經濟與管理論叢(Journal of Economics and Management), 2013, Vol. 9, No. 2, 107-135

Do Students Behave Rationally in Multiple Choice Tests? Evidence from a Field Experiment

María Paz Espinosa[*]

Departamento de Fundamentos del Análisis Económico II, University of the Basque Country, (UPV/EHU), Spain

Javier Gardeazabal

Departamento de Fundamentos del Análisis Económico II, University of the Basque Country, (UPV/EHU), Spain

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. In psychometrics, penalty and reward scoring rules are considered equivalent. However, experimental evidence indicates that students behave differently under penalty or reward scoring rules. These differences have been attributed to the different framing (penalty versus reward). In this paper, we model students' behavior in multiple choice tests as a choice among lotteries. We show that strategic equivalence among penalty and reward scoring rules holds only under risk neutrality. Therefore, risk aversion could be an alternative explanation to the previously found differences in students' behavior when confronted with penalty and reward scoring rules. We suggest the use of a modified penalty scoring rule which is equivalent to the reward rule for whatever risk attitudes students might have. To disentangle the effect of framing and risk aversion on students' behavior we design a field experiment with three treatments, each one with a different scoring rule. Two of these scoring rules are equivalent but have different framing, while the third is not equivalent but has the same framing as one of the other two. The

[*] Correspondence to: University of the Basque Country, UPV/EHU. BRiDGE, Departamento de Fundamentos del Análisis Económico II. Avenida Lehendakari Aguirre 83, 48015 Bilbao. Spain. Email: mariapaz.espinosa@ehu.es ; javier.gardeazabal@ehu.es. Financial support from the Spanish Ministry of Science and Innovation MICINN (ECO2009-09120 and ECO2012-35820) and from the Basque Government (IT-313-07) is gratefully acknowledged.

experimental results indicate that differences in students’ behavior are due to risk aversion and not due to different framing.

Keywords: scoring rules, risk aversion, field experiment
JEL classification: C93, D03, D81

1 Introduction

Multiple-choice tests are widely used as an evaluation tool,[1] their main advantages over constructed-response tests being that they guarantee wider content sampling and preclude measurement errors introduced by the grader. The main drawback 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 $\boldsymbol{S_n}$. 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 and is intended to reduce guessing behavior. Although rarely used, an alternative way of discouraging guessing is to reward omitted questions.

In the psychometric literature, scoring rules incorporating a correction for guessing in the form of a penalty for wrong answers ($\boldsymbol{S_P}$) and a reward for omitted questions ($\boldsymbol{S_R}$) were considered equivalent since one is in fact an affine transformation of the other. However, empirical evidence indicated that students behaved differently under both scoring rules. Bereby-Meyer *et al*. (2002) confronted students with scoring rules that penalized for wrong answers and rewarded for omissions. Their experimental evidence shows that students omitted more items when they were penalized for incorrect answers than when rewarded for omissions. These differences in students' behavior were thought to be associated with framing

[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.

and attributed to psychological factors.[3] An explanation of this non-equivalence result was advanced by Budescu and Bar-Hillel (1993) appealing to the different considerations 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 and the experimental results interpreted as evidence in favor of the theory (Bereby-Meyer *et al*., 2003).[4] Other theories of student behavior in multiple choice tests include Bernardo (1998) who assumes that students maximize the score or minimize the probability of failing the exam, and Burgos (2004) who uses Prospect Theory to postulate a utility function that assigns different values to losses and gains.

The psychometrics literature neglects the possibility that risk aversion could be the reason why students behave differently when confronted with scoring rules that penalize for incorrect answers or reward for omissions. For instance, Bereby-Meyer *et al*. (2002) confront students with scoring rules that penalized for incorrect answers and rewarded omissions, in both cases with positive expected score, and claim that "…guessing was clearly always the optimal strategy" (p. 323). These authors fail to consider the possibility that a risk averse student could omit an item with positive expected reward. Similarly, in Bereby-Meyer *et al*. (2003) experimental study with the penalty and reward scoring rules, the authors claim that "… answering all items was the dominant strategy for both rules" (p. 207) again neglecting the possibility that risk aversion might induce omissions. The implicit assumption in most of the psychometrics literature is that test takers are expected score maximizers. Our contribution in this paper is to provide a differentiation between students' risk preferences effects and framing effects on the number of omitted questions in multiple choice tests.

In this paper students' behavior in multiple choice tests is modeled as a choice over lotteries where risk considerations play an important role. We contribute three theoretical results which will be useful to disentangle the effect of framing and risk aversion. First, we show that when examinees are risk-averse the two scoring rules

[3] See Traub *et al*. (1969), Traub and Hambleton (1972) and Waters and Waters (1971).
[4] See Kahneman and Tversky (1979) on Prospect Theory.

imply a different trade-off when deciding whether to answer a question or not. Therefore, expected utility maximizers may behave differently under S_P than under S_R, even though they are linearly related. Second, we demonstrate that the two scoring rules are equivalent under risk neutrality. Third, we also show that the penalty rule can be modified so that it becomes strategically equivalent to the reward rule 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's validity and reliability and, in particular, the impact of psychological factors. Previous experiments confront subjects to scoring rules that are strategically equivalent only for risk neutral subjects. Therefore, the observed differences in behavior could be attributed to the framing of the scoring rule or to risk attitudes. Our experimental design allows us to distinguish between the two factors by confronting students to rules with different framing that are strategically equivalent for all types of risk attitudes.

To determine whether in this context risk aversion may be significant enough to give rise to the differences in observed behavior, we designed an experiment using several scoring rules. In a regular undergraduate Macroeconomics course, it was announced that exams would be graded with different scoring rules for different groups of students, so that all students knew well in advance the exact rule they would face in the exam.[5] The results of our field experiment indicate that under equivalent scoring rules, there are no significant differences in the number of omissions, even though one rule is framed as a penalty for wrong answers and the other as a reward for omissions, while differences in behavior are observed when the rules are not equivalent. Therefore, the results are consistent with rational student behavior. 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 and psychological factors did not seem to play a role in students' behavior. Note that, in this field experiment, subjects' decisions determined their grade on the course, so there was a strong incentive to take the right decisions. In summary, the evidence is consistent with expected utility maximization and supports the hypothesis that differences in behavior are due to risk aversion

[5] The experimental design described in Section 4 guarantees equal treatment for all students.

rather than psychological factors.[6] We also find that gender and students' knowledge are important determinants of the number of omissions.

We do not address the issue of the optimality of the grading procedures and the theory and empirical evidence in this paper are in no way intended to justify or recommend the use of any particular scoring rule.[7] Our work is mainly a contribution to the study of risk attitudes and rationality within a particular context. Nevertheless, a better understanding of the incentives behind these rules could be a useful first step for studying the optimal way of designing multiple-choice tests.

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. In Section 5 we report the results of the field experiment and perform the statistical analysis. Section 6 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, $u(s)$, representing her valuation of the score, s, obtained in the exam. We assume this valuation is such that $u'(s) \geq 0$.[8] We do not restrict the second derivative of the utility function, so students could be risk averse, risk neutral or risk loving. Note also that the utility function is independent of the scoring rule. Students may have different preferences and different levels of knowledge.

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

$$S_n = r \ .$$

Some scoring rules incorporate a correction for guessing feature. Typically, there is a penalty of $1/M-1$ points for each incorrect answer. This scoring rule yields a final score:

[6] There are numerous laboratory experiments documenting different types of deviations from rationality, but field experiments are scarce. See Bertrand *et al.* (2005), Haan *et al.* (2002), Haigh and List (2005) and List and Millimet (2005) for notable exceptions.

[7] See Espinosa and Gardeazabal (2010).

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

$$S_P = r - \frac{w}{M-1},$$

where r and w are the number of right and wrong answers, respectively. An alternative rule for discouraging guessing is to give $1/M$ points for each omitted question. This scoring method yields a final score:

$$S_R = r + \frac{o}{M},$$

where o is the number of questions omitted.

There are three important features of these scoring rules:

(i) First, 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. However, S_P is framed in terms of losses for wrong answers while S_R is framed in terms of gains.

(ii) Second, S_p and S_R are linearly related as:

$$S_R = \frac{N}{M} + \frac{M-1}{M} S_P. \quad (1)$$

(iii) Third, both scoring rules use values of the penalty for wrongs and reward for omissions such that the expected value of randomly answering an item equals the value of omitting. Consider an examinee who has no clue about the answer to an item and selects an answer randomly. Thus, the probability of a right answer is $1/M$ and the probability of failing the item is $M-1/M$. Under S_P, the expected value from answering is:

$$\left(\frac{1}{M}\times 1\right) + \left(\frac{M-1}{M}\times\left(-\frac{1}{M-1}\right)\right) = 0,$$

which is equal to the gain from omitting. Under S_R the expected value from answering the item is:

$$\left(\frac{1}{M}\times 1\right) + \left(\frac{M-1}{M}\times 0\right) = \frac{1}{M},$$

which is equal to the gain from omitting $1/M$.

We believe that features (ii) and (iii) of the scoring rules might be the reason why both rules have been considered equivalent in the psychometric literature.

A contribution of this paper is the modeling of multiple choice tests as lotteries. Item i can be viewed as a gamble in which a student has probability q_i of getting the right answer and the probability of failing the item is $1-q_i$. Of course, these probabilities depend on the knowledge of the student and the difficulty of the item. Assume the student answers only item i, leaving $N-1$ items unanswered. Denote by $s(r,w,o)$ the score obtained from r rights, w wrongs and o omissions. Let $\ell\{i\}$ denote the lottery induced by answering only item i, i.e. obtaining a score $s(1,0,N-1)$ with probability q_i and $s(0,1,N-1)$ with probability $1-q_i$. 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 lottery. If the student evaluates lotteries according to the Expected Utility Theory, then:

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

where $u(\cdot)$ is the student's valuation of the score.

Let $\ell\{i,j\}$ denote the compound lottery induced by answering items i and j, leaving $N-2$ items unanswered. The payoffs and probabilities of this lottery are given in Table 1 where q_i and q_j are the probabilities of getting the right answer to items i and j, respectively. The utility derived from this lottery is:

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

Table 1: Scores and Probabilities

Scores	Probabilities
$s(2,0,N-2)$	$q_i q_j$
$s(1,1,N-2)$	$q_i(1-q_j)+(1-q_i)q_j$
$s(0,2,N-2)$	$(1-q_j)(1-q_i)$

In an exam with N items any subset of items is a compound lottery. Let $\mathrm{L}(N)$ 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:

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

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 perfectly rational student would choose the best compound lottery in $\mathrm{L}(N)$. Formally, she would maximize the expected utility over the set of all possible compound lotteries:

$$\begin{array}{l}\max U(\ell) \\ \ell \in \mathrm{L}(N)\end{array} \tag{2}$$

In our model a rational test taker is expected to answer items to maximize expected utility. This calculation is difficult to perform for various reasons. First, the analytic solution to the problem is not straightforward and, second, it requires estimates of the probabilities of answering items correctly. Of course, subjects taking multiple-choice tests are in no way assumed to literally perform such calculations in real exams. Our model of rational behavior should be understood as a theory and not a rule of behavior or a description of cognitive decision processes (e.g., McKenzie, 2003).

3 Theoretical Results

In this section we analyze whether penalizing for wrong answers and rewarding omissions induce the same behavior on examinees. Penalizing for wrong answers (S_P) and rewarding omissions (S_R) have been wrongly considered equivalent, probably because one is just an affine transformation of the other (see equation (1)). Therefore, authors have focused on the different framing, namely losses (S_P) and gains (S_R), e.g. Bereby-Meyer *et al.* (2003).

In order to be precise about the equivalence between scoring rules we introduce the following definition.

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

This section presents three results. The first is that penalizing wrong answers and rewarding omissions, as defined in the previous section, are not in general strategically equivalent.

Proposition 1. For risk-averse exam takers, S_P and S_R are not strategically equivalent.

To show this, it is sufficient to find an example where a risk averse exam taker would behave differently under S_P than under S_R. Consider an exam with one item, $N=1$, two alternatives, $M=2$, a student with a concave valuation such as $u(s)=\sqrt{a+s}$, with $a \ge N$, and probability q_i of getting the right answer. In this case, a student is faced with a set of two lotteries, $\{\ell\{0\},\ell\{1\}\}$, that is, omitting the item and answering the item. Under S_P 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_R the examinee obtains 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 student with $a=1$ and a probability of answering correctly of $q_i=0.6$, both inequalities hold and therefore the student would omit under S_P and answer under S_R. 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_P and S_R 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_P, 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:

$$\sum_i q_i^{\ell} s_{pi}^{\ell} \ge \sum_i q_i^{\ell'} s_{pi}^{\ell'}, \tag{3}$$

for all $\ell' \in L(N)$, where s_{pi}^{ℓ} are the scores under S_P of all possible outcomes in lottery ℓ and q_i^{ℓ} are the associated probabilities of those outcomes, so that $\sum_i q_i^{\ell}=1$. Multiplying by $M-1/M$ and adding N/M to both sides of (3) we get:

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

for all $\ell' \in L(N)$. Using equation (1), the previous equation can be written as:

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

for all $\ell' \in L(N)$. In words, the student chooses lottery ℓ in the set $L(N)$ under S_P if and only if she also chooses it under S_R. This completes the proof of equivalence.

Scoring rules S_P and S_R 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_P and S_R. In order to measure the effects of framing, it is necessary to use scoring rules that are strategically equivalent for all types of risk attitudes. For that purpose, we propose a modified scoring rule with penalty denoted S_P^*:

$$S_P^* = \frac{M-1}{M} r - \frac{1}{M} w + \frac{N}{M}.$$

Notice that the modified penalty scoring rule can be written as $S_P^* = (S_p + pN)/(1+p)$ where $p = 1/M - 1$ is the penalty, so it is an affine transformation of the standard penalty rule. Intuitively, the modified penalty rule gives a startup score of N/M to all students and then subtracts 1/M for each wrong answer. This is in fact the same as rewarding for omissions, as the following proposition demonstrates.

Proposition 3. For each and every type of risk preferences, S_P^* and S_R are strategically equivalent.

For S_P^* and S_R to be strategically equivalent the student should answer the same set of items under the two scoring rules, i.e. the solution to problem (2) should be the same. To prove this, we simply have to show that the set $L(N)$ is identical under S_P^* and S_R. We do this in two steps. First, note 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 the right answers plus wrong answers plus omissions we have that:

$$S_P^* = \frac{M-1}{M} r - \frac{1}{M} w + \frac{N}{M} = \frac{M-1}{M} r - \frac{1}{M}(N - r - o) + \frac{N}{M} = r + \frac{o}{M} = S_R,$$

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

Corollary 1. If scoring rules S_i and S_j are strategically equivalent, so are λS_i and λS_j for $\lambda > 0$.

This follows from the fact that if the two scoring rules are strategically equivalent, they must yield the same payoffs and therefore, after multiplying their scores by a positive constant, the payoffs of the two scoring rules would also be identical.

The strategic equivalence of scoring rules S_P^* and S_R allows us 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 under S_P^* and S_R, then differences in behavior are due to framing.

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 penalty and reward scoring rules that are strategically equivalent for all risk attitudes, to determine whether there are any differences which could be attributed to framing. Third, we also try to determine which variables influence students' decisions to omit items.

We conducted a field experiment by grading students with different scoring rules in the exams of a regular course. The payoff in terms of grade is particularly appropriate for two reasons. First, grades generate stronger incentives than small amounts of money used in other contexts. 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 2. Subjects were second year undergraduate students pursuing a bachelor’s degree in Economics, enrolled for Intermediate Macroeconomics in the Spring of 2005. The students' performance on the course was evaluated using five multiple-choice tests; each of the 5 exams was worth 20 points summing up to a total of 100 points. 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.

Table 2: Description of Sessions/Exams

Session/Exam	Date	Treatments (Group)	Participants
1	March 9	S_P (W), S_P^*(B), S_R(Y)	177
2	March 21	S_P(B), S_P^*(Y), S_R(W)	169
3	April 13	S_P(Y), S_P^*(W), S_R(B)	162
4	May 4	S_n(B,W,Y)	152
5	May 20	S_n(B,W,Y)	148

W: white, B: blue, Y: yellow; S_P: penalty, S_P^*: modified penalty, S_R: reward

The experiment had three treatments: penalty for incorrect answers, S_P, modified penalty for incorrect answers, S_P^*, and reward for omissions, S_R. Given the parameters of the exams, $N = 10$ and $M = 4$ the scoring rules are as follows: $S_P = r - \frac{1}{3}w$, $S_P^* = 2.5 + \frac{3}{4}r - \frac{1}{4}w$ and $S_R = r + \frac{1}{4}o$. For the sum of scores obtained in all exams to add up to 100, all test scores are multiplied by 2,[9] so that the scoring rules presented to students were: $S_P = 2r - \frac{2}{3}w$, $S_P^* = 5 + \frac{3}{2}r - \frac{1}{2}w$ and $S_R = 2r + \frac{1}{2}o$. The rules were also presented to the subjects in a table; for example, students in the *reward* treatment were told that they would be graded according to the following table (see the experimental instructions in the appendix):

Scoring Rule	
Right	+2
Wrong	0
Omit	+0.5

These rules were posted on the course website along with other useful information so that students were well aware of the scoring methods.

The experimental design guaranteed equal treatment for students. The expected

[9] Notice that, by Corolary 1, this rescaling of scoring rules does not affect the strategic equivalence between pairs of scoring rules.

score is lower with scoring rule S_P than with rules S_R and S_p^*. Therefore, splitting subjects into three treatments (one for each rule) would favor students in treatments S_R and S_p^*. To treat all students equally each examinee was evaluated once with S_P, S_p^* and S_R in the first three exams. 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 assessed with a different scoring rule in each exam according to the design in Table 2. Therefore, group indicates a particular order in the administration of treatments. There are six possible permutations of treatments while only three groups. Had we split students in six groups, group size would be cut in half. Instead, we split students in three groups. A shortcoming of this experimental design is that the particular order of scoring rules might influence the results. That is why we control for groups in the regressions reported below.

The design of the experiment also takes into account that after the first three exams, students with low accumulated score have less incentive to omit, as a large number of omissions does not guarantee a passing grade (55%). For this reason, the fourth and fifth exams were graded with S_n (number right). 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. The analysis was restricted to the 160 students who took the first three exams. 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. However, according to the data shown in Table 3, groups are probabilistically equivalent. After each exam, students were told their scores.

Table 3: Group Characteristics

Group	Number of Subjects	Males	Females	Knowledge*	Proportion of Exams with No Omissions
Blue	54	27	27	6.47	0.34
Group	Number of Subjects	Males	Females	Knowledge*	Proportion of Exams with No Omissions
Yelow	57	29	28	6.18	0.26
White	49	25	24	6.57	0.40

*Knowledge=Average grade in a previous Macroeconomics course

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. When using penalty scoring, examiners are generally recommended to advise students not to omit if they can 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. Therefore, we did not give students any advice.

Under number-right scoring, rational students ought to respond to all items, and this is what happened in the last two exams, in which no student omitted a single item. Thus, the analysis is restricted to the first three exams. Table 4 reports basic descriptive statistics for the first three exams. 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.

Table 4: Omissions. Descriptive Statistics

	Observations	Mean	Std. Dev	Min.	Max.
Treatment			First Exam		
Penalty S_P	49	0.86	0.87	0	3
Modified Penalty S_P^*	54	1.11	1.11	0	5
Reward S_R	57	1.47	1.43	0	5
			Second Exam		
Penalty S_P	54	1.63	1.05	0	4
Modified Penalty S_P^*	57	2.04	1.53	0	6
Reward S_R	49	1.57	1.26	0	5
Treatment			Third Exam		
Penalty S_P	57	1.27	1.03	0	4
Modified Penalty S_P^*	49	0.65	0.90	0	4
Reward S_R	54	0.81	1.07	0	4

5 Experimental Results

As a preliminary way of assessing the effect of scoring rules on omissions, Figure 1 plots the histogram for omissions under the three rules. Omissions seem to be fairly similar across scoring rules, but we can see a difference between the shape of the histogram for the penalty scoring rule and the other two. The histograms for the reward and modified penalty are fairly similar and suggest that there is no difference between the strategically equivalent scoring rules (reward and modified penalty) despite their different framing. Additionally, note the difference in the shape of the histograms of the penalty scoring rule and the other two (reward and modified penalty) which are not strategically equivalent to the former.

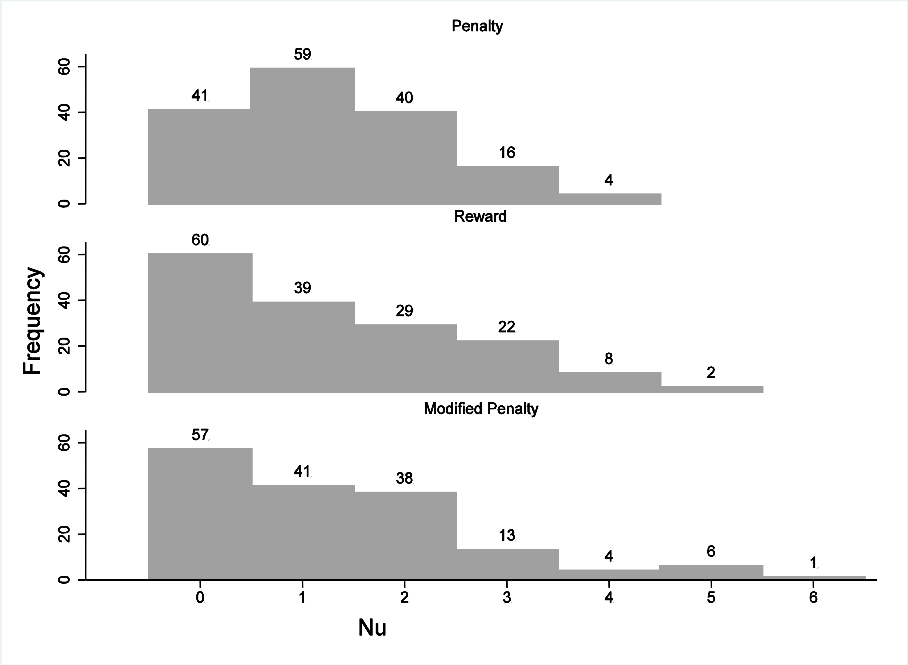

Figure 1: Histograms for Omissions under the Three Scoring Rules

Table 5 reports the Mann-Whitney test statistics for the null hypothesis that the samples of omissions from the first exam come from the same population. According to these results we can reject equality of distributions between penalty and reward and we cannot reject the hypothesis of equality of distributions between reward and modified penalty. The results are consistent with rational risk averse subjects. Rejection of the null of equality of distributions between penalty and reward scoring rules, which are not strategically equivalent, 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 modified penalty and reward is also consistent with the results in Proposition 3. These two scoring rules are strategically equivalent and therefore should induce the same behavior in rational students regardless of their attitudes toward risk. It is worth noting that in the comparison between modified penalty and reward, framing differs (losses in the first scoring rule and gains in the second). Nevertheless, subjects did not show any significant reaction to the different framing.

Table 5: Omissions and Scoring Rules. Mann-Whitney Test

	First Exam
Penalty vs. Reward	-2.002 (0.0453)
Reward vs. Modified Penalty	1.133 (0.2572)

p-values in parentheses.

The evidence reported in Table 5 makes use of data only from the first exam. The reason for doing so is that only before the first exam, students had an accumulated score of zero, while after the first exam their accumulated scores from previous exams were different, reflecting differences in knowledge, luck and the differential effect of the scoring rules. From the second exam on, students with different scores might have had different behavior towards omission even when faced with the same scoring rule. Therefore, the results of the first exam are the only ones not affected by the scores in past exams. In other words, after the first exam, the different treatment groups are not probabilistically equivalent because their accumulated score in previous exams could be different. Therefore, the Mann-Whitney test in Table 5 cannot be used to identify the effect of treatment effects of scoring rules. However, we can still carry out inference using the results of the second and third exams under the assumption of conditional mean independence, that is, conditional on a set of covariates, treatments and potential omissions are mean independent.

In addition to the treatment effects of the scoring rules, the number of omissions could be affected by a number of covariates:

1. The *accumulated score* in previous tests might influence students' behavior. After the first exam the grades are revealed and this may affect the way in which the compound lotteries in the second and third exams are evaluated. A student with a high accumulated score might decide to omit more (or less) 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, the score obtained in the first exam in the second exam, and the sum of the scores in the first and second exams in the third one.
2. *Knowledge*, and in particular knowledge of Macroeconomics, should determine the behavior of students. All else being 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 a previous Macroeconomics course either in the previous semester (Fall 2004) or in a previous year.[10] We also use as an alternative measure of knowledge the grade obtained in the last two exams (scored with number right), *knowledge2*.
3. The difficulty of the *exam* should definitely influence the number of omissions. 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 *et al*., 1999, Cadsby and Maynes, 2005 and Scotchmer 2008). Since scoring rules S_P and S_R are not equivalent for risk-averse subjects, gender might affect the number of omissions.[11] To account for

[10] Three students did not have a grade in a previous Macroeconomics course, which further restricted our sample to 157 students.

[11] Furthermore, it has been documented that instructions concerning guessing behavior may affect

these differences we include a gender dummy variable.

5. 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.
6. The order in which students face the scoring rules is determined by the *group* (Blue, Yellow and White). To control for order effects and unobservable differences in group characteristics we include a set of group dummies.

Next we apply formal statistical procedures to take into account the effect of these factors on students' behavior. By controlling for these factors we can pool the second and third exam data together with the first exam and obtain more reliable statistical results. We do this in two steps. First, we analyze the effect of scoring rules (treatments) and other covariates in the decision whether to omit or not to omit. We define a binary variable taking value 1 if the individual has omitted at least one item and 0 if none has been omitted and use logistic regression. Second, we analyze the effect of scoring rules and other covariates on the number of omissions and use count data regression. In both cases we use the same reference group, a female student from section 5, in the blue group, graded with the reward rule S_R in the third exam. Thus, coefficient estimates are to be interpreted as differences with respect to this reference point.

Table 6 presents the results of a logistic regression. Columns (1) to (2) report the results using alternative measures of knowledge. Columns (3) and (4) include exam-specific dummies and group specific dummies. The penalty scoring rule always has a positive and significant coefficient estimate. The modified penalty scoring rule also has a positive coefficient estimate across specifications, but it is never significantly different from the reference point (reward scoring rule). These results indicate that for the decision whether to omit or not, subjects behave consistently with our theoretical results for risk averse individuals. Proposition 1 shows that the penalty and reward rules are not strategically equivalent for risk averse individuals, and the empirical evidence indicates that subjects behave differently when confronted to these scoring rules. These differences in behavior

gender-related differences in multiple-choice test scores (see Prieto and Delgado, 1999).

could be due to risk aversion, but also to a different framing. However, Proposition 3 shows that, despite the different framing, modified penalty and reward are strategically equivalent for all types of risk attitudes. The empirical evidence indicates that when confronted with these strategically equivalent rules, subjects do not behave in a significantly different manner, despite their different framing.

Table 6: Logistic Regression. Dependent Variable: Indicator (1) Omits (0) No Omits

	(1)	(2)	(3)	(4)
Penalty	0.650**	0.741***	0.684**	0.694***
	(0.266)	(0.274)	(0.268)	(0.264)
Modified Penalty	0.048	0.207	0.079	0.115
	(0.254)	(0.262)	(0.255)	(0.254)
Male	-0.795***	-0.876***	-0.888***	-0.869***
	(0.218)	(0.227)	(0.218)	(0.217)
Accumulated Score	0.116***	0.105***	0.049	0.124***
	(0.032)	(0.032)	(0.081)	(0.032)
Accumulated Score Squared	-0.005***	-0.005***	-0.003	-0.006***
	(0.001)	(0.001)	(0.002)	(0.001)
Knowledge	0.017			
	(0.399)			
Knowledge Square	-0.013			
	(0.029)			
Section 1	0.229	-0.036	0.096	0.181
	(0.303)	(0.321)	(0.308)	(0.304)
Section 2	0.976***	0.739**	1.066***	1.037***
	(0.350)	(0.366)	(0.348)	(0.345)
Section 3	-0.085	-0.259	-0.089	0.022
	(0.321)	(0.350)	(0.318)	(0.315)
Section 4	0.294	-0.131	0.336	0.322
	(0.494)	(0.536)	(0.480)	(0.482)
Knowledge 2		0.095		
		(0.124)		
Knowledge 2 Squared		-0.003		
		(0.002)		
Exam 2			0.941	
			(0.706)	
Exam 3			0.0712	
			(0.883)	
White				-0.206
				(0.261)
Yellow				0.447*
				(0.264)
Constant	1.022	0.110	0.535	0.465
	(1.406)	(1.600)	(0.328)	(0.344)
Log-Likelihood	-261.43	-246.02	-262.45	-263.64
Observations	471	441	480	480

Standard errors in parentheses. One, two and three asterisks represent ten, five and one percent significance level.

The logistic regressions reported in Table 6 include other covariates. Gender has a significant effect on the decision whether to omit, males omit less than females.

The effect of the accumulated score in previous exams has an inverted U shape. The probability of omitting at least one item is first increasing and beyond a point decreasing in the accumulated score in previous exams. Knowledge does not appear to have a significant effect no matter the way we measure it. The dummy *Section 2* is significant, which might capture differences in instructors' teaching expertise. The exam-specific dummies which capture unobserved exam-specific characteristics such as difficulty, are not statistically significant. Finally, the dummy corresponding to the Yellow group is marginally significant at the 10 percent. Recall that the color group indicates a particular order in the administration of the treatments and, as argued above, a particular order in the administration of the treatments might affect subjects behavior.

Next we analyze the influence of the scoring rules on the number of omissions. The dependent variable, the number of omitted items by a student in a particular exam is a count variable. Therefore, we use Poisson regression for inference. Table 7 reports Poisson regression estimates where, in addition to the regressors used in Table 6, we have included interaction terms between the scoring rules and the other covariates. The reason for including these interactions is to correct for the lack of randomization after the first exam. According to the treatment effects literature, under the assumption of conditional mean independence, regressions of the outcome variable on the treatment dummy are enlarged with a set of covariates and the interaction of the treatment dummy and the covariates. Column (1) includes all the covariates and interaction terms. Column (2) excludes the interactions between scoring rules and sections and Column (3) excludes the exam dummies. In all cases the penalty scoring rule has a negative and significant coefficient estimate, whereas the modified penalty also has positive coefficient but it is never significant. In accordance with the results reported in Table 5, subjects confronted with the penalty scoring rule omitted less items than those confronted with the reward scoring rule while those confronted with the modified penalty scoring rule did not omitted significantly more than those confronted with the reward scoring rule. Again, subjects behave differently when facing scoring rules that are not strategically equivalent. This difference in behavior could be explained by risk aversion or framing. However, subjects behaved similarly when facing strategically equivalent scoring rules, hence ruling out framing as an explanation of the differences in behavior.

In Table 6 the sign of the penalty coefficient is positive so that subjects are more likely to be omitters under penalty than under reward (see also the histogram in Figure 1). In Table 7, the coefficient of penalty is negative, which indicates that this scoring rule lowers the number of omissions (compared to reward). Thus, under penalty more subjects omit, but they omit less than under reward.

Table 7: Poisson Regressions. Dependent Variable: Number of Omissions

	(1)	(2)	(3)	(4) Male	(5) Female
Penalty	-1.159**	-1.566***	-1.808***	-0.602	-2.375***
	(0.586)	(0.555)	(0.519)	(0.939)	(0.687)
Modified Penalty	-0.252	-0.692	-0.541	-0.194	-0.556
	(0.509)	(0.430)	(0.412)	(0.833)	(0.441)
Male	-0.486***	-0.476***	-0.476***		
	(0.153)	(0.155)	(0.155)		
Accumulated Score	0.055*	0.056**	0.074***	0.100***	0.055***
	(0.029)	(0.028)	(0.017)	(0.032)	(0.019)
Accumulated Score Squared	-0.003***	-0.003***	-0.003***	-0.004***	-0.002**
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Knowledge	0.070	0.052	0.051	0.747**	-0.055
	(0.075)	(0.078)	(0.077)	(0.379)	(0.078)
Knowledge Squared	-0.018***	-0.019***	-0.019***	-0.065**	-0.012
	(0.006)	(0.007)	(0.006)	(0.026)	(0.008)
Exam 1	0.075	0.069			
	(0.333)	(0.327)			
Exam 2	0.296	0.267			
	(0.222)	(0.215)			
Section 1	0.212	0.140	0.149	0.091	0.185
	(0.222)	(0.123)	(0.123)	(0.197)	(0.146)
Section 2	0.456**	0.232*	0.244**	0.285	0.266*
	(0.214)	(0.124)	(0.123)	(0.219)	(0.137)
Section 3	0.444*	0.0834	0.0869	0.0935	0.112
	(0.230)	(0.142)	(0.143)	(0.215)	(0.186)
Section 4	0.387	0.202	0.202	-0.055	0.340*
	(0.313)	(0.155)	(0.159)	(0.339)	(0.180)
White	0.035	0.120	0.379*	0.254	0.454*
	(0.292)	(0.281)	(0.222)	(0.376)	(0.260)
Yellow	0.344	0.425	0.704***	0.568	0.745***
	(0.290)	(0.285)	(0.224)	(0.411)	(0.234)
Penalty * Male	0.369*	0.348*	0.348*		
	(0.199)	(0.201)	(0.201)		

Table 7: Poisson Regressions. Dependent Variable: Number of Omissions (Continued)

	(1)	(2)	(3)	(4) Male	(5) Female
Modified Penalty * Male	0.200	0.188	0.198		
	(0.219)	(0.220)	(0.221)		
Penalty * Accumulated Score	0.035	0.029	0.025	0.036	0.024
	(0.024)	(0.024)	(0.023)	(0.039)	(0.029)
Modified Penalty *	-0.019	-0.027	-0.048**	-0.038	-0.052**
Accumulated Score	(0.026)	(0.025)	(0.021)	(0.038)	(0.023)
Penalty * Knowledge	0.099	0.136**	0.140**	-0.018	0.231***
	(0.067)	(0.064)	(0.064)	(0.117)	(0.075)
Modified Penalty *	0.089	0.131*	0.143**	0.093	0.158**
Knowledge	(0.074)	(0.071)	(0.070)	(0.129)	(0.076)
Penalty * Exam 1	0.489	0.432	0.636	0.776	0.621
	(0.561)	(0.552)	(0.481)	(0.763)	(0.618)
Penalty * Exam 2	0.273	0.367	0.914***	0.814	0.930**
	(0.511)	(0.501)	(0.318)	(0.539)	(0.373)
Penalty * Section 1	-0.084				
	(0.297)				
Penalty * Section 2	-0.230				
	(0.280)				
Penalty * Section 3	-0.571*				
	(0.331)				
Penalty * Section 4	-0.290				
	(0.358)				
Modified Penalty * Section 1	-0.101				
	(0.312)				
Modified Penalty * Section 2	-0.415				
	(0.325)				
Modified Penalty * Section 3	-0.503				
	(0.342)				
Modified Penalty * Section 4	-0.244				
	(0.423)				
Constant	0.111	0.350	0.179	-2.627*	0.469
	(0.428)	(0.426)	(0.351)	(1.387)	(0.353)
Log-Likelihood	-640.81	-643.45	-644.26	-310.26	-327.19
Observations	471	471	471	240	231

Robust standard errors in parentheses. One, two and three asterisks represent ten, five and one percent significance level.

Table 7 also indicates that males tend to omit significantly less than females. To further investigate the effect of gender on omissions, Columns (4) and (5) report Poisson regressions for males and females respectively. In both cases the modified penalty is not significantly different from the reference point (reward scoring rule), indicating than neither males nor females are affected by the framing of the scoring rules. However, the penalty scoring rule dummy is significant for females (Column (5)) but not for males (Column (4)). A possible explanation of this result is that males behave as risk neutral individuals, who see penalty and reward as strategically equivalent, while females behave as risk averse individuals who do not consider the penalty and reward scoring rules as strategically equivalent.

Table 7 also reports coefficient estimates for the other covariates. The effect of the accumulated score in previous exams on omissions has an inverted U shape. Knowledge is a significant determinant of the number of omissions. Some of the dummies for sections and color groups are also significant determinants of the number of omissions. Finally, some interaction terms are significant. However, these interaction terms do not have a clear interpretation, as they are included in the regression to account for the lack of randomization after the first exam.

To sum up, the experimental results indicate that the behavior of examinees in multiple-choice tests is consistent with rationality and risk aversion and does not seem to be affected by framing. To interpret this result it is important to note, first, that the results come from a field experiment and subjects were participating in real exams so that their incentives to behave rationally were very high. Second, the scoring rules were known well in advance, so the subjects had enough time to think what would be optimal for them to do under any of the rules. If the stakes had not been so high or the rules had been announced by surprise just before the test, it is likely that framing would have played a more important role.

Concerning risk attitudes, our results confirm that individuals do not exhibit risk neutrality on average since behavior under penalty is significantly different from behavior under modified penalty. However, when we look at the results splitting the sample into males and females, males tend to behave as risk neutral individuals while females act as risk averse persons.

6 Concluding Remarks

In this paper we show that scoring rules that penalize for wrong answers and those that reward for omissions are not strategically equivalent for risk averse individuals, although in the psychometrics literature they have been considered equivalent, under the implicit assumption that individuals are risk neutral.

Our main research question is whether subjects behave rationally or they are affected by framing, for any type of risk attitudes. We designed the experiment to be able to test rationality, on the one hand, and the presence of risk aversion, on the other. We propose a modification of the penalty rule that makes the two scoring rules (penalty and reward) strategically equivalent for any type of risk attitude. By confronting students with equivalent rules with different framing (modified penalty and reward) and non-equivalent scoring rules with the same framing (penalty and modified penalty), it is possible to distinguish the effect of risk aversion from that of psychological factors related to the different framing of the rules (gains versus losses).

Our field experiment shows significant differences in students' behavior when they are evaluated with penalty for wrong answers and reward for omissions. In addition, we find no significant differences between modified penalty and reward scoring rules. This evidence is consistent with expected utility maximizing behavior of risk averse students. In addition to the scoring rule, other significant determinants of the number of omissions are the accumulated score in previous exams, knowledge and gender.

Our results may be of interest to examiners and theorists. First, this study has shown that risk aversion is an important factor in real exams. A useful implication of this result for examiners is that any scoring rule that penalizes for wrong answers or rewards for omissions does introduce a bias against risk averse students. As long as there is a link between risk attitudes and social group (gender, etc.) this issue may have some practical relevance. However, the solution is not necessarily the elimination of penalties or rewards since that would increase the random component of grades in multiple-choice tests.

Second, even though the optimality of scoring rules is beyond the scope of this paper, a better understanding of scoring rules, the incentives that they provide and students' reactions to penalties and rewards are likely to be relevant for the optimal way of designing multiple-choice tests. Our main empirical result is that risk aversion appears as an important factor in real exams, and therefore this variable should not be ignored in any study of the optimality of scoring rules.

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 instructions, each treatment had one more instruction regarding the scoring rule used for that treatment.

TREATMENT $\boldsymbol{S_P}$

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

$$\textit{score} = 2 \text{ x } (\textit{rights} - \frac{\textit{wrongs}}{3}) = (2 \text{ x } \textit{rights}) - (0.66 \text{ x } \textit{wrongs}).$$

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

Scoring Rule	
Right	+2
Wrong	-0.66
Omit	0

TREATMENT S_P^*

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

$$score = 5 + 1.5 \text{ x } (rights - \frac{wrongs}{3}) = 5 + (1.5 \text{ x } rights) - (0.5 \text{ x } wrongs).$$

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

Scoring Rule	
Right	+1.5
Wrong	-0.5
Omit	0

TREATMENT S_R

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

$$score = 2 \text{ x } (rights + \frac{omits}{4}) = (2 \text{ x } rights) + (0.5 \text{ x } omits).$$

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

Scoring Rule	
Right	+2
Wrong	0
Omit	+0.5

References

Becker, W. E. and C. Johnston, (1999), "The Relationship between Multiple Choice and Essay Response Questions in Assessing Economics Understanding," *The Economic Record*, 75, 348-357.

Bereby-Meyer, Y., J. Meyer, and O. M. Flascher, (2002), "Prospect Theory Analysis of Guessing in Multiple Choice Tests," *Journal of Behavioral Decision Making*, 15, 313-327.

Bereby-Meyer, Y., J. Meyer, and D. V. Budescu, (2003), "Decision Making under Internal Uncertainty: The case of Multiple Choice Tests with Different Scoring Rules," *Acta Psychologica*, 112, 207-220.

Bernardo, José M., (1998), "A Decision Analysis Approach to Multiple Choice Examinations," In: Girón, F. J. (ed.), *Applied Decision Analysis*, Boston, Kluwer, 195-207.

Bertrand, M., D. S. Karlan, S. Mullainathan, E. Shafir, and J. Zinman, (2005), "What's Psychology Worth? A Field Experiment in the Consumer Credit Market," *Working Papers 918*, Economic Growth Center, Yale University.

Bredon, G., (2003), "Take-Home Tests in Economics," *Economic Analysis and Policy, Queensland University of Technology, School of Economics and Finance*, 33, 52-60.

Budescu, D. and M. Bar-Hillel, (1993), "To Guess or Not to Guess: A Decision-Theoretic View of Formula Scoring," *Journal of Educational Measurement*, 30, 277-291.

Burgos, A., (2004), "Guessing and Gambling," *Economics Bulletin*, 4, 1-10.

Byrnes, J. P., D. C. Miller, and W. D. Schafer, (1999), "Gender Differences in Risk Taking: A Meta-Analysis," *Psychological Bulletin*, 125, 367-383.

Cadsby, C. B. and E. Maynes, (2005), "Gender, Risk Aversion, and the Drawing Power of Equilibrium in an Experimental Corporate Takeover Game," *Journal of Economic Behavior and Organization*, 56, 39-59.

Chan, N. and P. Kennedy, (2002), "Are Multiple-Choice Exams Easier for Economics Students? A Comparison of Multiple-Choice and 'Equivalent' Constructed-Response Exam Questions," *Southern Economic Journal*, 68, 957-

971.

Espinosa, M. P. and J. Gardeazabal, (2010), “Optimal Correction for Guessing in Multiple-Choice Tests,” *Journal of Mathematical Psychology*, 54, 415-425.

Kahneman, D. and A. Tversky, (1979), “Prospect Theory: An Analysis of Decision under Risk,” *Econometrica*, 47, 263-291.

Haan, M., B. Los, Y. Riyanto, and M. V. Geest, (2002), “The Weakest Link - A Field Experiment in Rational Decision Making,” University of Groningen, Research Institute Systems, Organisations and Management, Research Report number 02F20.

Haigh, M. S. and J. List., (2005), “Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis,” *Journal of Finance*, 60, 523-534.

Heck, J. L. and D. E. Stout, (1998), “Multiple-Choice vs. Open-Ended Exam Problems: Evidence of Their Impact on Student Performance in Introductory Finance,” *Financial Practice and Education*, 8, 83-93.

List, J. and D. 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.

McKenzie, C., (2003), “Rational Models as Theories - Not Standards - of Behavior,” *Trends in Cognitive Sciences*, 7, 403-406.

Prieto, G. and A. 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.

Scotchmer, S., (2008), “Risk Taking and Gender in Hierarchies,” *Theoretical Economics*, 3, 499-524.

Siegfried, J. J., P. Saunders, E. Stinar, and H. Zhang, (1996), “How is Introductory Economics Taught in America?,” *Economic Inquiry*, 34, 182-192.

Traub, R. E., R. K. Hambleton , and B. Singh, (1969), “Effects of Promised Reward and Threatened Penalty on Performance of a Multiple-Choice Vocabulary Test,” *Educational and Psychological Measurement*, 29, 847-861.

Traub, R. E. and R. 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.

Walstad, W. B. and W. Becker, (1994), “Achievement Differences on Multiple-Choice and Essay Tests in Economics,” *The American Economic Review*, 84, 193-196.

Waters, C. W. and L. K. Waters, (1971), “Validity and Likability Ratings for Three Scoring Instructions for Multiple-Choice Vocabulary Tests,” *Educational and Psychological Measurement*, 31, 935-938.