Do students behave rationally in multiple-choice tests? Evidence from a field experiment.*

María Paz Espinosa and Javier Gardeazabal
University of the Basque Country†

22nd December 2005

Abstract

A disadvantage of multiple-choice tests is that students have incentives to guess. To discourage guessing, it is common to use scoring rules that either penalize wrong answers or reward omissions. These scoring rules are considered equivalent in psychometrics, although experimental evidence has not always been consistent with this claim. We model students' decisions and show, first, that equivalence holds only under risk neutrality and, second, that the two rules can be modified so that they become equivalent even under risk aversion. This paper presents the results of a field experiment in which we analyze the decisions of subjects taking multiple-choice exams. The evidence suggests that differences between scoring rules are due to risk aversion as theory predicts. We also find that the number of omitted items depends on the scoring rule, knowledge, gender and other covariates.

Keywords: field experiment, risk aversion, scoring rules, multiple-choice tests.

JEL: A200, C930, D800.

*We thank Kathrin Pokorny for her useful suggestions concerning an earlier version and participants in the ESA meeting at Montreal for their comments. We are indebted to Amagoia Sagasta and Arantza Ugidos for their help in conducting the experiment and to Susana Márquez for her assistance. Financial support from Ministerio de Ciencia y Tecnología (BEC2003-02084, SEC2003-04826) and Universidad del País Vasco (9/UPV 00035.321-13560/2001, 13511/2001) is gratefully acknowledged.

†Departamento de Fundamentos del Análisis Económico II. Avenida Lehendakari Aguirre 83, 48015 Bilbao, Spain. E-mail: mariapaz.espinosa@ehu.es, javier.gardeazabal@ehu.es.

1. Introduction

Multiple-choice tests are widely used as an evaluation tool.[1] Their main advantages over constructed-response tests are that their use guarantees a wider sampling of the content and precludes error of measurement introduced by the grader. Their main disadvantage is that multiple-choice tests may encourage guessing, which adds an error term to test scores and lowers test reliability in measuring students' knowledge.[2] This is the case when the test score is the number of right answers, hereafter S_1. When students are evaluated with the number-right scoring rule, they will of course answer all questions whether they know the answer or not. Thus, the score includes an error component coming from those questions in which a student gets the right answer by chance. To mitigate this problem, examiners quite often use a formula scoring rule that penalizes wrong answers (S_2) and is intended to reduce guessing behavior. Although it is used only on rare occasions, an alternative way of discouraging guessing is to reward omitted questions (S_3).

In the psychometric literature, scoring rules incorporating a correction for guessing in the form of a penalty for wrong answers (S_2) and a reward for omitted questions (S_3) are considered equivalent since one can be rewritten as a linear function of the other.[3] However, empirical evidence indicates that they may yield different results. These differences in students' behavior were thought to be associated to different framing and attributed to psychological factors.[4] An explanation of this non-equivalence result was advanced by Budescu and Bar-Hillel (1993) appealing to the different consideration that opportunity costs (failure to win) and out-of-pocket costs (paying a penalty) have for individuals. According to this view, examinees should guess more when they are rewarded for omissions than when penalized for wrong answers, given that it is easier to forgo a gain than to incur a loss. More recently this idea has been formalized using prospect theory by Bereby-Meyer, Meyer and Flascher (2002).

In this paper we model students' decisions as the choice between different compound lotteries to maximize expected utility and show several results related to the conditions for the equivalence of scoring rules. First, when examinees are risk averse the two scoring rules make them face a different trade-off when deciding whether to answer a question or not; therefore, expected utility maximizers could behave differently under S_3 than under S_2 even though they are linearly related. Second, we demonstrate that the two scoring rules are equivalent under risk neutrality. Third, we also show that when these two scoring methods are conveniently normalized they become strategically equivalent even under risk aversion.

These results are relevant for the design of experiments in psychometrics seeking to determine the effects of the scoring rule on the test' validity and reliability and, in particular,

[1] See Siegfried (1996) and Bredon (2003) to grasp the importance of the use of multiple choice tests in economics.

[2] See Walstad and Becker (1994), Heck and Stout (1998), Becker and Johnston (1999), Chan and Kennedy (2002), for a comparison of essay and multiple-choice tests in economics.

[3] We will show this relationship in Section 2.

[4] See Traub, Hambleton and Singh (1969), Traub and Hambleton (1972) and Waters and Waters (1971).

the impact of psychological factors. If examiners use the usual scoring rules, differences in observed behavior between penalty for wrongs and reward for omits could be due to psychological factors or risk aversion, that is, the impact of the two variables cannot be distinguished. However, if the scoring rules faced by subjects in such experiments are normalized so that they are in fact equivalent, rational expected utility maximizers should behave identically. If experimentally they do not, differences should then be attributed to irrationality.

To determine whether in this context risk aversion may be significant enough to give rise to the observed differences in behavior, we designed an experiment using scoring rules that incorporate the normalization proposed and the usual correction formula. In a regular undergraduate Macroeconomics course, it was made public that exams would be graded with different scoring rules for different groups of students, so that all students knew well in advance which rule they would face in the exam.[5] Then we looked at the number of omissions. The results of our field experiment indicate that once the scoring rules are normalized, there are no differences in behavior, while some differences are observed when comparing the outcomes with those of the usual scoring rule. Therefore, the results are consistent with rational behavior of students. This evidence suggests that individuals were not affected by the different framing of the scoring rules; when the scoring rules were strategically equivalent subjects adopted similar decisions. Thus, psychological factors did not seem to play a role in students' behavior. Note that subjects' decisions determined their grade on the course, so there was a strong incentive to take the right decisions.[6] In summary, the evidence is consistent with expected utility maximization and supports the hypothesis that differences in behavior might be due to risk aversion rather than psychological factors.[7] We also find that gender, the student's knowledge and the difficulty of the exam are important determinants of the number of omissions.

The rest of the paper is organized as follows. Section 2 lays out the preliminaries. Section 3 establishes the theoretical results. Section 4 describes the design of the experiment. Section 5 reports the experimental results and Mann-Whitney tests for the null hypothesis that students' behavior is identical under both types of rules. In Section 6 we use econometric models of count data to explain the number of omissions in the exams, controlling for knowledge, gender and other covariates. Section 7 concludes.

2. Preliminaries

Let N be the number of items in an exam and M the number of alternatives, one correct and $M-1$ incorrect. A student is defined by her level of knowledge and a function, $u(s)$,

[5]The experimental design, described in section 4, guaranteed an equal treatment of students.

[6]In a companion paper, Espinosa and Gardeazabal (2005), we report the results of other experiments with weaker incentives.

[7]There are numerous laboratory experiments documenting different types of deviations from rationality, but field experiments are scarce. See Bertrand, Karlan, Mullainathan, Shafir and Zinman (2005), Haan, Los, Riyanto and Van Geest (2002), Haigh and List (2005) and List and Millimet (2005) for notable exceptions.

representing her valuation of the score, s, obtained in the exam. We assume this valuation is such that $u'(s) \geqslant 0$.[8] Students may have different preferences and different levels of knowledge.

The simplest scoring rule is *number right*, denoted S_1, where the score is just the number of right answers r:

$$S_1 = r.$$

Some scoring rules incorporate a correction for guessing feature. Typically, there is a penalty of $\frac{1}{M-1}$ points for each incorrect answer. If a student does not know the answer to a question and considers all alternatives as equally likely, the expected value of answering is zero, exactly what the student gets from omitting the item. This scoring rule yields a final score:

$$S_2 = r - \frac{w}{M-1}$$

where r and w are the number of right answers and wrong answers, respectively. An alternative rule to discourage guessing is to give $\frac{1}{M}$ points for each omitted question. If a student does not know the answer to a question and considers all alternatives as equally likely, the expected value of answering is $\frac{1}{M}$, which is equal to the reward for omissions. This scoring method yields a final score:

$$S_3 = r + \frac{o}{M}$$

where o is the number of questions omitted. The reward for omissions and the penalty for wrong answers are intended to induce the same behavior in students: to discourage guessing when the student does not know the answer. Note that S_2 and S_3 are linearly related as

$$S_3 = \frac{N}{M} + \frac{M-1}{M} S_2. \tag{1}$$

Item i can be viewed as a gamble in which a student has probability q_i of getting the right answer, where the probability depends on the knowledge of the student and the difficulty of the item. It is reasonable to assume that the probability of getting the right answer is increasing with the knowledge of the student and decreasing with the degree of difficulty of the item. The probability of failing the item is $1-q_i$. Assume the student answers only item i, leaving $N-1$ items unanswered. There are two possible events: she may get the answer right or fail. Denote $s(r,w,o)$ the score obtained from r rights, w wrongs and o omissions. Let $\ell\{i\}$ denote the lottery induced by answering only item i,

$$\ell\{i\} = [q_i \circ s(1,0,N-1) \oplus (1-q_i) \circ s(0,1,N-1)].$$

Note that the scoring rule affects the values of the score, $s(1,0,N-1)$ and $s(0,1,N-1)$ but the probability q_i is rule-independent. Let $U(\ell\{i\})$ denote the utility derived from this

[8] This assumption does not exclude pass-fail exams in which $u' = 0$ until the pass score is reached.

lottery,

$$U(\ell\{i\}) = q_i u(s(1,0,N-1)) + (1-q_i)u(s(0,1,N-1)).$$

If the student answers items i and j, leaving $N-2$ items unanswered, she gets $s(2,0,N-2)$ when the two items are correctly answered, $s(1,1,N-2)$ when she gets one item wrong and $s(0,2,N-2)$ when she gets both wrong.[9] The payoffs and probabilities are given in the following table

Score and Probability

Score	Probability
$s(2,0,N-2)$	$q_i q_j$
$s(1,1,N-2)$	$q_i(1-q_j)+(1-q_i)q_j$
$s(0,2,N-2)$	$(1-q_j)(1-q_i)$

where q_i and q_j are the probabilities of getting the right answer to items i and j respectively. This lottery may be written as

$$\ell\{i,j\} = q_i \circ [q_j \circ s(2,0,N-2) \oplus (1-q_j) \circ s(1,1,N-2)] \oplus (1-q_i) \circ [q_j \circ s(1,1,N-2) \oplus (1-q_j) \circ s(0,2,N-2)].$$

The utility derived from this lottery is

$$U(\ell\{i,j\}) = q_i q_j u(s(2,0,N-2)) + (q_i(1-q_j)+(1-q_i)q_j)u(s(1,1,N-2)) + (1-q_i)(1-q_j)u(s(0,2,N-2)).$$

Similarly, in an exam with N items any subset of items is also a compound lottery. Let $\mathcal{L}(N)$ be the set of all compound lotteries in an exam with N items including a degenerate lottery, denoted by $\ell\{0\}$, which corresponds to omitting all items. For example, in an exam with two items, the set of all compound lotteries is

$$\mathcal{L}(N) = \{\ell\{0\}, \ell\{1\}, \ell\{2\}, \ell\{1,2\}\},$$

that is, the degenerate lottery which corresponds to omitting all items, the lottery consisting of answering only the first item, the lottery corresponding to answering only the second item and the lottery where the student answers both items.

A student chooses the best compound lottery in $\mathcal{L}(N)$. Formally, she maximizes the expected utility over the set of all possible compound lotteries

$$\begin{aligned} &\max U(\ell) \\ &\ell \in \mathcal{L}(N) \end{aligned} \tag{2}$$

This problem summarizes our assumptions on students' behavior. Bernardo (1998) and

[9] Without loss of generality we assume that all items yield the same payoffs.

Burgos (2004) are the only other attempts that we know of at modeling students behavior in multiple-choice test. Bernardo (1998) suggests that students maximize the score or minimize the probability of failing the exam. Burgos (2004) uses prospect theory to postulate a utility function that assigns different values to loses and gains. However, neither of these authors tested their assumptions empirically.

To give specific content to this gambling interpretation of items, let us consider the scoring rule that penalizes for wrongs, S_2. Under this scoring rule answering only item i is a lottery:

$$\left[q_i \circ 1 \oplus (1-q_i) \circ \left(-\frac{1}{M-1}\right)\right].$$

In this case, with probability q_i the student gets a score of one point and with probability $1-q_i$ she incurs the penalty $\frac{1}{M-1}$.

The reward for omissions scoring rule has a similar interpretation. Under S_3 the student has a sure payoff $\frac{N}{M}$ if she does not answer any item, that is, if she chooses the lottery $\ell\{0\}$. When the student answers item i she losses the sure reward for omitting $\frac{1}{M}$. If she gets the right answer she gains a score of 1 and if she fails she gets zero. Thus, answering only item i is a lottery

$$\left[q_i \circ \left(1+\frac{N-1}{M}\right) \oplus (1-q_i) \circ \left(\frac{N-1}{M}\right)\right].$$

Definition: Two scoring rules are *strategically equivalent* if they always induce the same behavior in a rational exam taker (an expected utility maximizer).

In this paper we check whether S_2 and S_3 are equivalent or not under different risk preferences and establish various results on the strategic equivalence of scoring rules.

3. Theoretical results

This paper contributes three results. The first is that, contrary to what is generally accepted in the educational measurement literature, penalizing wrongs and rewarding omits, as defined in the previous section, are not in general strategically equivalent.

Proposition 1: S_2 and S_3 are not strategically equivalent for risk averse exam takers.

To show this, it is sufficient to find an example where a rational exam taker would behave differently under S_2 than under S_3. Consider an exam with one item, $N=1$, two alternatives, $M=2$, a student with valuation $u(s)=\sqrt{a+s}$, with $a \geq N$, and probability q_i of getting the right answer. Answering this exam is a lottery $l_2 = [q_i \circ 1 \oplus +(1-q_i) \circ (-1)]$ while omitting yields a sure payoff of zero under S_2. The same exam under S_3 can be written as $l_3 = [q_i \circ 1 + (1-q_i) \circ 0]$ while omitting yields a sure payoff of $\frac{1}{2}$. Under S_2, the expected utility from answering is lower than that from omitting if $q_i\sqrt{a+1} + (1-q_i)\sqrt{a-1} < \sqrt{a}$. However, under S_3 a rational exam taker finds a higher expected utility from answering if $q_i\sqrt{a+1} + (1-q_i)\sqrt{a} > \sqrt{a+\frac{1}{2}}$. It is easy to verify that for $a=1$ and $q_i = 0.6$ both

inequalities hold. Facing this exam, the student would omit under S_2 and answer under S_3. This shows that the two scoring rules are not in general strategically equivalent.

Our second result states that for a particular type of risk attitude, the two scoring rules become strategically equivalent.

Proposition 2: For a risk neutral examinee, S_2 and S_3 are strategically equivalent.

To show this, it is necessary to prove that a risk neutral student would always make the same decision under either of the scoring rules. Under S_2, a risk neutral student would choose lottery ℓ whenever its expected payoff is at least as high as that of any other lottery, that is

$$\sum_i q_i^{\ell} s_{2i}^{\ell} \geq \sum_i q_i^{\ell'} s_{2i}^{\ell'} \tag{3}$$

for all $\ell' \in \mathcal{L}(N)$, where s_{2i}^{ℓ} are the scores under S_2 of all possible outcome in lottery ℓ and q_i^{ℓ} are the associated probabilities of those outcomes, so that $\sum_i q_i^{\ell} = 1$. Condition (3) is equivalent to

$$\sum_i q_i^{\ell} \left(\frac{N}{M} + \frac{M-1}{M} s_{2i}^{\ell} \right) \geq \sum_i q_i^{\ell'} \left(\frac{N}{M} + \frac{M-1}{M} s_{2i}^{\ell'} \right),$$

for all $\ell' \in \mathcal{L}(N)$. Using (1) we can write

$$\sum_i q_i^{\ell} s_{3i}^{\ell} \geq \sum_i q_i^{\ell'} s_{3i}^{\ell'},$$

for all $\ell' \in \mathcal{L}(N)$. In words, the student would choose lottery ℓ in the set $\mathcal{L}(N)$ under S_2 if and only if she would choose it under S_3. This completes the proof of equivalence.

Scoring rules S_2 and S_3 are equivalent for risk neutral students. However, experiments seem to indicate that students do not always behave identically under the two scoring rules. Our point is that risk preferences may have been dismissed as "other psychological factors" in the experiments designed to evaluate S_2 and S_3.

In this paper we propose a normalization of the scoring rules S_2 and S_3, denoted S_2^* and S_3^*. These normalized rules are as follows

$$S_2^* = \left(\frac{r - pw}{N} + p \right) \frac{1}{1+p},$$

$$S_3^* = \frac{1}{N} \left(r + \frac{p}{1+p} o \right),$$

where $p = \frac{1}{M-1}$ is the penalty. S_2^* and S_3^* lie in the unit interval, whereas S_2 may take negative values. Notice that S_2^* and S_3^* assign a positive score to examinees leaving all questions unanswered.

Proposition 3: S_2^* and S_3^* are strategically equivalent for all types of risk preferences.

For S_2^* and S_3^* to be strategically equivalent the student should choose the same set of items to answer under the two scoring rules, i.e. the solution to problem (2) is the same. For

that we simply have to show that the set $\mathcal{L}(N)$ is identical under S_2^* and S_3^*. We do this in two steps. First, notice that the probabilities of the different events are independent of the scoring rule. Second, since the number of items in the exam is equal to rights plus wrongs plus omits we have that

$$S_2^* = \left(\frac{r - pw}{N} + p\right)\frac{1}{1+p} = \left(\frac{r - p(N - r - o)}{N} + p\right)\frac{1}{1+p}$$

$$= \frac{(1+p)r + po}{N(1+p)} = \frac{r + \frac{p}{1+p}o}{N} = S_3^*,$$

so payoffs are also identical under both scoring rules. This completes the proof of equivalence.

The normalized scoring rules S_2^* and S_3^* would allow the experimenter to isolate the effect of psychological factors from that of risk preferences since they induce the same behavior in rational students. If students do not behave identically in the experiment, then differences in behavior are due to psychological factors.

4. Experimental design and procedures

The objective of our experiment is threefold. First, we test whether students behave differently with the standard scoring rules, so that our results are comparable to previous findings. Second, we compare the results when students face normalized scoring rules, to determine whether differences observed in behavior are due to attitudes towards risk or other factors. Third, we also try to determine what variables influence students' decision to omit items.

We chose to conduct a field experiment by grading students using different scoring rules in the exams of a regular course, so that the subject pool would not introduce any bias into the results. Furthermore, the payoff in terms of grade is particularly appropriate in this case for two reasons. First, grades generate stronger incentives than small amounts of money. Second, attitudes towards risk concerning grades may be different from behavior towards risk when money is involved.

The experiment was conducted at the University of the Basque Country, Spain. The salient features of the experiment are shown in Table 1. Subjects were second year undergraduate students pursuing a B.A. degree in Economics, enrolled for Intermediate Macroeconomics II in the Spring of 2005. The students' performance on the course was evaluated using five multiple-choice tests, each accounting for 20% of the total grade. Each exam covered the material in one chapter, except the first exam, which covered the first two (shorter) chapters. Each exam had ten items and each item had four possible answers, one correct and three incorrect.

The experiment had three treatments: penalty for incorrect answers, S_2, normalized penalty for incorrect answers, S_2^*, and normalized reward for omissions, S_3^*. Given the parameters of the exams these rules were as follows: $S_2 = 2r - \frac{2}{3}w$, $S_2^* = 5 + \frac{3}{2}r - \frac{1}{2}w$ and $S_3^* = 2r + \frac{1}{2}o$.

Notice that for any given student and exam, the expected score is lower with scoring rule S_2 than with scoring rules S_2^* and S_3^*. This difference in expected values precludes certain types of experiment designs. For instance, the most natural design would be to split subjects into three treatments, but this design would favor students in treatments with normalized scores. To preserve the student's right to be treated equally in the course grading we used the following design. At the beginning of the course, students were randomly assigned to three groups -Blue, Yellow and White- with 62, 62 and 61 students respectively. Students were told to which group they had been assigned and that each group was going to be evaluated with a different scoring rule in each exam according to the design in Table 2. In each of the first three exams we used penalty for incorrect answers, normalized penalty for incorrect answers and normalized reward for omissions; the last two exams were graded using number-right. Thus, all students were evaluated with S_2, S_2^*, and S_3^* once each and twice with S_1. These rules were posted on the course website along with other useful information before the beginning of the course, so that students were well aware of scoring methods.

At the beginning of the course 185 students were enrolled. The number of students taking exams decreased during the course: 177 students took the first exam, 169 the second and 162 the third. After each exam, students were told the score obtained. Before the first exam, all students had a score of zero, while after the first exam their scores were different, reflecting differences in knowledge, luck and the differential effect of the scoring rules. In the second exam, students with different scores might have had different behavior towards omission even when faced with the same scoring rule and more so if they faced different ones. Therefore, the results of the first exam are the only ones not affected by the scores in past exams.

Notice that the order in which students face scoring rules could also be a relevant factor in their behavior, because the score in previous exams could matter and this depends on the scoring rule. Since the order in which students face the scoring rules is determined by the group, the group could potentially influence the observed behavior of students. Table 3 summarizes group characteristics. Even though group assignment was random, there are some differences in the average knowledge of the students and the proportion of exams with no omissions.

Furthermore, the fact that there was a passing grade (55%) has some implications for the design of the experiment. As we approach the end of the course, students with low accumulated scores have less incentive to omit, as a large number of omissions does not guarantee a passing grade. That is why we used the number right scoring rule in the last two exams.

Appendix A contains the instructions given in all exams, which included a set of general instructions and a treatment-specific instruction regarding the scoring rule. In the educational measurement literature it is common practice to give students advice regarding omissions. However, we did not give students any such advice. When using penalty scoring, it is typically recommended that examiners advise students not to omit in case they may

rule out one or more answers as incorrect. The idea is that students should answer when they have partial knowledge and the expected value of answering is positive. However, a risk averse student may optimally decide not to respond to an item with positive expected value. Unless all students were risk neutral, no good general advice could be provided, as the students' optimal behavior depends on their degree of risk aversion.

5. Experimental results

Under number right scoring, rational students ought to respond all items, and this is what happened in the last two exams where no student omitted a single item. Thus, the analysis is restricted to the first three exams. All statistics reported correspond to the variable *number of omitted items*. The statistical analysis would be equivalent had we used the number of answered items, in particular the Mann-Whitney test statistics reported below would be numerically identical up to a sign transformation. Table 4 reports basic descriptive statistics for the first three exams. The statistical analysis is restricted to the 160 students who took the first three exams. Unfortunately, even though we designed the experiment with equal initial group sizes, due to drop outs the number of students in each group varies from 49 students in the Blue group to 57 students in the Yellow group. The average number of omissions varies between exams and scoring rules. Some students answered all questions, no matter what scoring rule they faced. On the other hand, no student omitted all questions in any exam, because the passing grade (55%) required some answers.

Table 5 reports the Mann-Whitney test statistics for the null hypothesis that the samples come from the same population. According to these results we can reject equality of distributions between penalty and normalized reward in the first and third exams and between penalty and normalized penalty in the third exam. We cannot reject the hypothesis of equality of distributions between normalized reward and normalized penalty.

The results are broadly consistent with the theoretical predictions. Rejection of the null of equality of distributions between penalty and normalized reward and penalty and normalized penalty scoring rules is consistent with the behavior of risk averse expected utility maximizers (see Proposition 1). Furthermore, the fact that we cannot reject the null of equality between normalized penalty and normalized reward is also consistent with the results in Proposition 3. The two scoring rules are strategically equivalent and therefore should induce the same behavior in rational students regardless of their attitudes toward risk.

To sum up, the empirical results are consistent with rational subjects: when facing rules that are strategically equivalent we never reject the hypothesis of equality of distributions, but when facing rules that are equivalent under risk neutrality but may induce different behavior under risk aversion (or risk loving) the distributions are different in half of the cases and similar in the other half.

5.1. Should covariates be controlled for?

In this subsection we check whether the conclusions reached in the light of the tests of equality of distributions are valid or, on the contrary, we need to control for other covariates.

The statistics in Table 4 indicate that results vary from exam to exam. The largest average number of omissions corresponds to normalized reward in the first exam, normalized penalty in the second and penalty in the third. These differences across exams are also found in Table 5: omissions tend to be significantly higher under normalized reward than penalty in the first exam and the opposite is true in the third exam. Further evidence that results are exam-dependent is given in Table 6, where we report the Mann-Whitney test statistics to test the null hypothesis of equality of distributions between exams for a given scoring rule. The results are clear: in seven out of nine cases we reject the null of equality of distributions. Exams are not independent from one another. Rather, each one is part of a sequence of tests for students' evaluation during the course.

Notice that the behavior of rational students should depend on their accumulated score. To see why this should be the case, note that after the first exam the grades are revealed so that for each student this realization of the score is incorporated into her objective function. This may affect the way the compound lotteries in the second and third exams are evaluated.

According to this, in the second and third exams differences in students' behavior could depend not only on the rule used but also on their previous scores. Therefore, the first exam is the only one where the only difference among groups is the scoring rule used and the tests of equality of distributions are properly applied. The following section applies formal econometric procedures to take into account the effect on students' behavior of several factors, including the score obtained in previous exams.

6. Econometric analysis

As posited in the discussion above, to go further in our analysis of the data we need to control for several factors which may influence omission behavior. Our dependent variable, the number of omissions will be explained as a function of variables which we now explain:

1. As argued above, the *accumulated score* in past exams on the course should influence students' behavior. A student with a high accumulated score might decide to omit more items than a student with a low accumulated score. To investigate this possibility we include in the regression the accumulated score, which is set to zero in the first exam, to the score obtained in the first exam at the second exam, and to the sum of the scores in the first and second exams at the third one.

2. *Knowledge*, and in particular knowledge of Macroeconomics, should determine the behavior of students. All else equal, a student with greater knowledge should omit less than a less knowledgeable student. In the regressions reported below we include a proxy for knowledge of the subject: the grade obtained in Intermediate Macroeco-

nomics I either in the previous semester (Fall 2004) or in a previous year.[10]

3. The difficulty of the *exam* should definitely influence the number of omissions in one exam. For a given set of students, a more difficult exam ought to be reflected in a higher number of omissions. Even though we tried to write exams of ex-ante similar difficulty, it could be the case that exams had different degrees of difficulty. To account for this possibility in the regression, we include a set of dummy variables indicating the exam (first, second or third) which capture unobserved characteristics of exams, constant for all individuals.

4. Some studies have shown a link between risk attitudes and *gender* (see for example Byrness, Miller and Schafer, 1999 and Cadsby and Maynes, 2005). Since scoring rules S_2^* and S_3^* are not equivalent for risk averse subjects, gender might effect the number of omissions. Furthermore, it has been documented that instructions concerning guessing behavior may affect gender-related differences in multiple-choice test scores, see Prieto and Delgado (1999). To account for gender differences we include a gender dummy variable.

5. The main proposition of this paper is that the *scoring rule* used in an exam is an important factor that can affect students' attitude towards omission. In the regressions we include a set of dummy variables capturing the rules used.

6. Students are distributed into five *sections* with three different instructors. Different teaching expertise could induce differences in the performance of students. To control for any such differences we include dummy variables for sections.

7. Even though assignment to *groups* (Blue, Yellow and White) was random, we control for unobservable differences in group characteristics by including a set of group dummies. Table 3 reports observable group characteristics such as gender distribution and average knowledge. Gender is evenly distributed among groups, while there are small differences in average knowledge and also in the proportion of exams with no omissions.

8. We also include interactions and the continuous variables squared. This specification can be interpreted as a quadratic approximation to a nonlinear regression equation.

The dependent variable of our analysis, the number of omissions, can be classified as count data. To analyze the role of the explanatory variables we have made use of four different regression models for count data: Poisson, Negative Binomial (NB), Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB). Omissions have a Poisson distribution

[10]Three students did not have a grade for Intermediate Macroeconomics I, which further restricted our sample to 157 students.

with mean λ if the probability of z omissions is

$$P(z) = \begin{cases} \frac{e^{-\lambda}\lambda^z}{z!} & for \quad z = 0, 1, 2, ... \\ 0 & otherwise \end{cases}$$

The Poisson regression model assumes that the parameter λ depends on a vector X of explanatory variables exponentially, that is,

$$\lambda = e^{X'\beta}, \tag{4}$$

where β is a vector of parameters. Even though the Poisson model can describe many empirical situations, there are cases which it does not fit well. A property of the Poisson model is that the mean and the variance of omissions are equal to λ. As shown in Table 4, the number of omissions may not satisfy this property of the Poisson model. To allow for different mean and variance of omissions we fit the NB model where

$$\lambda = e^{X'\beta+u} \tag{5}$$

and u follows a Gamma distribution with parameters $(1/\alpha, \alpha)$. Under these assumptions, the mean number of omissions is λ and the variance of omissions is $(1 + \alpha)\lambda$. If $\alpha = 0$ the NB model boils down to the Poisson model.

A further criticism of the Poisson model is that it typically predicts a lower number of zeros than observed in the data. In our data set, one third of the exams had zero omissions. To account for this characteristic of the experimental data we fit the ZIP regression model. A simple interpretation is that this model classifies individuals as "omitters" or "non omitters" using a logit model. Data from "omitters" are used to predict the number of omissions for each individual (which can be zero or a positive integer lower than the number of items) as in the standard Poisson model. Therefore, the probability of observing an individual omitting z items is

$$P(z) = \begin{cases} \omega + (1-\omega)e^{-\lambda} & if \quad z = 0 \\ (1-\omega)\frac{e^{-\lambda}\lambda^z}{z!} & if \quad z = 1, 2, ... \\ 0 & otherwise \end{cases}$$

where the parameter λ is expressed as a function of a set of explanatory variables as in (4) and

$$\omega = \frac{e^{Z'\gamma}}{1 + e^{Z'\gamma}},$$

where Z is a vector of variables that determine whether an observation corresponds to the group of "omitters" or "non-omitters" and γ is a vector of parameters. The Poisson distribution is a particular case of a ZIP distribution when $\omega = 0$. The ZINB is a generalization of the ZIP model with λ as in (5).

For the empirical analysis we pool the observations of all the exams and use the set of variables previously described as covariates. To interpret coefficient estimates it is necessary to identify the reference group: a female student from section 1 graded with the penalty rule (without normalization) in the first exam. Thus, coefficient estimates are to be interpreted as differences with respect to the reference group.

Table 7 presents the estimated Poisson, NB, ZIP and ZINB models. The results are very similar quantitatively and significance of coefficients is uniform across models. To choose among these models we run two model selection tests which can be found in Table 8. According to Vuong's test we reject the Poisson model in favor of the ZIP and the likelihood ratio test cannot reject the ZIP model against the more general ZINB model. Therefore, hereinafter we report the ZIP results.

Table 9 searches for the appropriate specification by sequential elimination of non-significant regressors from column (1) to (4). Table 9 shows how the number of omissions is related to all the factors listed above. The scoring rule does in fact have a significant effect on the number of omissions. Being graded with the reward for omissions tends to increase the number of omissions. The inflate part of the ZIP regression indicates that males tend to be non omitters. Although males omitters, omit significantly more than females. Group is also a significant factor in the determination of omissions.

The accumulated score in previous exams and this variable squared are significant. Thus, the effect of the accumulated score on the number of omissions has an inverted U shape. The number of omissions is first increasing in the accumulated score and then decreasing beyond a point. Knowledge and knowledge squared are also significant explanatory variables and, as in the case of the accumulated score, the relationship is concave. Thus the number of omissions is first increasing and then decreasing in knowledge.

There are also several interaction terms that are significant determinants of the number of omissions: accumulated score and gender, reward for omissions and accumulated score, normalized penalty and accumulated score, and lastly reward for omissions and the proxy for knowledge.

In order to determine whether the three rules have a differential effect on omissions within the ZIP framework we run the likelihood ratio tests described in Appendix B. According to the second row of Table 10, we reject the hypothesis of no differences in behavior under penalty and normalized reward. Furthermore, the third row of Table 10 indicates that we cannot reject the hypothesis of no differences in behavior under normalized penalty and normalized reward. These results are consistent with rationality. In addition, there are differences between penalty and normalized penalty but they are not large enough to reject the null of no differential behavior under penalty and normalized penalty (the p-value is slightly over 10 per cent).

7. Concluding Remarks

In this paper we show that scoring rules which penalize for wrong answers and reward for omissions are, contrary to what is generally accepted, not strategically equivalent. We propose a normalization mechanism that makes the two scoring rules strategically equivalent. By confronting students with scoring rules with and without normalization, it should be possible to distinguish the effect of risk aversion from that of psychological factors. Our field experiment shows significant differences in students' behavior when they are evaluated with penalty for wrong answers and reward for omissions, but found no significant differences between normalized scoring rules. This evidence is consistent with expected utility maximizing behavior of students. In addition to the scoring rule, other significant determinants of the number of omissions are: the accumulated score in previous exams on the same course, knowledge, the difficulty and other unobserved characteristics of the exam and gender.

Appendix A: Instructions

The original instructions were given in Spanish and Basque: what follows is a translation. All treatments included the following general instructions.

- *Read all instructions carefully.*
- *You are not allowed to talk during the exam. If you have a question, raise your hand.*
- *Write down your name and ID number on the answer sheet and on this exam.*
- *At the end of the exam you must hand in both this exam and your answer sheet.*
- *This exam has 10 items.*
- *Each item has four possible answers and only one is correct.*
- *You have 30 minutes.*

In addition to these general instruction, each treatment had one more instruction regarding the scoring rule used for that treatment.

Treatment S_2.

- *Your score will be given by the following formula*

$$score = 2 \times \left(rights - \frac{wrongs}{3}\right) = (2 \times rights) - (0.66 \times wrongs).$$

That is, your score depends on the number of rights, wrongs and omits. Each item will be valued according to the following table

Scoring Rule	
Right	+2
Wrong	−0.66
Omit	0

Treatment S_2^*.

- *Your score will be given by the following formula*

$$score = 5 + 1.5 \times \left(rights - \frac{wrongs}{3}\right) = 5 + (1.5 \times rights) - (0.5 \times wrongs).$$

That is, your score depends on the number of rights, wrongs and omits. Each item will be valued according to the following table

Scoring rule	
Right	+1.5
Wrong	−0.5
Omit	0

TREATMENT S_3^*.

- *Your score will be given by the following formula*

$$score = 2 \times \left(rights + \frac{omits}{4} \right) = (2 \times rights) + (0.5 \times omits).$$

That is, your score depends on the number of rights, wrongs and omits. Each item will be valued according to the following table

Scoring Rule	
Right	+2
Wrong	0
Omit	+0.5

Appendix B: Likelihood ratio tests in the ZIP model.

The unrestricted ZIP model is

$$\lambda = e^{X'\beta} = e^{X'\beta + D_3\theta_3 + D_4\theta_4 + X'D_3\eta_3 + X'D_4\eta_4} \tag{6}$$

where D_3 and D_4 are dummy variables for normalized penalty and normalized reward respectively (penalty is the reference group) and X is a vector containing all the other explanatory variables. Let L_1 be the (maximized) value of the likelihood function when the ZIP model is specified as in (6). Notice that this value, L_1, would be the same if the reference group were normalized penalty or normalized reward. Under the null hypothesis that normalized penalty and normalized reward have the same effect on omissions, that is, $\theta_3 = \theta_4$ and $\eta_3 = \eta_4$, the ZIP model is restricted to

$$\lambda = e^{X'\beta + (D_3+D_4)\theta_3 + X'(D_3+D_4)\eta_3}. \tag{7}$$

Let L_2 be the value of the likelihood function when the ZIP model is specified as in (7). In practice this amounts to fitting the restricted model simply including a single dummy variable constructed as the sum of two dummy variables D_3 and D_4. The likelihood ratio test of the null hypothesis that normalized penalty and normalized reward have the same effect on omissions is computed as $-2ln(L_2/L_1)$.

Under the hypothesis that penalty and normalized penalty have the same effect on omissions, that is, $\theta_2 = \theta_3$ and $\eta_2 = \eta_3$, the ZIP model is restricted to

$$\lambda = e^{X'\beta + (D_2+D_3)\theta_2 + X'(D_2+D_3)\eta_2}. \tag{8}$$

Let L_3 be the value of the likelihood function when the ZIP model is specified as in (8). The likelihood ratio test of the null hypothesis that penalty and normalized penalty have the same effect on omissions is computed as the ratio of the likelihoods $-2ln(L_3/L_1)$. The other likelihood ratio test is constructed similarly.

References

[1] Becker, William E. and Carol Johnston, (1999). "The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding," The Economic Record 75(231), 348-57.

[2] Bereby-Meyer, Yoella, Joachim Meyer, Oded M. Flascher. (2002). "Prospect Theory Analysis of Guessing in Multiple Choice Tests," Journal of Behavioral Decision Making 15(4), 313-327.

[3] Bernardo, José M. (1998). "A decision analysis approach to multiple choice examinations." In F. J. Girón (ed.) Applied Decision Analysis. Boston: Kluwer, 195-207.

[4] Bertrand, Marianne, Dean S. Karlan, Sendhil Mullainathan, Eldar Shafir and Jonathan Zinman. (2005). "What's Psychology Worth? A Field Experiment in the Consumer Credit Market," Working Papers 918, Economic Growth Center, Yale University.

[5] Bredon, George. (2003). "Take-Home Tests in Economics," Economic Analysis and Policy 33(1), 52-60.

[6] Budescu, David and Maya Bar-Hillel. (1993). "To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring," Journal of Educational Measurement 30(4), 277-291.

[7] Burgos, Albert. (2004). "Guessing and gambling," Economics Bulletin 4(4), 1-10.

[8] Byrnes, James P., David C. Miller and William D. Schafer. (1999) "Gender Differences in Risk Taking: A Meta-analysis," Psychological Bulletin 125(3), 367-383.

[9] Cadsby, C. Bram and Elizabeth Maynes. (2005). "Gender, Risk Aversion, and the Drawing Power of Equilibrium in an Experimental Corporate Takeover Game," Journal of Economic Behavior & Organization 56, 39-59.

[10] Chan, Nixon and Peter Kennedy. (1999). "Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and "Equivalent" Constructed-Response Exam Questions," Southern Economic Journal 68(4), 957-71.

[11] Espinosa, María Paz and Javier Gardeazabal. (2005) "Personal Characteristics and Behavior in Multiple-Choice Tests under Different Scoring Rules, mimeo.

[12] Haan, Marco, Bart Los, Yohanes Riyanto and Martin Van Geest. (2002). "The Weakest Link - A Field Experiment in Rational Decision Making," Experimental 0203001, Working Paper Archive EconWPA.

[13] Haigh, Michael S. and John List. (2005). "Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis," Journal of Finance LX(1), 523-534.

[14] Heck, Jean L., and David E. Stout. (1997). "Multiple-Choice vs. Open-Ended Exam Problems: Evidence of Their Impact on Student Performance in Introductory Finance," Financial Practice and Education 8, 83-93.

[15] List, John and Daniel Millimet. (2005). "Bounding the Impact of Market Experience on Rationality: Evidence from a Field Experiment with Imperfect Compliance," Departmental Working Papers 0505, Southern Methodist University, Department of Economics.

[16] Prieto, Gerardo. and Ana R. Delgado. (1999). "The role of instructions in the variability of sex-related differences in multiple-choice tests", Personality and Individual Differences 27, 1067-1077.

[17] Siegfried, John J., Phillip Saunders, Ethan Stinar, and Hao Zhang. (1996). "How is Introductory Economics Taught in America?" Economic Inquiry 34, 182-92.

[18] Traub, Ross E., Ronald K. Hambleton and Balwant Singh. (1969). "Effects of Promised Reward and Threatened Penalty on Performance of a Multiple-Choice Vocabulary Test," Educational and Psychological Measurement 29, 847-861.

[19] Traub, Ross E., and Ronald K. Hambleton. (1972). "The Effect of Scoring Instructions and Degree of Speededness on the Validity and Reliability of Multiple-Choice Tests," Educational and Psychological Measurement 32, 737-758.

[20] Walstad, William B. and William Becker. (1994). "Achievement Differences on Multiple-Choice and Essay Tests in Economics," The American Economic Review 84(2), 193-196.

[21] Waters, Carrie W. and Lawrence K. Waters. (1971). "Validity and Likability Ratings for Three Scoring Instructions for Multiple-Choice Vocabulary Tests," Educational and Psychological Measurement 31, 935-938.

Table 1: Description of sessions / exams

Session / Exam	Date	Treatments	Students
1	March 9, 2005	S_2, S_2^*, S_3^*	177
2	March 21, 2005	S_2, S_2^*, S_3^*	169
3	April 13, 2005	S_2, S_2^*, S_3^*	162
4	May 4, 2005	S_1	152
5	May 20, 2005	S_1	148

Table 2: Experimental design

Group	First exam	Second exam	Third exam	Fourth exam	Fifth exam
Blue	Norm. Penalty	Penalty	Norm. Reward	Number right	Number right
Yellow	Norm. Reward	Norm. Penalty	Penalty	Number right	Number right
White	Penalty	Norm.Reward	Norm. Penalty	Number right	Number right

Table 3: Group characteristics

Group	Number of Students	Males	Females	Average knowledge	Proportion of exams with no omissions
Blue	54	27	27	6.47	0.34
Yellow	57	29	28	6.18	0.26
White	49	25	24	6.57	0.40

Average knowledge is the average grade in a previous Macroeconomics course on a 10-point scale.

Table 4: Descriptive Statistics on Omissions

Treatment	Num. of Obs.	Mean	Std. Dev.	Min.	Max.
			First exam		
Penalty, S_2	49	0.86	0.87	0	3
Normalized Penalty, S_2^*	54	1.11	1.11	0	5
Normalized Reward, S_3^*	57	1.47	1.43	0	5
			Second exam		
Penalty, S_2	54	1.63	1.05	0	4
Normalized Penalty, S_2^*	57	2.04	1.53	0	6
Normalized Reward, S_3^*	49	1.57	1.26	0	5
			Third exam		
Penalty, S_2	57	1.27	1.03	0	4
Normalized Penalty, S_2^*	49	0.65	0.90	0	4
Normalized Reward, S_3^*	54	0.81	1.07	0	4

Table 5: Are results rule-independent?

	First Exam	Second exam	Third exam
Penalty vs. Norm. Reward	-2.002 (0.0453)	0.448 (0.6542)	2.663 (0.0077)
Penalty vs. Norm. Penalty	-1.015 (0.3099)	-1.243 (0.2137)	3.372 (0.0007)
Norm. Reward vs. Norm. Penalty	1.133 (0.2572)	-1.462 (0.1462)	0.611 (0.5411)

Mann-Whitney test (p-value)

Table 6: Are results exam-independent?

	Penalty S_2	Penalty Normalized S_2^*	Reward S_3^*
First vs. Second	-3.765 (0.0002)	-0.603 (0.5467)	-3.378 (0.0007)
First vs. Third	-2.126 (0.0335)	2.500 (0.0124)	2.317 (0.0205)
Second vs. Third	1.766 (0.0774)	3.289 (0.0010)	5.115 (0.0000)

Mann-Whitney test (p-value)

Table 7: Poisson, NB, ZIP and ZINB regressions.

Dependent variable: number of omissions:	Poisson	NB	ZIP	ZINB
Constant	-0.4796	-0.4796	-0.6938	-0.6938
	(0.4913)	(0.4914)	(0.5191)	(0.5191)
Normalized Reward	1.3321***	1.3320***	1.4196***	1.4197***
	(0.4345)	(0.4346)	(0.4395)	(0.4396)
Normalized Penalty	0.5830	0.5829	0.6102	0.6103
	(0.4450)	(0.4450)	(0.4362)	(0.4363)
Male	0.0577	0.0577	0.1219	0.1221
	(0.4635)	(0.4636)	(0.4677)	(0.4678)
Second Exam	0.1013	0.1013	0.4348	0.4350
	(0.4001)	(0.4001)	(0.3991)	(0.3991)
Third Exam	-0.0424	-0.0424	0.3518	0.3518
	(0.6090)	(0.6090)	(0.6212)	(0.6212)
Section 2	0.0808	0.0808	0.0723	0.0724
	(0.1111)	(0.1111)	(0.1075)	(0.1075)
Section 3	-0.0499	-0.0499	-0.0226	-0.0226
	(0.1302)	(0.1302)	(0.1269)	(0.1269)
Section 4	0.0742	0.0741	0.0872	0.0872
	(0.1528)	(0.1529)	(0.1491)	(0.1492)
Section 5	-0.1250	-0.1250	-0.1230	-0.1230
	(0.1209)	(0.1209)	(0.1190)	(0.1190)
Blue Group	-0.3337**	-0.3336**	-0.3251**	-0.3251**
	(0.1376)	(0.1376)	(0.1373)	(0.1373)
White Group	-0.2456*	-0.2456*	-0.2314	-0.2314
	(0.1437)	(0.1437)	(0.1424)	(0.1425)
Accumulated score	0.0853***	0.0853***	0.0696**	0.0696**
	(0.0313)	(0.0313)	(0.0312)	(0.0312)
Accumulated score, squared	-0.0023***	-0.0023***	-0.0017**	-0.0017**
	(0.0008)	(0.0008)	(0.0008)	(0.0008)
Knowledge	0.2238**	0.2237**	0.2475**	0.2475**
	(0.1111)	(0.1111)	(0.1147)	(0.1147)
Knowledge, squared	-0.0180**	-0.0180**	-0.0176**	-0.0176**
	(0.0077)	(0.0077)	(0.0078)	(0.0078)
Male * Normalized Reward	-0.4273**	-0.4272**	-0.3712*	-0.3712*
	(0.2158)	(0.2158)	(0.2114)	(0.2114)
Male * Normalized Penalty	-0.2311	-0.2311	-0.1703	-0.1704
	(0.2064)	(0.2064)	(0.2075)	(0.2075)
Male * Second Exam	0.4656	0.4656	0.2815	0.2815
	(0.3013)	(0.3013)	(0.3247)	(0.3248)

Table 7: Poisson, NB, ZIP and ZINB regressions (continued).

Male * Third Exam	0.5957	0.5957	0.4781	0.4781
	(0.4795)	(0.4795)	(0.4770)	(0.4771)
Male * Accumulated Score	-0.0332*	-0.0332*	-0.0474**	-0.0474**
	(0.0194)	(0.0194)	(0.0196)	(0.0196)
Male * Grade in Interm. Macro I	-0.0271	-0.0271	0.0243	0.0242
	(0.0641)	(0.0641)	(0.0660)	(0.0660)
Male * Blue Group	0.0676	0.0676	0.0789	0.0788
	(0.2103)	(0.2103)	(0.2152)	(0.2152)
Male * White Group	-0.0481	-0.0481	-0.0339	-0.0339
	(0.2186)	(0.2186)	(0.2131)	(0.2131)
Norm. Reward * Accumulated Score	-0.0207	-0.0207	-0.0258*	-0.0258*
	(0.0145)	(0.0145)	(0.0144)	(0.0144)
Norm. Penalty * Accumulated Score	-0.0371***	-0.0371***	-0.0417***	-0.0417***
	(0.0128)	(0.0128)	(0.0124)	(0.0124)
Norm. Reward * Knowledge	-0.1516**	-0.1516**	-0.1570**	-0.1570**
	(0.0675)	(0.0675)	(0.0685)	(0.0685)
Norm. Penalty * Knowledge	-0.0278	-0.0278	-0.0226	-0.0226
	(0.0689)	(0.0689)	(0.0676)	(0.0676)
Second Exam * Knowledge	-0.0156	-0.0156	-0.0482	-0.0482
	(0.0611)	(0.0611)	(0.0631)	(0.0631)
Third Exam * Knowledge	-0.0699	-0.0699	-0.1153	-0.1153
	(0.0853)	(0.0853)	(0.0888)	(0.0888)
Inflate			ZIP	ZINB
Constant			-17.9080***	-18.4791***
			(1.8108)	(1.8108)
Male			15.1305***	15.7060***
			(0.8074)	(0.8067)
Second Exam			1.9708	1.9727
			(1.8172)	(1.8226)
Third Exam			-12.0833***	-11.6337***
			(2.7537)	(2.7645)
Accumulated score			-0.4436	-0.4441
			(0.4091)	(0.4110)
Knowledge			0.3340	0.3340
			(0.2065)	(0.2065)

Heteroskedasticity robust standard errors in parentheses.

*, ** and *** mean significance at 1, 5 and 10 % respectively.

Table 8: Model Selection.

Test	Statistic	P-value
Poisson vs. ZIP Vuong's test	1.92	0.03
ZIP vs. ZINB Likelihood ratio test	0.01	0.47

Table 10: Likelihood Ratio Tests: equality of scoring rules.

Test	Statistic	P-value
Penalty vs. Normalized Penalty	7.7555	0.1010
Penalty vs. Normalized Reward	12.1747	0.0161
Normalized Penalty vs. Normalized Reward	4.6826	0.3214

Table 9: Specification search for the ZIP regression

Dependent variable: number of omissions.	(1)	(2)	(3)	(4)
Constant	-0.6938	-0.7385	-0.3534	-0.3935
	(0.5191)	(0.5282)	(0.3343)	(0.3535)
Normalized Reward	1.4196***	1.4469***	1.3549***	1.3880***
	(0.4395)	(0.4436)	(0.3951)	(0.4035)
Normalized Penalty	0.6102	0.6224	0.5561	0.6246
	(0.4362)	(0.4407)	(0.4461)	(0.4186)
Male	0.1219	0.2764	0.3604*	0.3629**
	(0.4677)	(0.4230)	(0.1871)	(0.1847)
Second Exam	0.4348	0.5377	0.1823	
	(0.3991)	(0.3955)	(0.2287)	
Third Exam	0.3518	0.5315	-0.2264	
	(0.6212)	(0.5871)	(0.2761)	
Section 2	0.0723	0.0761		
	(0.1075)	(0.1068)		
Section 3	-0.0226	-0.0234		
	(0.1269)	(0.1260)		
Section 4	0.0872	0.0831		
	(0.1491)	(0.1502)		
Section 5	-0.1230	-0.1255		
	(0.1190)	(0.1186)		
Blue Group	-0.3251**	-0.2905**	-0.2922**	-0.2744**
	(0.1373)	(0.1257)	(0.1280)	(0.1227)
White Group	-0.2314	-0.2449*	-0.2435*	-0.2417**
	(0.1424)	(0.1253)	(0.1275)	(0.1212)
Accumulated score	0.0696**	0.0638**	0.0709**	0.0853***
	(0.0312)	(0.0301)	(0.0288)	(0.0157)
Accumulated score, squared	-0.0017**	-0.0018**	-0.0019***	-0.0028***
	(0.0008)	(0.0007)	(0.0007)	(0.0005)
Knowledge	0.2475**	0.2466**	0.1811**	0.1900**
	(0.1147)	(0.1159)	(0.0799)	(0.0836)
Knowledge, squared	-0.0176**	-0.0172**	-0.0169***	-0.0167***
	(0.0078)	(0.0077)	(0.0059)	(0.0062)
Male * Normalized Reward	-0.3712*	-0.3802*	-0.3722*	-0.3909*
	(0.2114)	(0.1995)	(0.1998)	(0.2038)
Male * Normalized Penalty	-0.1703	-0.1810	-0.1935	-0.2142
	(0.2075)	(0.2052)	(0.2115)	(0.2000)
Male * Second Exam	0.2815			
	(0.3247)			

Table 9: Specification search for the ZIP regression (continued).

Male * Third Exam	0.4781			
	(0.4770)			
Male * Accumulated Score	-0.0474**	-0.0298***	-0.0307***	-0.0295***
	(0.0196)	(0.0100)	(0.0102)	(0.0101)
Male * Knowledge	0.0243	0.0136		
	(0.0660)	(0.0648)		
Male * Blue Group	0.0789			
	(0.2152)			
Male * White Group	-0.0339			
	(0.2131)			
Norm. Reward * Accumulated Score	-0.0258*	-0.0259*	-0.0269*	-0.0284*
	(0.0144)	(0.0144)	(0.0150)	(0.0150)
Norm. Penalty * Accumulated Score	-0.0417***	-0.0417***	-0.0421***	-0.0390***
	(0.0124)	(0.0125)	(0.0130)	(0.0131)
Norm. Reward * Knowledge	-0.1570**	-0.1605**	-0.1439**	-0.1479***
	(0.0685)	(0.0690)	(0.0620)	(0.0633)
Norm. Penalty * Knowledge	-0.0226	-0.0242	-0.0115	-0.0237
	(0.0676)	(0.0686)	(0.0717)	(0.0660)
Second Exam * Knowledge	-0.0482	-0.0494		
	(0.0631)	(0.0631)		
Third Exam * Knowledge	-0.1153	-0.1153		
	(0.0888)	(0.0897)		
Inflate	ZIP			
Constant	-17.9080***	-19.6587***	-17.5292***	-17.4379***
	(1.8108)	(1.6998)	(1.7830)	(1.8334)
Male	15.1305***	17.0105***	15.3652***	15.4566***
	(0.8074)	(0.7269)	(0.8670)	(1.1337)
Second Exam	1.9708	1.9431	2.0917	4.6357
	(1.8172)	(2.2679)	(4.1732)	(4.6306)
Third Exam	-12.0833***	-13.6500***	-12.5173*	-12.5173
	(2.7537)	(3.7097)	(7.2203)	(8.1652)
Accumulated score	-0.4436	-0.4671	-0.5176	-1.3397
	(0.4091)	(0.5410)	(1.1115)	(1.5176)
Knowledge	0.3340	0.3224	0.2506	0.2261
	(0.2065)	(0.2019)	(0.1971)	(0.2056)

Heteroskedasticity robust standard errors in parentheses.

*, ** and *** mean significance at 1, 5 and 10 % respectively.