ISSN 1988-088X


Department of Foundations of Economic Analysis II
University of the Basque Country
Avda. Lehendakari Aguirre 83
48015 Bilbao (SPAIN)
http://www.ehu.es/FAEII

# *DFAE-II WP Series*

2007-08

MARÍA PAZ ESPINOSA & JAVIER GARDEAZABAL

*Optimal Correction for Guessing in Multiple-Choice Tests*

# Optimal Correction for Guessing in Multiple-Choice Tests

María Paz Espinosa and Javier Gardeazabal
University of the Basque Country

**Abstract**

Building on Item Response Theory we introduce students' optimal behavior in multiple-choice tests. Our simulations indicate that the optimal penalty is relatively high, because although correction for guessing discriminates against risk-averse subjects, this effect is small compared with the measurement error that the penalty prevents. This result obtains when knowledge is binary or partial, under different normalizations of the score, when risk aversion is related to knowledge and when there is a pass-fail break point. We also find that the mean degree of difficulty should be close to the mean level of knowledge and that the variance of difficulty should be high.

"Religion, politics and formula scoring are areas where two informed people often hold opposing ideas with great assurance." (Lord, 1975, p. 7)

Multiple-choice tests have some advantages over constructed-response tests, such as a wider sampling of content and the prevention of grading errors. On the other hand, guessing behavior on the part of examinees affects the scores obtained in these tests. To avoid this disadvantage of multiple-choice tests, a correction for guessing formula is commonly used. Penalizing incorrect answers reduces the incentive for guessing and subjects may leave items unanswered.

Educational and psychological researchers use Item Response Theory (IRT) to evaluate tests and item scores based on the mathematical relationship between abilities and item responses. Building on this framework, we propose a theory of students' behavior in multiple-choice tests that explains why they might find it optimal to answer some items and leave others blank. We analyze whether formula scoring discriminates against risk-averse subjects. The model developed in this paper is solved by simulation and allows us to make certain claims about the optimal design of multiple choice tests.

---

Financial support from Spanish Ministry of Education and Science (grant SEJ2006-06309) is acknowledged. Mailing address: Dpto. Fundamentos del Análisis Económico II, Avda. Lehendakari Aguirre 83, 48015 Bilbao, Spain.
E-mail: mariapaz.espinosa@ehu.es javier.gardeazabal@ehu.es

Suppose a test has $N$ items and each item has $M$ possible choices. The score, $s$, is a function of the number of rights, $r$, wrongs, $w$, and the penalty, $p$: $s = r - pw$ with $p = \frac{1}{M-1}$. If knowledge was binary, students either select the right answer or pick one randomly. The expected value of a pure guess is equal to the expected value of omitting.

It is typically argued that the correction for guessing formula eliminates the measurement error induced by guessing. However, the use of formula scoring has been extensively criticized. Here we point out two criticisms. First, it has been argued that the assumption that students either know or do not know the answer is very strong, as many students do have partial knowledge and are able to rule out one or more distractors. This argument is backed up by studies (e.g., Bliss, 1980) showing that students who omit under formula scoring would have chosen more correct answers than under chance expectations. This point is best explained in the context of partial knowledge, that is, when a student is able to rule out some alternatives but cannot determine the correct answer from among the remaining ones. Under partial knowledge and with a penalty $\frac{1}{M-1}$, the expected value of guessing is greater than the value of omitting so it could be argued that the penalty does not really prevent guessing. Even in the case of no knowledge, the expected values of guessing and omitting are equal, so there is no assurance that the student will not guess. Second, under formula scoring it is typically recommended that instructors advise students to answer all questions on which one or more answers can be eliminated and omit those on which no answer can be ruled out. When the exam has a pass-fail threshold and the student's knowledge is very low, the best strategy is to guess. The instructor is thus giving advice that some students should ignore for their own benefit. With these caveats in mind, it is not surprising that many practitioners recommend using not formula scoring but number-right scoring, that is, setting the score equal to the number of right answers by using a penalty of zero.

All the discussion has centered on two values of the penalty: those favoring number right scoring recommend $p = 0$ while those favoring formula scoring recommend $p = \frac{1}{M-1}$. Both sides have good arguments in support of their positions but surprisingly no one, to the best of our knowledge, has suggested a different value for the penalty. To reach a conclusion on the optimal value of the penalty, it is necessary to nest both types of scoring rule within a more general framework where number-right and formula scoring are just two particular cases.

In this paper we discuss the following trade-off: under no penalty (or a low penalty) for wrong answers, students have an incentive to guess, thus increasing measurement error, but a high penalty does not discourage guessing uniformly: students with high risk aversion are more easily discouraged, so penalties introduce a bias in favor of risk takers. Therefore, if risk aversion is correlated with gender, knowledge or social group, a high penalty would discriminate against these groups of students, e.g., Prieto and Delgado (1999).

Most papers assume that the test taker is an expected score maximizer (e.g., Budescu & Bar-Hillel, 1993), but in many situations individuals behave as risk averse. Bliss (1980) and Albanese (1986, 1988) provide indirect evidence showing that examinees answered more questions under number-right scoring than under formula scoring and that they answered the added items correctly at a rate beyond chance. This is not consistent

with expected score maximizers, but rather with risk-averse test takers who omit questions with a positive expected score to avoid risk. Other possible explanations are the influence of directions given before the exam (e.g., Budescu & Bar-Hillel, 1993), framing of the scoring rule (e.g., Bereby-Meyer, Meyer, & Flascher, 2002) and personality (e.g., Avila & Torrubia, 2004).

If the goal is to maximize the expected score students will answer whenever the expected value of answering turns out to be higher than or equal to the payoff for omitting. This behavior corresponds to a risk-neutral agent, i.e. someone who is indifferent between a gamble with expected value $x$ and getting $x$ for sure. This behavioral assumption is to say the least questionable, since individuals do pay for risk avoidance in numerous real situations, for instance, when they buy insurance. We argue that depending on the students' attitudes towards risk, the optimal penalty need not be the typically used $\frac{1}{M-1}$ value nor zero. In fact, our simulations show that the optimal penalty is above $\frac{1}{M-1}$.

## A binary knowledge model

In a binary knowledge model students either know the answer to a question for sure or cannot discriminate between the different alternatives. This assumption is extreme but useful for establishig a benchmark. In this case we can define the true score as the number of questions to which the examinee knows the answer, while the observed score is the score obtained in the test. The examiner would like the grading procedure to elicit the true score or at least order students according to it.

A penalty for wrong answers affects the correlation between the observed score and the true score by introducing a trade-off between measurement error and bias illustrated in the following example. Consider an exam with two questions and two alternatives each. There is a population of $T$ students, all of them with exactly the same knowledge. All students know the answer to the first question and have no clue about the second one. Half of them are risk-neutral and derive utility from their score $u(s) = 2+s$. The other half are risk-averse and have a utility function given by $u(s) = \sqrt{2+s}$. All students maximize the expected utility of the score. The penalty for a wrong answer is $p = 0.9$ while a right answer yields 1 point. Students consider whether to answer the question or not. The risk-neutral students find that guessing yields a higher expected payoff than omitting: $\frac{1}{2}(3+1)+\frac{1}{2}(3-p) > 3$. However, the risk-averse students will not answer the question, since the expected utility of guessing is lower than the utility of omitting: $\frac{1}{2}\sqrt{(3+1)}+\frac{1}{2}\sqrt{(3-p)} < \sqrt{3}$.

In this example, the examiner faces two problems due to different attitudes of students towards risk and guessing behavior. On the one hand, expected scores are not equal for all students who have the same knowledge. In fact, scores are biased against risk-averse students. On average, risk-neutral students obtain higher scores than risk-averse ones (1.05 vs. 1.00). On the other hand, actual scores will exhibit variance even though all students have the same knowledge; this is the measurement error due to guessing. We would like students with the same knowledge to get the same score. In this example, on average $\frac{T}{4}$ students get a score of 2, $\frac{T}{2}$ students a score of 1 and $\frac{T}{4}$ a score of 0.1: grades do not reflect

knowledge. In general, it is not desirable for grades to have such high variance for equally knowledgeable students.

If we make the penalty zero, bias disappears: all students guess and all students have the same expected score. In this case risk-averse students are not discriminated against. However, the variability of actual scores remains: on average $\frac{T}{2}$ students will guess the right answer and obtain a final score of 2 and $\frac{T}{2}$ students will get it wrong and obtain a final score of 1. This dispersion of grades is due to guessing.

In this example the dispersion of grades, as measured by the standard deviation, decreases with the penalty level and falls to zero for penalty levels higher than 1. It is possible to eliminate the measurement error by making the penalty high enough so that nobody guesses and all students obtain the true score. Unfortunately, as we will see in the next section, a high penalty is not the solution when there is a probability of making a mistake when the student knows the right answer or when there is partial knowledge.

We now formalize these ideas in a binary knowledge model.

**The test**. The test has $N$ items. Each item has $M$ possible answers, one correct and $M-1$ incorrect. Item $i$ has a degree of difficulty $b_i$, a real number. The penalty for each incorrect answer is $p$. Denote the vector of difficulties $b = (b_1, ...., b_N)'$. Therefore, a test is completely characterized by the set $\{N, M, p, b\}$.

**Students**. There is a population of $T$ students who may differ on the level of knowledge or ability, a real number $\theta_t$, and the utility they derive from the score, as given by a continuous function $u_t(s_t)$, strictly increasing in the score, $u_t'(s_t) > 0$ and concave $u_t''(s_t) \leq 0$. Therefore, a student is completely characterized by $\{\theta_t, u_t(s_t)\}$. When the utility function is strictly concave the student is risk-averse. When the utility function is affine, we say the student is risk-neutral.

**Binary knowledge**. For each question $i$, student $t$ observes the difference $\theta_t - b_i$. The students knows the right answer if and only if $\theta_t - b_i > 0$. The difference $(\theta_t - b_i)$ represents the outcome of confronting the student's ability or knowledge $\theta_t$ and the item's difficulty $b_i$.

**True score**. For a given exam $\{N, M, p, b\}$ and student $\{\theta_t, u_t(.)\}$, define the true score, $k_t$, as the number of questions to which examinee $t$ knows the answer, $k_t = \sum_{i=1}^{N} 1(\theta_t - b_i > 0)$ where $1(A)$ is equal to one when $A$ is true and zero otherwise.

**No mistakes**. Examinees do not make mistakes. When a student knows the answer, the probability of getting the right answer is one; when the student does not know the answer, the probability is $\frac{1}{M}$.

**Score**. The score is written as $s_t = r_t - pw_t$. Let $n_t$ be the number of answered questions. Since the number of answers equals the number of rights plus the number of wrongs, $n_t = r_t + w_t$, the score can be written as $s_t = (1+p)r_t - pn_t$.

**Utility maximization**. Students answer as many questions as necessary to maximize the expected value of the utility derived from the score. They choose the number of answers $n_t$ to maximize $E(u_t(s_t(n_t)))$. This assumption includes expected score maximizers as a special case, but also includes other types of behavior which are usually left out of the analysis.

**Objective**. The examiner sets the penalty $p$, and the distribution of difficulties $b_i$ with the objective of maximizing Pearson's correlation between knowledge and score

$$\rho_P(\theta,s) = \frac{T\sum_{t=1}^{T}(s_t-\overline{s})(\theta_t-\overline{\theta})}{\sqrt{\sum_{t=1}^{T}(s_t-\overline{s})^2}\sqrt{\sum_{t=1}^{T}(\theta_t-\overline{\theta})^2}}.$$

where $\overline{s}$ and $\overline{\theta}$ are the sample means of the score and knowledge. Maximizing the correlation between knowledge and the score is a sensible objective. However, if the examiner is interested in ranking the students according to their knowledge, maximizing Spearmans' correlation coefficient

$$\rho_S(s_t,\theta_t) = 1 - \frac{6\sum_{t=1}^{T}(rank(s_t)-rank(\theta_t))^2}{T(T^2-1)}$$

would be the appropriate objective.

Under these conditions the examiner may obtain a perfect correlation between the score and the true score.

**Proposition 1** If $p > \frac{1}{M-1}$ then $\rho_P(k,s) = 1$.

**Proof** If the student answers only the $k$ questions for which $\theta_t - b_i > 0$, the score is $s(k) = k$. Answering one more question yields an expected score:

$$E(s(k+1)) = k + \frac{1}{M} - \frac{M-1}{M}p.$$

First, consider risk neutral students. If $p > \frac{1}{M-1}$, then $s(k) > E(s(k+1))$ and therefore a risk-neutral student, i.e. an expected score maximizer, will not answer the $(k+1)$-th question. Second, consider a risk-averse student. Since, utility is concave for risk averse students, it follows from Jensen's inequality that $E(u(s(k+1))) < u(E(s(k+1)))$. Since utility is strictly increasing, $u(E(s(k+1))) < u(s(k))$. Thus, risk-averse students will not answer the $(k+1)$-th question either. All examinees will answer only the questions they know and omit the questions they do not know. Hence, the observed score, $s = k$, is perfectly correlated with the true score, $k$, i.e. $\rho_P(k,s) = 1$. QED

Note that a penalty $\frac{1}{M-1}$ would yield $\rho_P(k,s) = 1$ when $u''(s) < 0$ and a penalty lower than $\frac{1}{M-1}$ could also achieve the objective of maximizing the correlation between the true score and the observed score as long as all students are sufficiently risk-averse. However, $p > \frac{1}{M-1}$ achieves the objective for any degree of risk aversion. Since individuals do not make mistakes and the penalty is never enforced, there is no loss in setting $p > \frac{1}{M-1}$.

The previous proposition refers to the correlation between the true score $k$ and the observed score $s$. The following propositions show results on the correlation between knowledge $\theta$ and the score $s$.

**Proposition 2** If $p > \frac{1}{M-1}$, students are labeled according to knowledge $\theta_1 < \theta_2 < ... < \theta_T$ and the exam has at least $T-1$ questions with difficulties $b_1 < b_2 < ... < b_{T-1}$ such that $\theta_1 < b_1 < \theta_2 < b_2 < ... < b_{T-1} < \theta_T$, then $\rho_S(\theta,k) = 1$.

**Proof** Consider first an exam with exactly $T-1$ questions. With $p$ greater than $\frac{1}{M-1}$, all students answer all the questions that they know and leave the rest blank. Our choice of difficulties yields scores $s_t = t-1$, therefore $rank(s_t) = t = rank(\theta_t)$ for all $t$. Thus, $\rho_S(\theta,k) = 1$. Next, consider exams with more than $T-1$ questions. First, consider an exam with $T$ questions. All students such that $\theta_t > b_T$ will answer item $T$ and get one extra point and all students with $\theta_t < b_T$ will omit the item. Therefore, adding more questions to the original exam changes grades but it does not change the ranking of students or Spearman's rank correlation. The argument is exactly the same for exams with $T+1$ questions or more. QED

**Proposition 3** If $p > \frac{1}{M-1}$, $T$ is fixed and difficulties $b_i$ are random draws from a distribution with the same support as the distribution of $\theta_t$, then as $N$ tends to infinity $\rho_P(\theta,s)$ tends to one.

**Proof** From Proposition 1, students answer only the items they know. Let $b^{max}_{N,t}$ be the maximal element of the set $\{b_i | b_j < \theta_t, j = 1,...,N\}$. Since the support of $b_i$ and $\theta_i$ are the same, $..... \geq \theta_t - b^{max}_{N-2,t} \geq \theta_t - b^{max}_{N-1,t} \geq \theta_t - b^{max}_{N,t} \geq 0$. Therefore, as $N$ tends to infinity, $b^{max}_{N,t} - \theta_t$ tends to 0. Since $T$ is fixed, this is true for all students. Thus $\rho_P(\theta,s)$ tends to one. QED

To sum up, in the binary knowledge model the examiner can reach the maximum correlation between scores $s$ and knowledge $\theta$ as long as the exam has a high enough number of questions and the penalty is above $\frac{1}{M-1}$. As we will see in the next section this objective is harder to achieve in the case of partial knowledge.

## A partial knowledge model

The assumption of binary knowledge is convenient but unrealistic. Quite often examinees can rule out some of the alternatives but they are not sure which of the other alternatives is the right answer. In order to capture this possibility we make use of a latent variable model. For student $t$ and item $i$ define a latent variable $y_{ti} = \theta_t - b_i - v_{ti}$, where $v_{ti}$ is a zero mean random variable with distribution function $F(v_{ti})$. The probability that student $t$ knows the answer to item $i$ is $P(y_{ti} \geq 0) = F(\theta_t - b_i)$. Let us define a random variable $z_{ti}$ such that $z_{ti} = 1$ if item $i$ is correctly answered and equal to zero otherwise.

**Partial knowledge** The probability of answering correctly item $i$ is $P(z_{ti} = 1) = c_i + (1-c_i)P(y_{ti} \geq 0)$ where $c_i$ can be interpreted as the probability of a pure guess.

**Increasing difficulty**. Items are ordered in increasing degree of difficulty, i.e. $P(y_{t1} \geq 0) \geq P(y_{t2} \geq 0) \geq ..... \geq P(y_{tN} \geq 0)$. We are assuming that the order of items does not affect students' results. This is consistent with the results obtained in the experiment carried out by McLeod, Zhang, and Yu (2003).

Students have to decide when it is optimal to stop answering items. The score obtained by student $t$ can be written as $s_t = (1+p)\sum_{i=1}^{n_t} 1(z_{ti} = 1) - pn_t$. Therefore, the expected score is $E(s_t) = (1+p)\sum_{i=1}^{n_t} P(z_{ti} = 1) - pn_t$.

First we explore the optimal choice when the student is an expected score maximizer. Suppose the student has already answered item $n_t - 1$ and faces the problem of deciding whether to answer item $n_t$ or not. If the goal is to maximize the expected score, the student

will answer question $n_t$ if the expected value of answering the $n_t$ questions is greater than the expected value of answering $n_t - 1$ questions, that is $E(s(n_t)) > E(s(n_t - 1))$. It is easy to verify that this occurs when $P(z_{tn_t} = 1) > \frac{p}{1+p}$. In other terms, an expected score maximizer will answer any question $i$ for which the following inequality holds

$$\frac{P(z_{ti} = 1)}{1 - P(z_{ti} = 1)} > p. \tag{1}$$

The left-hand side of (1) is the odds ratio, that is, the probability of answering item $i$ right over the probability of answering it wrong. An expected score maximizer will answer question $i$ when the odds ratio is greater than the penalty parameter. Notice that for a student with no knowledge, $P(z_{ti} = 1) = \frac{1}{M}$ and therefore the odds ratio is the critical point $\frac{1}{M-1}$.

The optimal decision of a risk-averse student, however, is not as simple as that of a risk neutral student. A risk-averse examinee will compare the expected value of the utility of answering item $n_t - 1$ with the expected value of the utility of answering one more item. The student will answer item $n_t$ if

$$E\left(u\left((1+p)\sum_{i=1}^{n_t} 1(z_{ti} = 1) - pn_t\right)\right) > E\left(u\left((1+p)\sum_{i=1}^{n_t-1} 1(z_{ti} = 1) - p(n_t - 1)\right)\right).$$

This optimality condition does not provide us with a simple rule such as (1).

In a partial knowledge model, whether the student knows the answer is not well defined. This prevents examiners from getting a perfect ordering of the students according to their true score, as was possible in the case of binary knowledge. With partial knowledge it is not possible to characterize the optimal test so easily. In the following sections we provide a numerical solution.

## Numerical procedure

In this section we provide a numerical solution to the problem of selecting the penalty and other features of the test to maximize Pearson's correlation between knowledge and score. First, we make several parametric assumptions. Then, we describe a procedure for obtaining a numerical solution.

**Extended power utility** Students choose how many items to answer to maximize expected utility

$$E(u(s)) = E\left(\frac{(a+s)^{1-\phi}}{1-\phi}\right),$$

where $a > 0$, $\phi > 0$ and $u(s) = \ln(a+s)$ if $\phi = 1$.

**Log-normal risk aversion**. The distribution of the risk aversion measure $\phi$ among the population of students is log-normal with parameters $(\mu, \sigma^2)$ and independent of knowledge.

**Normal knowledge**. The distribution of knowledge, $\theta$, among the population of students is normal with parameters $\left(\mu_1, \sigma_1^2\right)$ and independent of risk aversion.

**Normal difficulty**. The examiner can set exams whose difficulties are random draws from a normal distribution with parameters $\left(\mu_2,\ \sigma_2^2\right)$.

**Logistic noise** Following Item Response Theory (IRT), we assume that the distribution of $v_{ti}$ is logistic. The three-parameter-logistic item-response-model assumes that the probability of a right answer is

$$P(z_{ti}=1) = c_i + \frac{1-c_i}{1+\exp(-(\theta_t - b_i)/g_i)} \tag{2}$$

where the variance of $v_{ti}$ is $\frac{\pi^2}{3}g_i$ and $g_i$ is inversely related to the item's discrimination. The higher the value of $g_i$ the lower the relevance of the difference between knowledge and difficulty in determining the probability of knowing the answer to the item.

In the numerical solution we rescale the score by adding $pN$ and dividing by $(1+p)N$, so that it takes values in the interval $[0,1]$. The score of student $t$ is

$$s_t = \left(\frac{r_t - pw_t}{N} + p\right)\left(\frac{1}{1+p}\right). \tag{3}$$

Since the normalized score is a linear function of the original score, this normalization does not alter the correlation between knowledge and the score.

For a given student $\{\theta_t, \phi_t\}$ and test $\{N, M, p, b\}$ the solution to the student's optimization problem can be found as follows:

1. First we compute the set of possible values of the score. For instance, with two questions, $N=2$, the set of possible values of the normalized score is $\{s(0,0), s(0,1), s(1,1), s(0,2), s(1,2), s(2,2)\} = \left\{\frac{p}{1+p}, \frac{p}{2(1+p)}, \frac{1+2p}{2(1+p)}, 0, \frac{1}{2}, 1\right\}$, where $s(r,n)$ is the score with $r$ rights out of $n$ answers.

2. Second, we compute the probabilities associated with each value of the score by first computing the probability of answering correctly item $i = 1, ....., N$, using equation 2 and then the probability of getting $r$ rights out of $n$ answers

$$P(r,n) = P(r,n-1)P(z_{tn}=0) + P(r-1,n-1)P(z_{tn}=1)$$

for $r = 1, ..., n$, $n = 1, 2, ..., N$ and $P(0,n) = P(0,n-1)P(z_{tn}=0)$.

3. Third, we compute the value of $n$ that maximizes expected utility

$$E(u(s(n))) = \sum_{j=1}^{n} P(j,n)u(s(j,n)).$$

To do this we simply evaluate the expected utilitiy at all posible values $n = 1, ..., N$.

## Numerical results

This section presents the results of the simulation. We analyze the relationship between omissions and knowledge by solving the problem for three students whose risk aversion parameter is $\phi = 20.09$ (the median of the distribution of $\phi$ used later) and who have three different values of knowledge $\{-1.28, 0, 1.28\}$ (the 0.1, 0.5 and 0.9 quantiles of the distribution of $\theta$ used later). The utility function parameter $a$ is set to unity for all students. We set these three students the same exam with $N = 100$ questions and $M = 4$ alternatives each, whose difficulties $b_i$ are random draws from a normal distribution with parameters $\mu_2 = 0$ and $\sigma_2^2 = 1$. The item parameter $g_i$ is fixed at $\frac{3}{\pi^2}$, hence $v_{ti}$ has unit variance. This is repeated for penalty values in the interval $[0, 2]$. Figure 1 graphs the optimal number of omissions. There are three aspects of this graph that deserve attention. First, the number of omissions increases with penalty and decreases with knowledge. Second, all three students are equally risk-averse but start omitting at different penalties. The two least knowledgeable ones start omitting for penalties below the critical value $\frac{1}{M-1} = 0.33$ whereas the most knowledgeable student starts to omit at penalty values higher than 0.33. Third, for the least knowledgeable student and the average one, omissions increase sharply for penalty values in the interval $[0.3, 0.4]$ and then for higher penalty values omissions increase at a lower rate.

Next we focus on the relationship between omissions and risk aversion. Figure 2 graphs the optimal number of omissions for three students with the same level of knowledge $\theta = 0$ and three different degrees of risk aversion, the risk neutral case with $\phi = 0$, the median value of the distribution of risk aversion coefficients $\phi = 20.09$ and a highly risk-averse individual with $\phi = 100$. The difference in omissions between the most risk-averse and the least is zero for penalties below 0.23; it increases to about 20% for a penalty value of 0.33 approximately and then it decreases to about $5\% - 8\%$ and remains at that level for higher penalties. Comparing Figures 1 and 2 we can conclude that risk aversion influences the number of omissions, but knowledge influences it much more.

*The role of difficulty and penalty*

Next we simulate 30 exams using the parameters of Table 1.

Table 1: Parameter values used in the simulations.

| | | | | | | $\phi_t$ | | $\theta_t$ | | $b_i$ | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $T$ | $N$ | $M$ | $a$ | $c_i$ | $g_i$ | $\mu$ | $\sigma$ | $\mu_1$ | $\sigma_1$ | $\mu_2$ | $\sigma_2$ |
| 100 | 100 | 4 | 1 | 0.25 | $\frac{3}{\pi^2}$ | 3 | 1 | 0 | 1 | 0 | 1 |

For each exam we compute the correlation between the knowledge of each student and the score obtained in the exam and then compute their average. We explore how the average correlation between score and knowledge depends on the degree of difficulty and the penalty, first within the binary knowledge model and then in the partial knowledge model.

The binary knowledge model is a special case of the partial knowledge model where

the variance of the disturbance term is zero. Therefore, we simulate the binary knowledge model setting $g_i = 0$ for all $i$. Figure 3 shows the contour lines of the correlation surface for the binary knowledge model when we let the penalty take values in the interval $[0,2]$ and the mean of the distribution of difficulty, $\mu_2$, takes values in the interval $[-2,2]$. Correlation between knowledge and score is maximized when the penalty is equal to or greater than $\frac{1}{M-1} = 0.33$ and when the mean of the distribution of difficulties is in the neighborhood of zero. These results are consistent with Propositions 1 and 3: correlation is maximized for penalties greater than or equal to $\frac{1}{M-1}$ but it does not reach one because the number of items in the exam is finite.

Next we consider the partial knowledge model by setting the variance of the disturbance term $v_{ti}$ to unity. Figure 4 plots the iso-correlation curves. As was the case with the binary knowledge model, maximum correlation between knowledge and score is attained for values of $\mu_2$ in the neighborhood of zero, that is, the mean value of the distribution of knowledge. Unlike the binary knowledge model, correlation between knowledge and the score decreases if the penalty is set too high. Nevertheless, the optimal values of $p$ are quite high: between 0.33 and 1.

**Result 1** A relatively high penalty is better than using number-right scoring when the objective is to attain a high correlation between knowledge and the score.

This result is in sharp contrast with the belief that number-right is superior to formula scoring, (e.g., Bar-Hillel, Budescu, & Attali, 2005).

When the variance of the disturbance term $v_{ti}$ is set higher, which means that the discrimination parameter $g_i$ is also higher, the results are very similar. Notice that the correlation between score and knowledge is maximized for values of the mean difficulty near the value of the mean knowledge. We have verified that this result also obtains for different values of mean knowledge.

**Result 2** The maximum correlation between knowledge and score is attained when the mean value of difficulty is near the mean value of knowledge.

So far we have assumed that risk aversion and knowledge are independent. If we allow risk aversion to be related to the level of knowledge the correlation between the score and knowledge remains very much the same.

**Result 3** Whether risk aversion is related to knowledge does not matter for maximizing correlation between score and knowledge.

We now relax the assumption that both knowledge and difficulty have unit variance. Figure 5 shows contour lines of the correlation surface when the variance of the distribution of difficulty is high, $\sigma_2^2 = 2$ and Figure 6 shows the iso-correlation lines when the variance of difficulty is low, $\sigma_2^2 = 0.5$. When the variance of the distribution of difficulty is high, changes in the mean, $\mu_2$, have less effect on the level of correlation than when it is low.

**Result 4** Correlation between score and knowledge is higher when the variance of difficulty is high.

*Are risk averse individuals discriminated against?*

Some researchers and practitioners recommend not penalizing for wrong answers. They claim that psychological factors or personal characteristics may influence the decision of students to omit questions on which they have partial knowledge and a positive expected reward from answering. Our analysis shows that penalizing for wrong answers does have an effect on the number of items omitted, as shown in Figure 1. Specifically, more risk-averse students omit more than less risk-averse ones. Risk-averse students omit items with positive expected score, so their expected score should be lower than the expected score of risk-neutral students who answer those items with positive expected gain. However, we show next that the difference in scores obtained by risk-averse and risk-neutral students is not large. To show this we simulate the scores obtained by two individuals: a risk-neutral individual with parameter $\phi = 0$ and a risk-averse one with parameter $\phi = 100$ both with the same level of knowledge, a low value of $-1.28$, in 500 exams of 100 questions each. Figure 7 shows the score plotted against the value of penalty. The scores of the risk-neutral and risk-averse individuals are very similar. Nevertheless, when penalty values are in the interval $[0.2, 0.4]$ the score of the risk-averse individual is significantly lower than the score of the risk-neutral one. Although significant, this discriminatory difference is quantitatively small, peaking at 3.1 per cent of the score when penalty is 0.3 (at this point the more risk-averse individual omits 69.3 percent more items than the less risk-averse one). We have repeated this simulation with more knowledgable students, but then the difference in omissions between the risk-averse and the risk-neutral individual decreses and so does the difference in the score. Therefore, it is fair to say that formula scoring does not discriminate against knowledgeable, risk-averse examinees and only discriminates against risk-averse examinees with little knowledge. More importantly, the difference in the score almost dissapears for penalty values either smaller or larger than the typically used $\frac{1}{M-1} = 0.33$.

**Result 5** Formula scoring discriminates against risk-averse students. This discrimination is quantitatively important for examinees with little knowledge, but almost negligible for students with high or average levels of knowledge.

## Robustness

The results reported so far rely on several assumptions, three of which are relaxed in this section to see whether the results are robust. Although the theoretical results were obtained with a non-normalized score, the numerical results use a normalized formula. The linear rescaling of equation (3) does not affect the value of the correlation; however, a simpler normalization truncates the score at zero:

$$s_t = \max\left\{\frac{1}{N}\left(r_t - pw_t\right), 0\right\}.$$

We have verified that the correlation surface is almost unaffected by this choice of normalization.

As a second robustness check we see whether using Spearman's rank-correlation coefficient affects the results obtained so far. Maximizing the rank-correlation is a reasonable objective when the examiner only wants to rank students and does not care about the specific measure of knowledge. For instance, this would be the right objective if we wished to select a fixed number of people from among a group of candidates. In this case, the correlation between knowledge and the score is slightly higher than when using Pearson's correlation coefficient, with the optimal penalty and mean difficulty being very similar.

As a third robustness check we consider exams with a passing grade. So far we have assumed the exam has no pass-fail break point. In practice many exams require students to reach a minimum passing score. It is reasonable to argue that students value all scores below the passing point as zero. We will assume that utility takes the functional form

$$u(s) = \begin{cases} \frac{a^{1-\phi}}{1-\phi} & if \quad s < \underline{s} \\ \frac{(a+s)^{1-\phi}}{1-\phi} & if \quad s \geq \underline{s} \end{cases}$$

where $\underline{s}$ is the passing score. Under this assumption, students value any score below $\underline{s}$ at $u(0)$ and utility is increasing in the score only if $s \geq \underline{s}$. Figure 8 shows the contour lines of the correlation surface when exams have a passing score of $\underline{s} = 0.75$. The correlation exhibits its maximum at higher penalty levels than when no passing grade is imposed. The reason is that less knowlegable students will answer more questions than without a passing score.

Next we consider pass-fail exams. Bernardo (1998) analyzes the case of exams to select a group of people, such as the MIR exam for entry to graduate medical school in Spain. He argues that exam takers might reasonably behave so as to maximize the probability of attaining the minimum passing score, which is equivalent to minimizing the probability of failing the exam. This behavior can also be accommodated in our model by assuming there are only two outcomes, pass and fail, so that the utility function is:

$$u(s) = \begin{cases} u_f & if \quad s < \underline{s} \\ u_p & if \quad s \geq \underline{s} \end{cases}$$

where $u_f < u_p$. Therefore expected utility is

$$E(u(s)) = P(s < \underline{s})u_f + P(s \geq \underline{s}))u_p = u_f + P(s \geq \underline{s})(u_p - u_f).$$

Since $u_f < u_p$ and both are fixed values, maximizing expected utility is equivalent to maximizing the probability of passing the exam, $P(s \geq \underline{s})$. The correlation between the score and knowledge is very similar to that reported in Figure 8.

**Result 6** The existence of a passing grade or pass-fail scoring induces risk-loving behavior on the part of examinees. To compensate for this behavior, the optimal penalty is higher.

## Conclusions

Combining basic IRT and the theory of decision under uncertainty, we have developed a model of students' optimal behavior. The model can be simulated and allows us to draw some conclusions on the optimal design of a multiple-choice test. The simulations suggest that maximum correlation between knowledge and score is attained when the mean difficulty is near the mean level of knowledge and penalty is well above $\frac{1}{M-1}$. Therefore, in this model number-right scoring is not superior to formula scoring. This result is robust to a number of deviations from the benchmark model. It obtains under binary or partial knowledge, under different normalizations of the score, when risk aversion is related to knowledge and when there is a pass-fail break point.

Finally, our model provides a framework for further research on the optimality of different features of multiple-choice tests. In particular, the model could be easily extended to account for other types of individual behavior. For instance, it could easily adopt the extended IRT model of San Martn, Pino, and De Boeck (2006) where the guessing parameter (or probability of correct guess) does not only depend on the item, but also on the ability of individuals. Another interesting extension would follow Bereby-Meyer et al. (2002), relaxing the assumption of expected utility maximization by assuming that the utility of rewards is smaller than the disutility of penalties.

## References

Albanese, M. A. (1986). The correction for guessing: a further analysis of Angoff and Schrader. *Journal of Educational Measurement*, *23*, 225-235.

Albanese, M. A. (1988). The projected impact of the correction for guessing on individual scores. *Journal of Educational Measurement*, *25*, 149-157.

Avila, C., & Torrubia, R. (2004). Personality, expectations, and response strategies in multiple-choice question examinations in university students: A test of Gray's hypothesis. *European Journal of Personality*, *18*, 45-59.

Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind and Society*, *4*, 2-12.

Bereby-Meyer, Y., Meyer, J., & Flascher, O. M. (2002). Prospect theory analysis of guessing in multiple choice tests. *Journal of Behavioral Decision Making*, *15*, 313-327.

Bernardo, J. M. (1998). A decision analysis approach to multiple choice examinations. In F. J. Girn (Ed.), *Applied decision analysis* (p. 195-207). Boston: Kluwer.

Bliss, L. B. (1980). A test of Lord's assumption regarding examinee guessing behavior on multiple-choice tests using elementary school students. *Journal of Educational Measurement*, *17*, 147-153.

Budescu, D., & Bar-Hillel, M. (1993). To guess or not to guess: a decision-theoretic view of formula scoring. *Journal of Educational Measurement*, *30*, 227-291.

Lord, F. M. (1975). Formula scoring and number-right scoring. *Journal of Educational Measurement*, *12*, 7-11.

McLeod, I., Zhang, Y., & Yu, H. (2003). Multiple-choice randomization. *Journal of Statistics Education*, *11*.

Prieto, G., & Delgado, A. R. (1999). The role of instructions in the variability of sex-related differences in multiple-choice tests. *Personality and Individual Differences, 27*, 1067-1077.

San Martn, E., Pino, G. del, & De Boeck, P. (2006). Irt models for ability-based guessing. *Applied Psychological Measurement, 30*, 183-203.

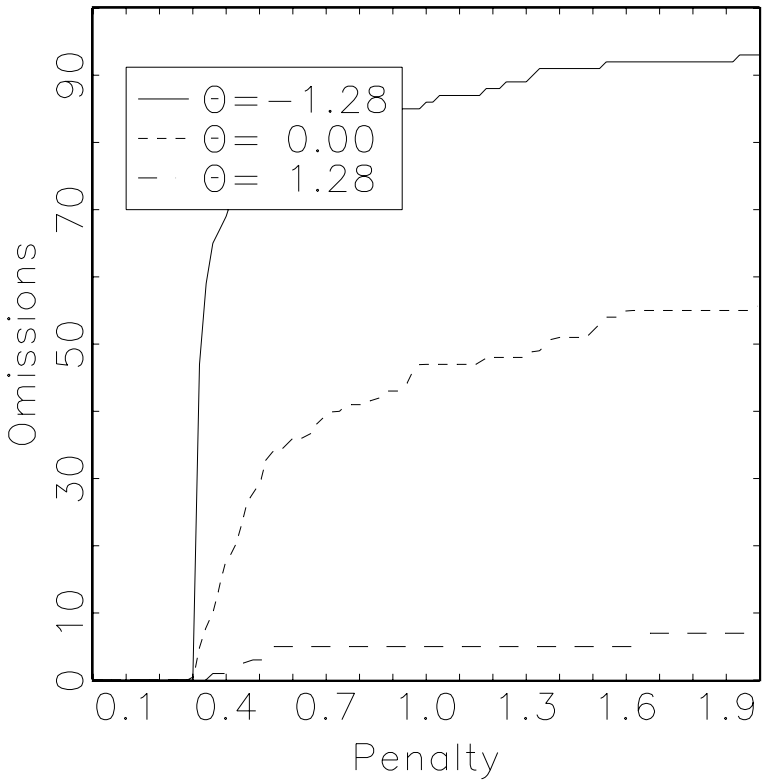


*Figure 1.* Omissions as a function of penalty for three different levels of knowledge and risk aversion fixed at the median value.

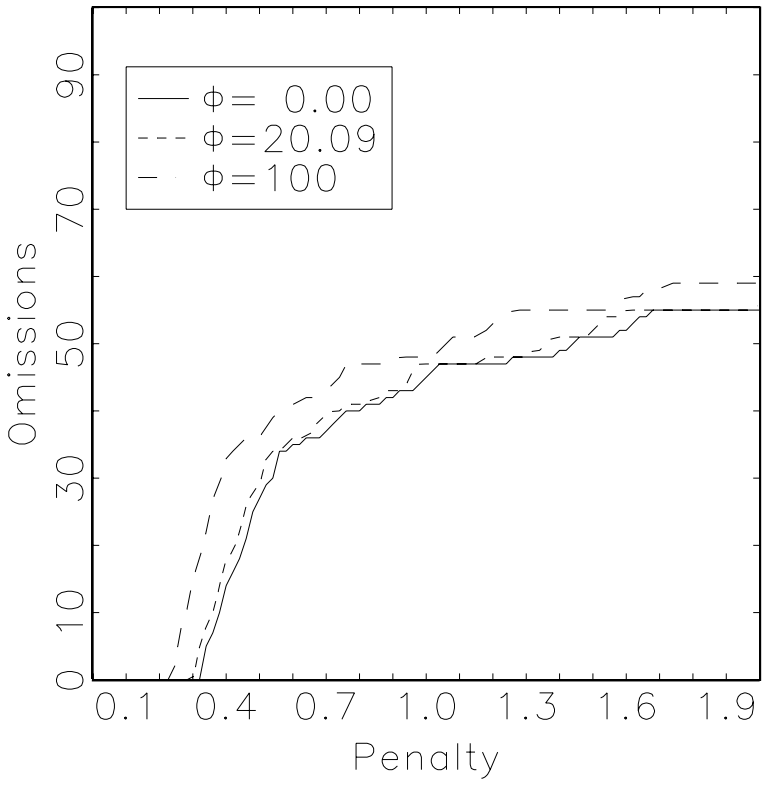


*Figure 2.* Omissions as a function of penalty for three different values of risk aversion and knowledge fixed at the median value.

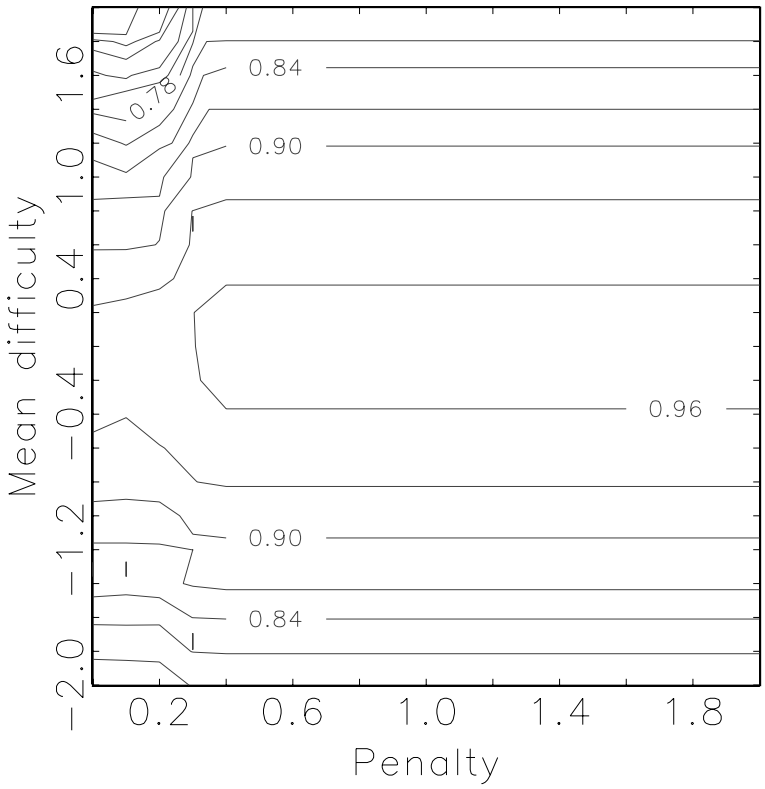


*Figure 3.* Contour lines of the correlation between score and knowledge as a function of mean difficulty and penalty in the binary knowledge model.

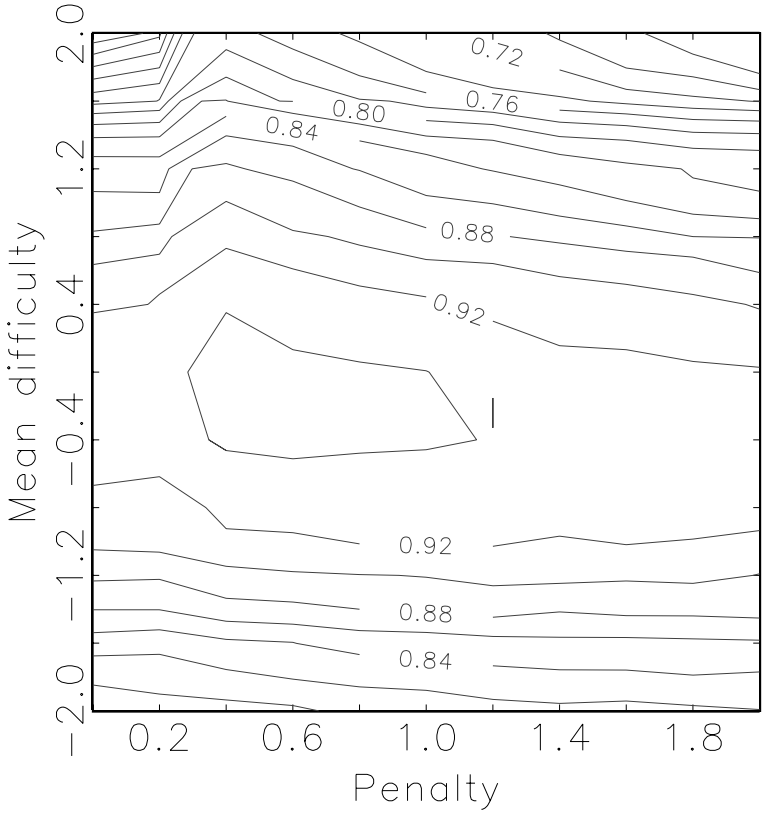


*Figure 4.* Contour lines of the correlation between score and knowledge as a function of mean difficulty and penalty in the partial knowledge model.

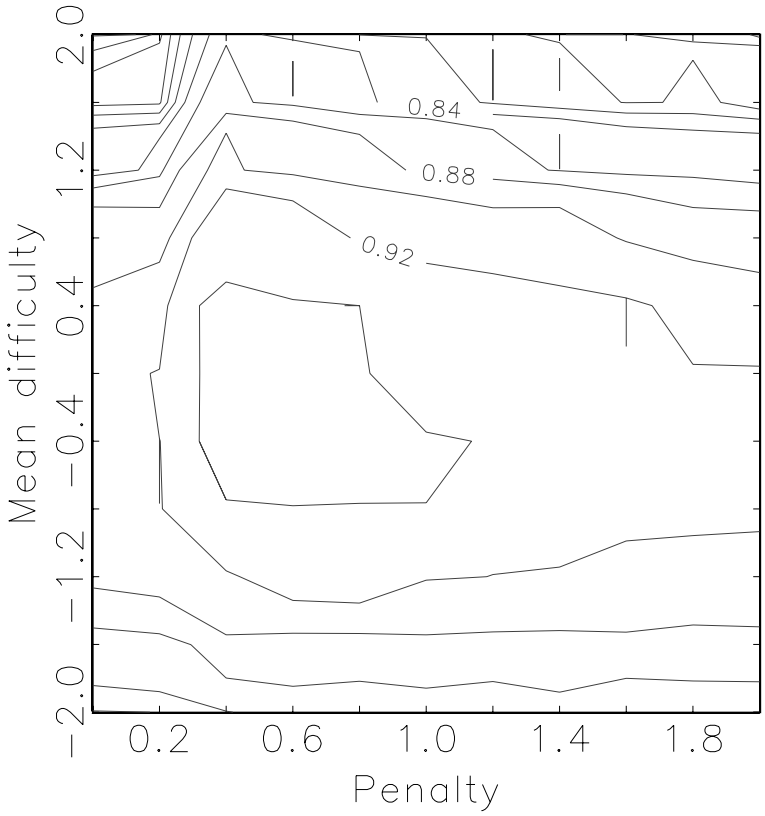


*Figure 5.* Contour lines of the correlation between score and knowledge as a function of mean difficulty and penalty when the variance of difficulty is twice the variance of knowledge.

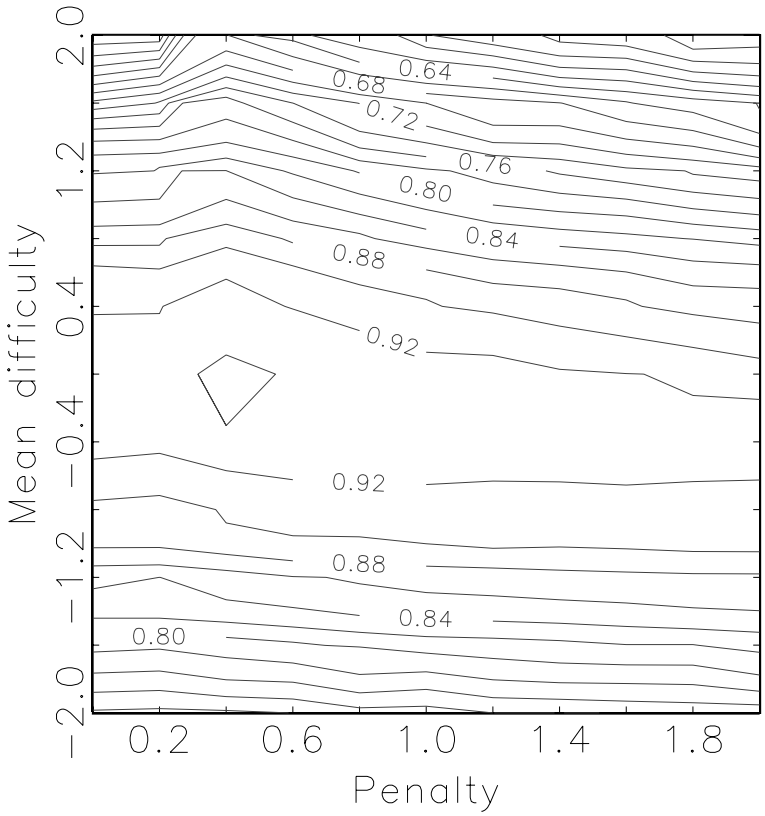


*Figure 6.* Contour lines of the correlation between score and knowledge as a function of mean difficulty and penalty when the variance of difficulty is half the variance of knowledge.

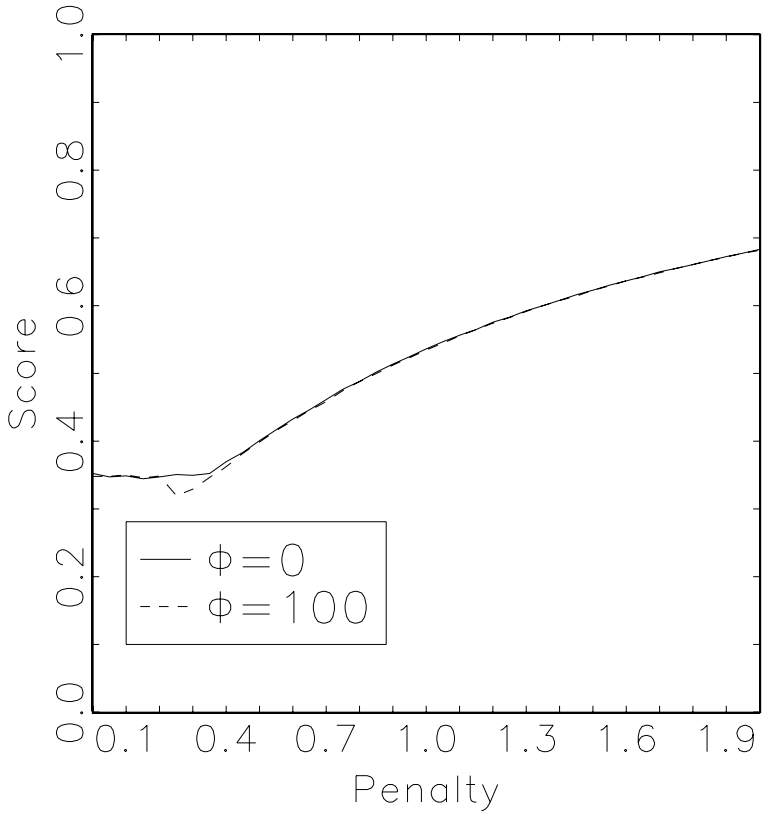


*Figure 7.* Score as a function of penalty for two individuals with low knowledge.

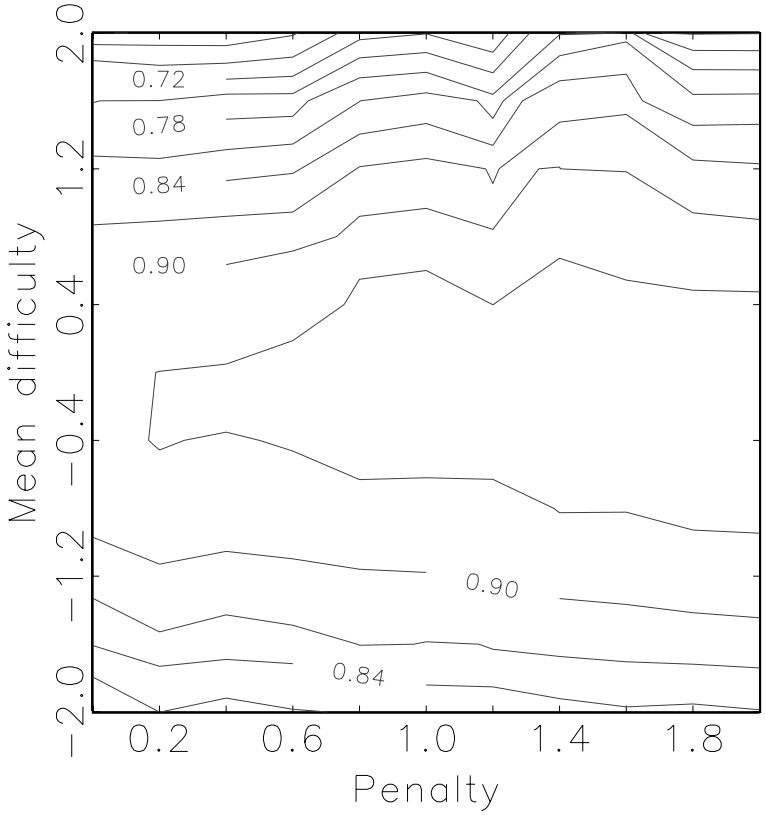


*Figure 8.* Contour lines of the correlation between score and knowledge as a function of mean difficulty and penalty in exams with a pass-fail breakdown