unity. Equation (12) becomes

$$\frac{db}{y - b} = \frac{-\lambda dm}{m} \,, \tag{16}$$

and Equation (13) can be re-written as

$$-\ln(y - b) = -\ln(m) - (\lambda - 1)\ln(m) + C \,, \tag{17}$$

the ratio of the two logarithm arguments in which is seen to have a positive slope.

DATA AND METHODOLOGY

The author's spring and fall classes taught in 2003 and 2004, excepting night classes, were the source of data for the analysis, for a total of 6 microeconomics classes and 8 macroeconomics classes. Each class incorporated four in-term exams and a comprehensive final. The four exams counted collectively for 60% of the grade and the final 40%. The values of *m* and *b* for the four in-term exams were tabulated for the study.

The values of *m* and *b* for the microeconomics exams are shown in Figure 6a and for the macroeconomics exams are shown in Figure 6b.

{FIGURES 6a and 6b Here}

Each data point is marked with a numeral indicating whether it is the first, second, third, or fourth exam of the course.

The coordinates of the upper left corner of the chart are selected to represent the curve that would be given for an exam for a class of infinite size in which all the students turned in completely randomized answer sheets. Such an exam would be considered worthless as an assessment tool by the students. Thus, exams whose coordinates fell closer to the upper left corner can be regarded as having done a worse job of assessment than those that fell closer to the lower right corner.

The macroeconomics data show a lower average value of *m*, indicating a wider distribution of scores, than the data for microeconomics. The difference in *m* is due partly to a difference in composition of the student populations for the two courses, as macroeconomics is likely to have a greater proportion of non-business majors, who are likely to invest less effort in the course. Also, the two courses are identical in what material is covered by the first exam, and most students who take both courses take macroeconomics first, so most of the students taking the first microeconomics exam had been exposed to the material twice [Linn, 2005]. So for the statistical tests, data for Exam 1 were excluded from consideration. Figures 7a and 7b show the distribution of curves with data from Exam 1 excluded.

{FIGURES 7a and 7b Here}

The data with Exam 1 excluded still show a difference between the microeconomics scores and the macroeconomics scores, which difference is accounted for by a difference in populations, as before mentioned, or by a difference in difficulty of the material.

THE RESULTS

Figures 8a and 8b show the points computed for the statistical test for reaction to disincentive among the highest scorers. The mean scorers were used to represent the typical student, so the value of y was set to 77. Two data points were excluded for microeconomics and one data point for macroeconomics for which Step 2 of the score transformation was not employed.

{FIGURES 8a and 8b Here}

Table 1 shows the t-statistics for the two courses. These statistics were generated using Web-based software provided by Wessa (2005).

Course	Sample Size	Coefficient	Constant	t for coefficient
Microeconomics	16	-0.020 (0.205)	2.66 (.226)	-0.099
Macroeconomics	23	0.46 (0.106)	2.16 (0.113)	4.39

Table 1. Statistics for the Test for Reaction to High Scorer Disincentive.

For macroeconomics, the statistics are significant, but for microeconomics, they are not significant. Why the difference?

One possible interpretation is that high scorers in macroeconomics exhibit disincentives but high scorers in microeconomics do not do so. As the macroeconomics courses have a larger ratio of non-business majors to business majors, and non-business majors—even those who are high scorers—are likely to invest less effort in the course, the likelihood of the high scorer having a positive β might be greater.

Tables 2a and 2b show the distributions of high scorers of Exams 2, 3, and 4 in the two samples by major. As many students had not yet declared their major at the time they were taking macroeconomics or microeconomics, a student was inferred to be a business major if he or she had taken or eventually took both courses. Ties in high score in a particular exam were resolved in favor of the student who showed more good scores in his or her other exams. In three instances in microeconomics, and four instances in macroeconomics, the same individual was the highest scorer in two of the three exams.

Highest Scorers	Business Majors	Non-Business or Unknown	Total
Two of Three Exams	3	0	6
One of Three Exams	6	4	10

Table 2a. Distribution of Highest Scores by Major, Microeconomics

Highest Scorers	Business Majors	Non-Business or Unknown	Total
Two of Three Exams	2	2	8
One of Three Exams	8	7	15

Table 2b. Distribution of Highest Scores by Major, Macroeconomics

Macroeconomics shows a lower proportion of high scorers who are business majors compared to those who are non-business majors or whose major is unknown.

Another possible interpretation is that the significant results for macroeconomics are due to some other unidentified factor, perhaps related to the difference in difficulty of the material.

CONCLUSION AND APPLICATIONS

The linear transformation, or "curve", that macroeconomics students select for their scores show that they expect the highest scorers to exhibit disincentives, since the latter are rewarded less by the curve. A remedy is to allow the highest score to curve to a value greater than 100%. Not only will high-scorer disincentive be reduced, but the marginal return to effort for all students will be increased, with the result that all students will work harder. If the hypothesis is correct that the difference in results between microeconomics and macroeconomics is that more of the macroeconomics high scorers have a greater propensity towards leisure, raising the marginal return to effort will have a greater impact in the macroeconomics than in the microeconomics courses.

REFERENCES

Linn, J.B., Presentation, "An Analysis of Score Differences of Business and Non-Business Majors for Microeconomics and Macroeconomics Principles", University of Kentucky Annual Teaching Conference, Lexington, KY 26 March 2005.

Wessa, P., Free Statistics Software, Office for Research Development and Education, version 1.1.17, URL http://www.wessa.net/, 2005.

Figure 1. Distribution of Effort, Identical Utility Functions

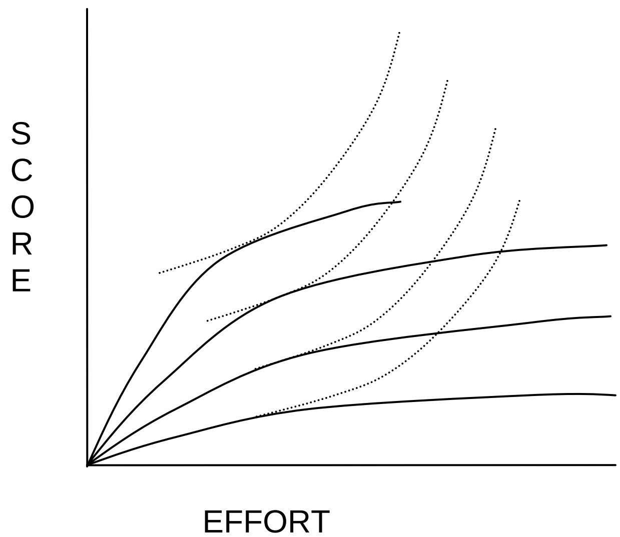

Figure 2. Distribution of Effort, Same Production Function

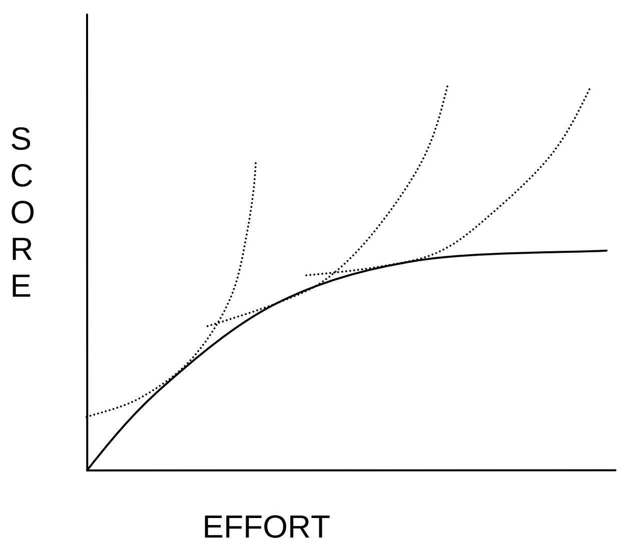

Figure 3. Substitution and Income Effects of Score Transformation

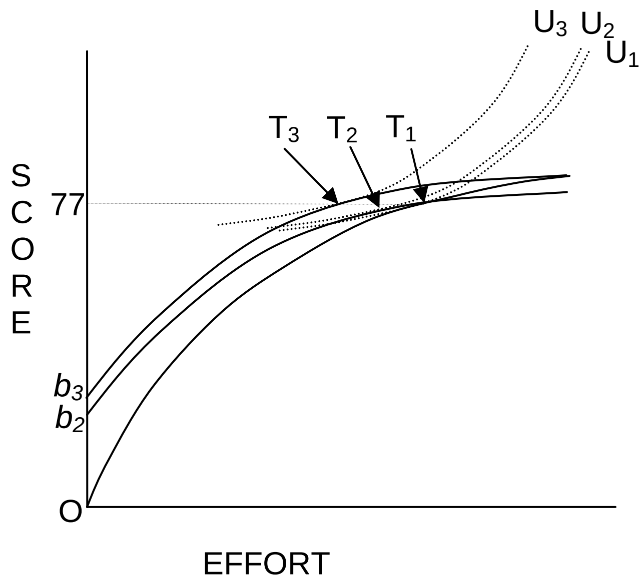

Figure 4. Transformation Function for the Highest-Ranking Student

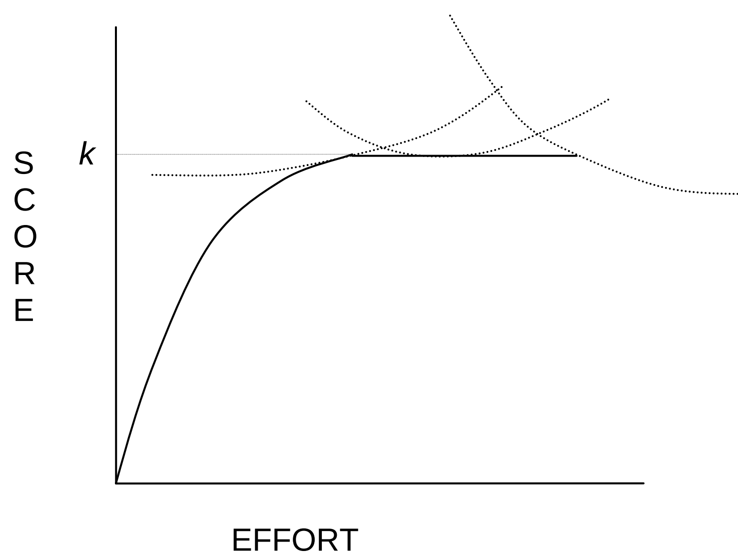

Figure 5: Production Function for a Multiple-Choice Exam

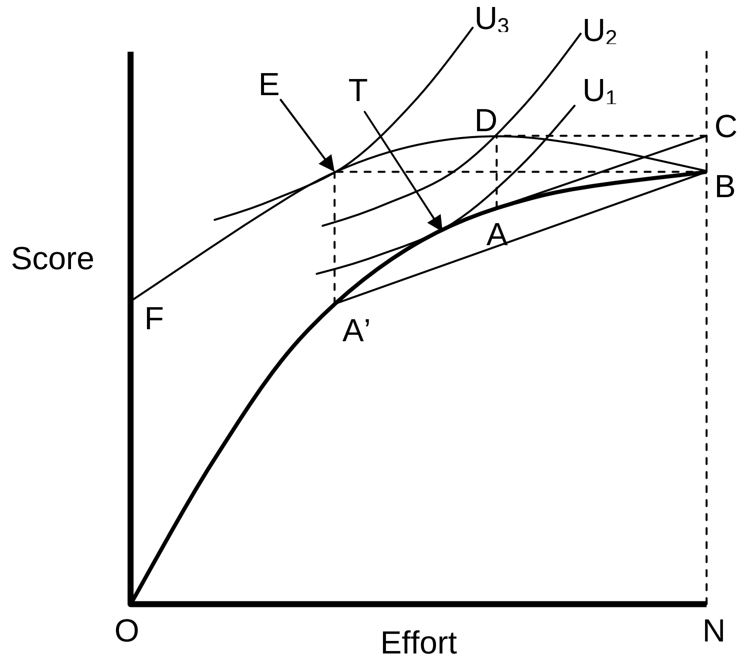

Figure 6a. Distribution of Curves for Microeconomics

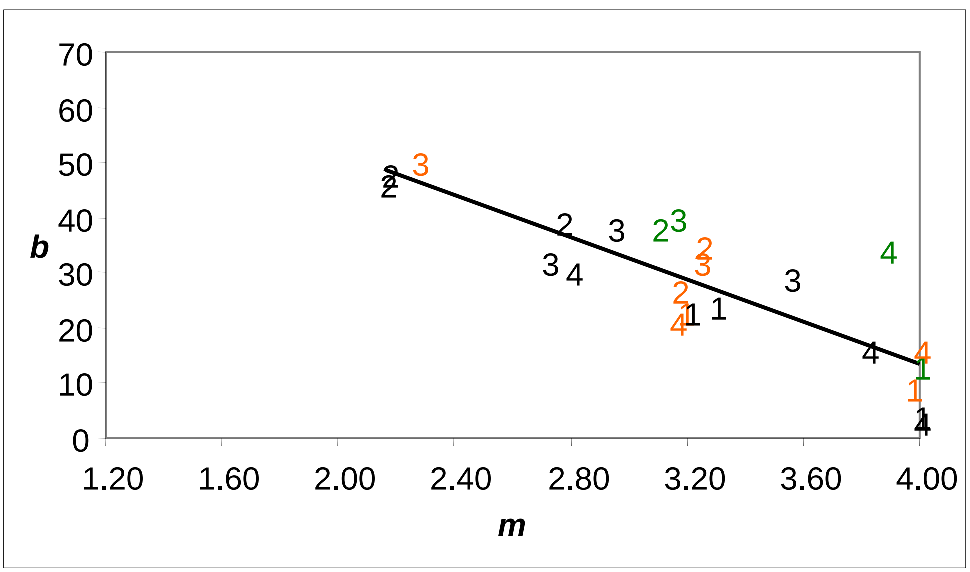

Figure 6b. Distribution of Curves for Macroeconomics

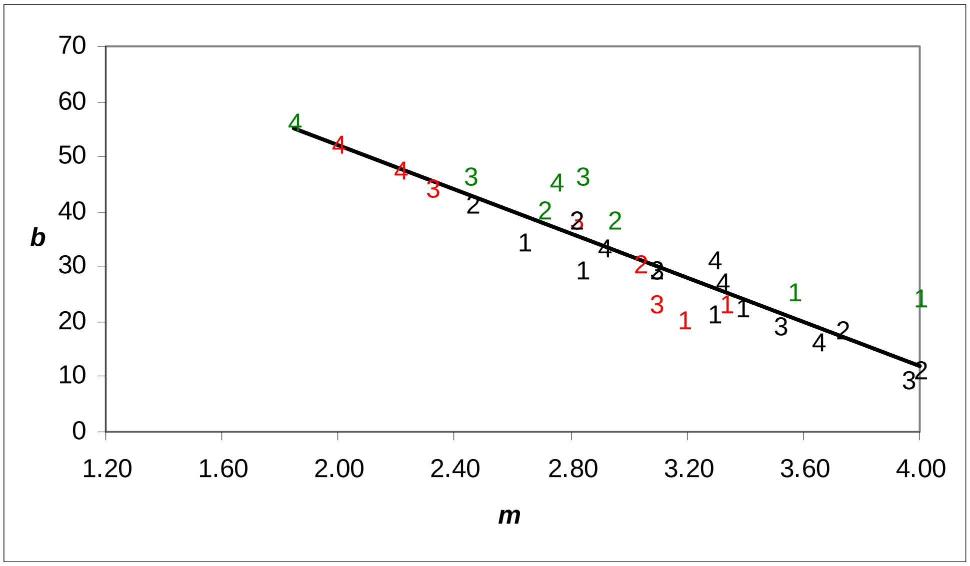

Figure 7a. Microeconomics Curve Distribution, Exams 2 – 4.

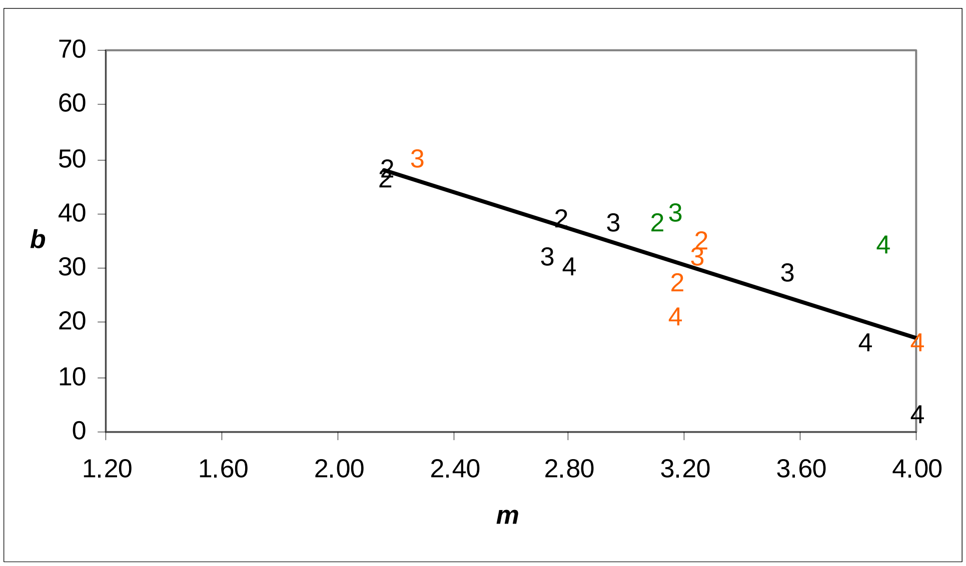

Figure 7b. Macroeconomics Curve Distribution, Exams 2 – 4.

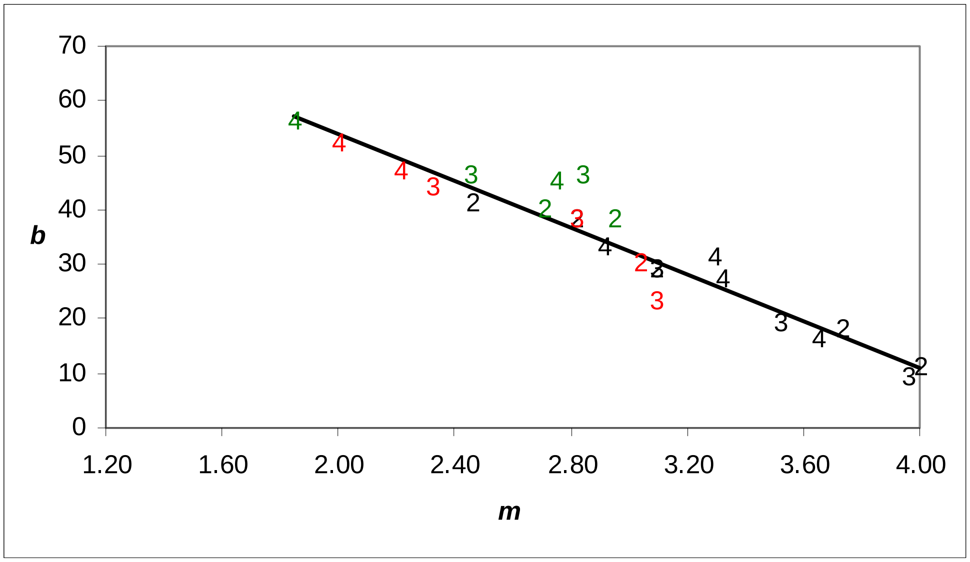

Figure 8a. Test for Reaction to High Scorer Disincentive, Microeconomics.

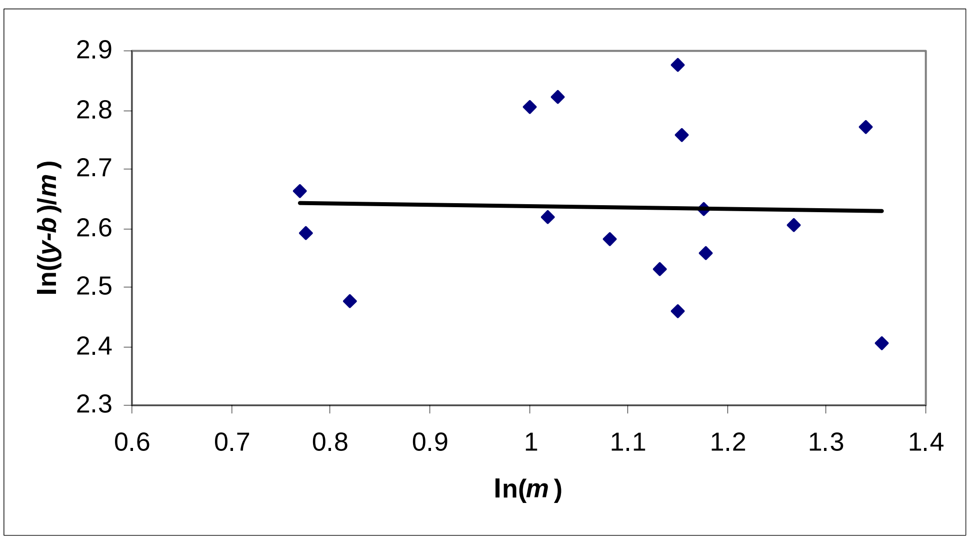

Figure 8b. Test for Reaction to High Scorer Disincentive, Macroeconomics.