Response of Economics Principles Students to a Change in High-Scorer Disincentives

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. The author used a linear transformation to "curve" the scores to a maximum not greater than 100 percent for the first half of the study, then changed the possible maximum to 110 percent in the second half of the study, to test if students when they select their curve transformation react to disincentives among the highest scorers. Results show that macroeconomics students reacted to changes in high scorer behavior but microeconomics students did not do so. The difference in outcomes is attributable a greater aversion to effort among macroeconomics students than microeconomics students. A highest score of 110 is overgenerous for both courses; highest scores of 106 and 104 are better for microeconomics and macroeconomics respectively.

The Status of the Research
In a previous paper, the author (Linn, 2006) described how a linear transformation of raw test scores on student effort produced different results for exams in microeconomics and macroeconomics. The maximum allowed adjusted score was 100 percent. Students in macroeconomics behaved in a manner consistent with their being aware that the highest scorers received less benefit from the score transformation and adjusted their own levels of effort accordingly, but students in microeconomics did not do so. The author set for the hypothesis that microeconomics students invest more effort in the course and are less likely to be influenced by the behavior of the high scorers. In the paper in hand, the researcher changed the transformation so that high scorers could earn up to 110 percent on their adjusted scores. In effect, high scorers could "bank" up to 10 points against possible lower scores on later exams.

During the course of the study, the author became aware that students were cheating by peeking at each other's papers, so he began preparing several versions of each exam—same questions, but answer choices arranged randomly—first in the largest classes, then all of them. Since the author changed the anti-peeking regime during the study, a new dummy variable had to be incorporated into the model to describe the change in regime, and the author also had to rework the original model to incorporate peeking.

In the parts of the paper that follow, the change in the transformation function is described, the model is re-worked to incorporate peeking, and the results of the study are set out.

Change of the Score Transformation
In the first half of the study, extending from the spring of 2003 to the fall of 2004, the students were informed in the syllabus that raw score transformations were to be done in two steps:

1. If the mean score in terms of percentage of questions answered correctly is less than 77 percent, the percentage score of all test takers is increased by the difference of 77 percent and the mean score percentage.

2. If Step 1 results in a percentage value of the highest score greater than 100 percent, a linear transformation of the raw scores is made so that the mean is 77 percent and the maximum is 100 percent.

The purpose of the 100 percent threshold for application of Step 2 was to prevent a curve in which the lower-scoring students received less of an increase in their scores than the higher-scoring students.

In the second half of the study, extending from Spring 2005 to Fall 2006, the maximum allowed score in Step 2 was changed from 100 percent to 110 percent.

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

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

where h is the highest score, k is the value to which the highest score is curved, and $\bar{x}$ is the mean. In the first part of the study, k took the value of 100, in the second part, 110.

The Model with No Cheating

In the model with no cheating, test takers maximized utility functions whose arguments were test score and effort. The number of students taking the exam (class size) was assumed to be sufficiently large that effort on the margin by any one student had negligible effect on the mean score. The students took multiple choice exams in which guessing was not penalized, for which the utility function was assumed to take the form

$$U = U\left(y(x(n) + c(N-n)), n\right) \qquad (3)$$

where y is the score transformation, x is the number of items answered correctly, N is the number of items on the exam, n is the number of items attempted, and c is the average likelihood of a choice being correct. The expected raw percentage score of someone who turned in a completely randomized answer sheet would be c. The expected opportunity cost of attempting to answer a question would be c times the value of the question, so the net marginal product of attempting a question would be $(1 - c)$ times the value of the question.

The elasticized parameters for the two arguments of the utility function were defined as

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

All students were assumed to have a positive benefit, α , from higher test scores. Students for whom leisure is a normal good would have positive values of β ; students who enjoy the challenge of taking test questions would have negative values of β .
Utility maximization resulted in the first-order condition for *n*:

$$\alpha y'(x'(n) - c) = \beta. \qquad (5)$$

If we assume that the items on the exam are arranged in increasing order of difficulty so that the marginal product of effort is monotonically decreasing, a student's utility-maximizing effort is shown in Figure 1.

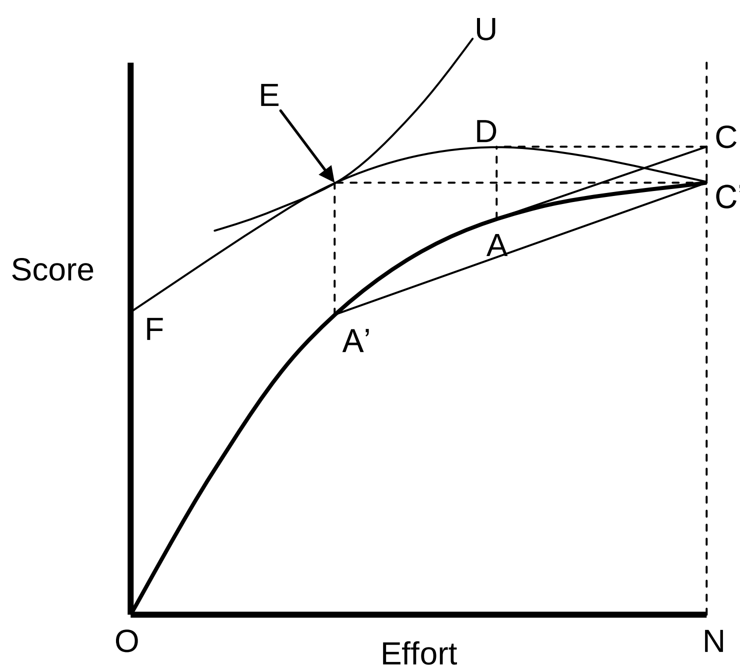

Figure 1: Production Function for a Multiple-Choice Exam, No Cheating

The marginal product of attempting the questions falls below the value of guessing at point A. The slope of AC is *c*. A student who initiated 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. This particular student with positive β would initiate guessing at point A', receive a raw score of C' which would translate to the tangency point E. Students with negative β would initiate guessing after passing point A and end up at a point to the right of point D. 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 would never guess.

Introducing Cheating to the Model

The effect of cheating on a question by looking at a neighbor's paper is an increase in productivity; in effect, it is theft of productivity. The marginal benefit of cheating on a question is the increase in productivity so obtained: what is the expected value of additional information from my knowing the answer on my neighbor's paper? The marginal cost is the expected impact of being seen cheating. Since cheating occurs, students assess this expected impact as being small. We will simplify the model by assuming that at the level of cheating done by students it is costless to them. Then no other changes need to be made to the model other than altering the production function.

We will assume that the cheater selects neighbors who will not have initiated guessing at the level of effort at which the cheater chooses to initiate guessing, because cheating confers no benefit on the margin if the neighbors have initiated guessing. The neighbors will have either a higher marginal productivity of effort than the cheater or a lower value of β .

Let us suppose that the exam begins with items to which the cheater knows the answers. Peeking confers no benefit for these because the expected value of the additional information is zero. As the difficulty of the items increases, the expected value of additional information becomes positive, but is bounded by the quality of the answers on the neighbors' papers. As the difficulty of the items increases, the quality of the answers on the neighbors' papers decreases until the point is reached where marginal return to effort of the neighbors falls to *c*. Beyond that point, the neighbors exhibit negative skill, but, paradoxically, negative skill conveys useful information to the cheater, so the expected value of additional information to the cheater increases again.

Figure 2 shows how cheating affects the production function. The original production function begins at the origin and goes through point A'. The cheater's production function begins at the origin and goes through point AA'. The dashed curved line parallel to the original production function and above it is the curve that the cheater's production function will intersect when the marginal return to effort of the neighbors has reached *c*. The non-cheater initiates guessing at point A' that transforms through point C' to tangency point E at utility U. The cheater initiates guessing at point AA' that transforms through point CC' to tangency point EE at utility UU. The utility contours are drawn here to show a student for whom the income effect of test scores is zero. Thus the change of effort shown is the substitution effect only. The cheater initiates guessing at a higher level of effort than the non-cheater because the opportunity cost of initiating guessing is higher for the cheater.

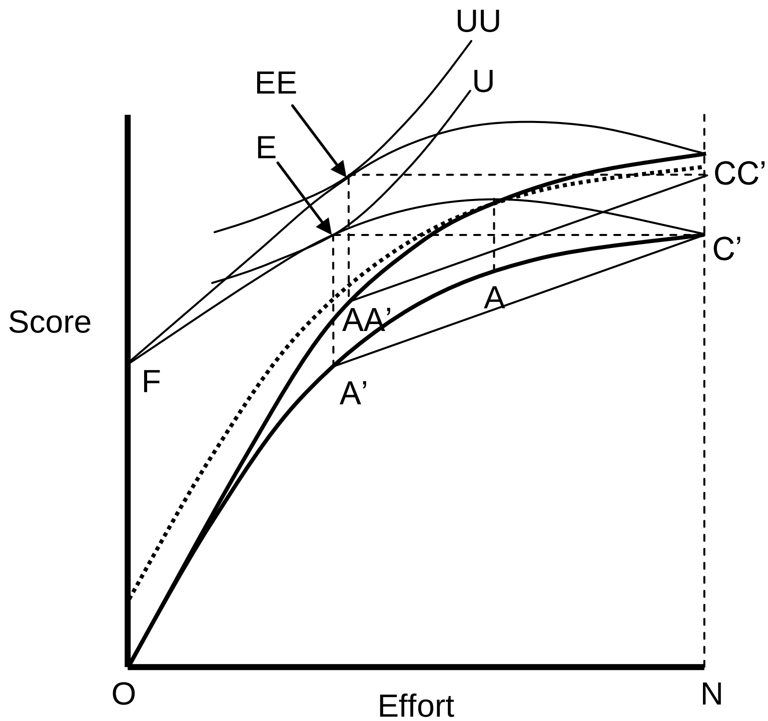

Figure 2: Substitution Effect of Cheating for a Multiple-Choice Exam

If the cheater's income elasticity for leisure is negative or weakly positive, the substitution effect is predominant and results in a net increase of effort. This is more likely to be the case if the student is a business major. If the cheater's income elasticity for leisure is strongly positive, the income effect is predominant, and results in a net decrease of effort. This is more likely to be the case if the student is not a business major. Thus an increase in the prevalence of cheating is likely to result in a divergence in performance of business majors and non-business majors, while reduction of cheating is likely to have the opposite effect.

Data and Methodology

The author's spring and fall classes taught in 2003 and 2004 constituted the first part of the study, and spring and fall classes taught in 2005 and 2006 constituted the second part of the study. Each part of the study had 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 percent of the grade and the final 40 percent. 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 3a and for the macroeconomics exams are shown in Figure 3b.

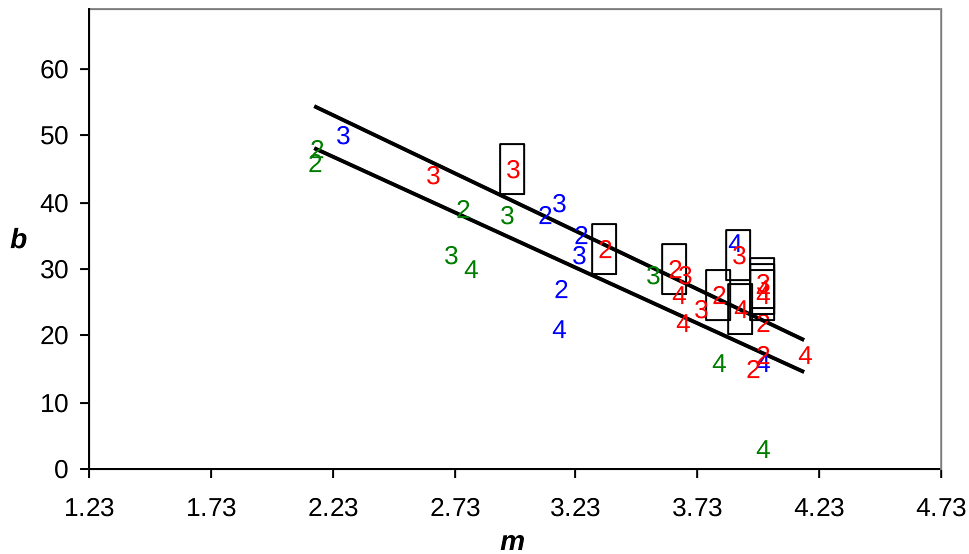

Figure 3a. Distribution of Curves for Microeconomics

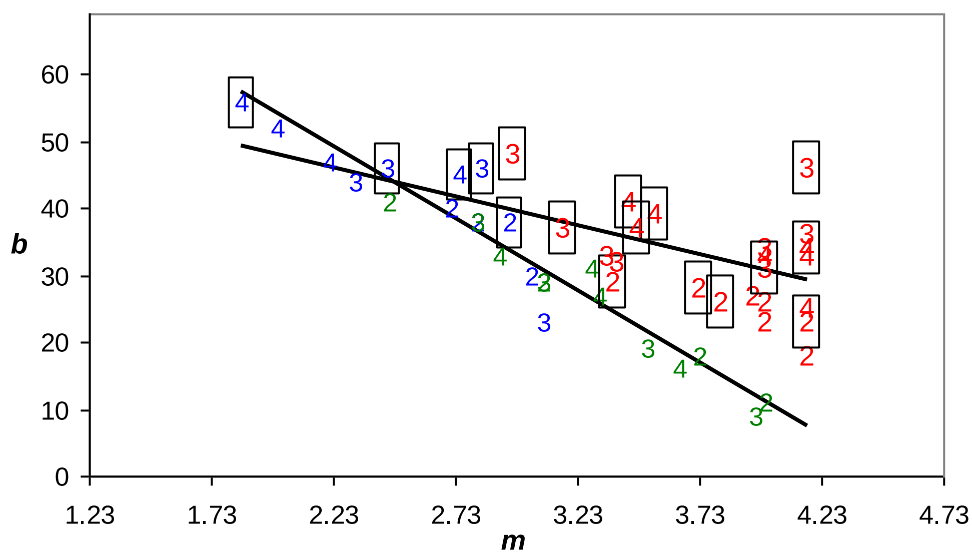

Figure 3b. Distribution of Curves for Macroeconomics

Each data point is marked with a numeral indicating whether it is the second, third, or fourth exam of the course. Data for the first exam were excluded because it covered substantially the same material in both courses, and caused a bias in favor of students taking their second course, as they were being exposed to the same material a second time. Numerals enclosed by rectangles indicate exams for which countermeasures against peeking were taken. The numerals in blue and green indicate the exams in the first half of the study and the numerals in red indicate the exams in the second half of the study. The green numerals represent exams given during the academic year 2003-2004, in which the instructor wrote his own examination questions. For exams at all other times the instructor used a database provided with the textbook.

The coordinates of the upper left corner of the charts (1.23, 69) were selected to represent the curve that would be given for an exam for a class of infinite size, for a value of *k* of 100, 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 equivalent coordinates for a value of *k* of 110 are (1.76, 66).

In each chart, the regression lines are for the two halves of the study. The slopes of the regression lines are consistent with the model's prediction of an inverse relationship between *b* and *m*, but the slopes are flattened somewhat because when Step 2 is not employed, *m* cannot be greater than 4.00 in the case of a 25-item exam or 4.17 in the case of exams for which one item was discarded. The students compensate by selecting a larger value of *b*. The centers of gravity for the data points shift to the right and downward in the second half of the study as would be expected as the students select larger values of *m* and smaller values of *b*.

The distributions of data points show that the effects of implementation of the no-peeking regime, the instructor preparing his own examination questions, and the non-use of Step 2 were not random, so dummy variables were introduced to deal with each of these in the regressions.

The Regressions

In Linn (2006), it was shown that if the student maximizes a utility functions whose arguments are test scores and leisure, the relationship of *b* to *m* is given by

$$-\ln(y-b) = -\ln(m) + C, \qquad (6)$$

where for students at the mean the value 77 is substituted for *y*. To test the hypothesis offered by Equation (6) the following regression was selected:

$$\begin{aligned} YBLN &= a_1 MLN + a_2 NOPEEK + a_3 K + a_4 BUSP \\ &+ a_5 SPR + a_6 OWNQ + a_7 HBUS + a_8 NOSTEP2 + C + \varepsilon \end{aligned} \qquad (7)$$

where *YBLN* is the natural logarithm of 77 less *b*, *MLN* is the natural logarithm of *m*, *NOPEEK* is a dummy variable whose value is 1 if countermeasures against peeking were taken for the exam and zero otherwise, *K* is a dummy variable whose value is 1 if *k* was 110 and zero otherwise, *BUSP* is the fraction of the class who were business majors at the time grades were issued, *SPR* is a dummy variable whose value is 1 for classes in the spring semester and zero otherwise, *OWNQ* is a dummy variable whose value is 1 when the instructor composed his own exam questions and zero otherwise, *HBUS* is a dummy variable whose value is 1 if at least one of the highest scorers was a business major and zero otherwise, and *NOSTEP2* is a dummy variable whose value is 1 if Step 2 was not used in the curve and zero otherwise. The statistical package available at Wessa (2007) was used for the statistical work.

For all the statistics with the exception of *NOPEEK* and *K* the one-tailed test was used. The appropriate signs for the tests are as follows:

For *MLN*, the sign is to be positive, as is clearly seen from Equation (6).

For *NOPEEK*, the sign is indeterminate. Peeking has an indeterminate result on effort, as has been seen, but it will increase the mean score and may also increase the highest score. In equation (2b), the value of *b* is inversely related to $\overline{x}$ but it is directly related to *h*. If the proportion of students in the class who peek is small, or the incentive of the high scorers to peek is high, peeking will contribute of a higher value of *b*. We should therefore see a larger difference of *y* and *b*, hence a lower value of *YBLN*. On the other hand, if the proportion of students in the class who peek is large, or the incentive of the high scorers to peek is small, peeking would result in lower values of *b* or a higher value of *YBLN*. Countermeasures to peeking will have the opposite effects on *YBLN* .

For *K* the sign is indeterminate. If, for the highest scorer, leisure is a normal good, students will expect a greater increase in *b* than would be the case if there was no income effect; if leisure is an inferior good for the highest scorer then there would be a lesser increase in *b*. A change in *k* should have no effect on student performance if there is not income effect for the high scorer, so positive income effect will result in a negative sign for the test and a negative income effect will result in a negative sign for the test.

For *BUSP* the sign is to be positive. Business majors are likely to invest more in the courses than non-business majors; an increased number of them will tend to raise the mean, which for a given value of *h* would show as lower values of *b*.

For *SPR* the sign is likely to be positive. In the spring semester, all of the students have had more test-taking experience and can be expected to perform better, resulting in lower values of *b*, except in the case—unlikely to happen—that the high scorers improve significantly more than the population in general.

For *OWNQ* the sign is to be positive. The data clearly indicate that the exam questions composed by the instructor tended to be less difficult than those in the database, hence, values of *b* would be expected to be lower.

For *HBUS* the sign is to be negative. Business majors are likely to invest more in the course than non-business majors; the value of *h* for a business major best scorer is likely to be higher than that of a non-business best scorer, so for a given mean, a business major high scorer will generate a smaller value of *m*, hence a larger value of *b*.

For *NOSTEP2* the sign is to be negative, as students select larger values of *b* when there is a ceiling on the value of *m*.

For microeconomics the results are

Variable	Parameter	S.E.	*t*	2 tailed p	1 tailed p
MLN	**1.205**	0.1402	8.60		**0.000**
NOPEEK	-**0.157**	0.0626	-2.50	**0.019**	
K	0.025	0.0574	0.43	0.671	
BUSP	-0.039	0.2383	-0.16		0.435
SPR	0.006	0.0416	0.15		0.440
OWNQ	**0.117**	0.0560	2.09		**0.023**
HBUS	0.046	0.0467	1.00		0.164
NOSTEP2	0.016	0.0540	0.29		0.387
C	**2.361**	0.1934	12.21		**0.000**

where the figures in bold type are significant at the five-percent level.

For macroeconomics the results are

Variable	Parameter	S.E.	*t*	2 tailed p	1 tailed p
MLN	**1.337**	0.1541	8.67		**0.000**
NOPEEK	-0.090	0.0537	-1.67	0.102	
K	**-0.170**	0.0682	-2.49	**0.017**	
BUSP	-0.004	0.2914	-0.01		0.494
SPR	0.046	0.0467	0.97		0.168
OWNQ	0.043	0.0675	0.64		0.263
HBUS	**-**0.011	0.0358	-0.31		0.378
NOSTEP2	**-0.167**	0.0542	-2.98		**0.003**
C	**2.291**	0.2001	11.55		**0.000**

where the figures in bold are significant at the five-percent level.

For neither macroeconomics or microeconomics did the proportion of business majors in the class, whether or not a business major was among the high scorers, or whether the class was taught in the spring or fall have a significant impact on the results. None of the other dummy variables had a significant impact on both macroeconomics and microeconomics; each had an impact only on one or the other.

In microeconomics, there was a negative relationship between *NOPEEK* and *YBLN*, so there was a positive relationship between the incidence of peeking and *YBLN*, which would be consistent, as before said, to a high proportion of students who peek or a low incentive of the high scorers to peek. In macroeconomics, the imposition of the no-peeking regime did not have a significant impact. The lack of significance in macroeconomics might be attributable to lower overall test scores and a lower marginal return to peeking.

The impact of changing *k* was significant for macroeconomics but not so for microeconomics. This implies, from what was before said, that high scorer moral hazard is indicated for macroeconomics but not for microeconomics. High scorer moral hazard is associated with a preference for leisure--or its equivalent, an aversion to effort. As the results seem not to indicate a difference between business majors and non-business majors in their effect on the statistics, the difference in results between microeconomics and macroeconomics may be laid to greater difficulty in learning the materials for the latter.

The instructor's composing his own exam questions resulted in a significant improvement of student scores in microeconomics, but not in macroeconomics. This is likely due to a greater difficulty of creating plausible wrong answers in microeconomics than in macroeconomics.

The effect of imposing an upper limit on m was significant in macroeconomics but not in microeconomics. The number of instances in which Step 2 was not employed increased much more for macroeconomics than microeconomics after k was increased to 110. Thus the high scorers in macroeconomics were exposed to more instances in which their substitution effect was limited; hence, in an atmosphere of positive preferences for leisure, the high scorers' income effects would overwhelm their substitution effect, resulting in lower highest scores. A regression of raw highest scores against the same set of dummy variables used in Equation (7) was performed, and it confirmed that in macroeconomics, *NOSTEP2* significantly lowered highest scores, while such was not the case for microeconomics.

Effect of the Curve Change on Transformed Score Variances

For a grading system based on student rank to have meaning, the standard deviation of the scores should fit with the grading system used (in this case, ten-point intervals for each letter grade, each interval one standard deviation). The following tables show the standard deviation of the microeconomics and macroeconomics weighted average scores (four exams and a final) for each half of the study. Assuming no change in the covariance of particular student scores on exams, the desired standard deviation of the average scores can be attained by manipulation of the standard deviations of the specific examination scores, as has been attempted here with a change in k. The following two tables show the results for the change in k for microeconomics and macroeconomics:

For microeconomics:

Regime	Number of Scores	Standard Deviation
k = 100	245	8.94
k = 110	286	10.78

And for macroeconomics:

Regime	Number of Scores	Standard Deviation
k = 100	354	9.35
k = 110	313	9.20

An increase in k increased the standard deviation of the scores in microeconomics but did not do so in macroeconomics. A change in k alone appears to be sufficient in bringing about a standard deviation of the desired size in microeconomics, but in macroeconomics a change of k should be supplemented with a change that increases the substitution effect. In microeconomics, a linear interpolation of the observed standard deviations suggests that a k

value of about 106 would bring about the desired standard deviation. In macroeconomics, the standard deviation is too small, but would need only two-thirds of the magnitude of the change needed for microeconomics. A value of *k* of 104 would do the job in macroeconomics if an offsetting change is made in the substitution effect. A simple way of boosting the substitution effect is to reduce the number of items on the examination without reducing the size of the database from which they are drawn; in effect increasing the marginal benefit of effort for each item.

Conclusion and Applications

This study tested for the existence of moral hazard among the highest scorers under a regime that rewarded them less than lower-scoring students when the raw scores were "curved". The remedy tested was to increase the value to which the highest scores could be curved from 100 to 110, subject to a ceiling on the substitution effect. The results showed that for microeconomics, the incidence of high scorer moral hazard was low. In macroeconomics, the incidence of moral hazard was higher, likely due to the material being more difficult to learn, but the effectiveness of the remedy to reduce moral hazard was limited because of the ceiling on the substitution effect, and in fact reduced highest scores because of an overgenerous income effect.

The study clearly shows that the curving regimes in the two courses should be different. Microeconomics highest scores should be curved to approximately 106. In macroeconomics, raising the curve for the highest scores confers no benefit unless the marginal return to effort per item is increased. The deficiency in standard deviation to be overcome in macroeconomics is about two-thirds of that in microeconomics, so a macroeconomics highest score is 104 is likely the best.

References

Linn, J. "Test Score Compensation, Moral Hazard, and Student Reaction to High Scorer Disincentives in Economics Principles", MBAA, Chicago, Ill. 16 Mar 2006.

Wessa, P. (2007). Free Statistics Software, Office for Research Development and Education, version 1.1.21, URL http://www.wessa.net/