Optimal Correction for Guessing in Multiple-Choice Tests

María Paz Espinosa and Javier Gardeazabal
University of the Basque Country

Abstract
MCQ examinations are frequently used but there is no consensus as to whether a penalty for wrong answers should be used or not. For example, in some countries examinations for medical licensing include MCQ sections with penalty while in others there is no penalty for wrong answers. We contribute to this discussion with the analysis of the effects of penalties. In a partial knowledge model there is a trade-off between bias and measurement error that is affected by the level of penalties. We build on IRT and introduce students’ optimal behavior. Our simulations indicate that the optimal penalty may be relatively high: even though it discriminates against risk averse students this effect is small compared with the measurement error that it prevents.
"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 sometimes 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 IRT, we propose a theory of students’ behavior in multiple-choice tests that might explain why they may find it optimal to answer some items and leave others blank.

Suppose a test has $N$ items and each item has $M$ possible choices. Different methods of correction for guessing have been proposed: elimination/inclusion scoring and confidence or probabilistic scoring, among others (e.g., Budescu & Bar-Hillel, 1993; Collet, 1971). Although the model developed in this paper allows for other types of scoring, we will focus on the most used scoring formula. The score, $s$, is a function of the number of rights, $r$, wrongs, $w$, and the penalty for wrongs, $p$: $s = r - pw$ with $p = \frac{1}{M-1}$. Under these conditions, the expected value of a pure guess, that is getting the right answer ($s = 1$) with probability $\frac{1}{M}$ and failing ($s = -\frac{1}{M-1}$) with probability $\frac{M-1}{M}$, is equal to the expected value of omitting ($s = 0$), that is, $(\frac{1}{M}) - (\frac{1}{M-1} \times \frac{M-1}{M}) = 0$. (e.g., Budescu & Bar-Hillel, 1993).

Students sometimes have partial knowledge and are able to rule out one or more distractors, (e.g., Bliss, 1980). Under partial knowledge and with a penalty $\frac{1}{M-1}$, the expected value of guessing is greater than the value of omitting. Sometimes subjects who have partial knowledge omit items with positive expected reward. Evidence on this was provided by Bliss (1980) who administered a MCT under formula scoring where some subjects left questions unanswered. When time was up, he allowed subjects to circle the correct answer in those questions left unanswered, this time with no penalty for wrongs. He found that subjects who omitted under formula scoring would have chosen more correct answers than under chance expectations. Highham (2007) ran two experiments using questions from actual SAT tests and with different penalty for errors (no penalty, penalty=0.25 and penalty=4); after the SAT was completed subjects were required to complete questions that were originally left unanswered using a pen with red ink (forced report condition). Comparing the free report condition to the forced report condition, Highham (2007) found that the decrease in the number of omissions increased both the raw and the corrected score and to a greater extent in the group facing the high penalty. These results are consistent with previous work on the accuracy of withheld answers (Bliss, 1980; Cross & Frary, 1977; Ebel, 1968; Harden, Brown, Biran, Dallas Ross, & Wakeford, 1976; Muijtjens, Mameren, & Hoogenboom, 1999; Sanderson, 1973; Sax & Collet, 1968; Sherriffs & Boomer, 1954; Slakter, 1968a, 1968b). This evidence is not consistent with expected score maximizers.
A possible explanation of this behavior can be advanced using Decision Theory under uncertainty (DT), (e.g., Arrow, 1971; Friedman & Savage, 1948; Pratt, 1964; von Neumann & Morgenstern, 1944). Unless a subject is completely sure about which of the alternatives is the correct answer, the decision whether to answer an item or leave it blank involves some uncertainty. Subjects face the dilemma of choosing between omitting and getting no reward or answering and bearing the risk of being penalized. Within the DT framework, subjects’ attitudes towards risk can be classified into three categories: risk aversion, risk neutrality and risk seeking. A test taker who omits an item with positive expected reward is called risk averse. Those who answer all items with positive expected reward are labeled risk neutral. A risk seeking individual would answer items with negative expected reward. Risk aversion is a type of behavior that is observed in many real situations. A leading example of risk aversion is the fact that people choose to buy insurance to hedge against several risks such as a car accident, house fire or burglary. However, within the educational measurement literature, most papers assume that the test taker is an expected score maximizer and therefore risk neutral (e.g., Budescu & Bar-Hillel, 1993). If subjects’ goal were to maximize the expected score, they would answer whenever the expected value of answering turns out to be higher than or equal to the payoff for omitting. However, the evidence previously mentioned indicates that examinees do not behave as expected score maximizers. Moreover, although the experiments were not aimed at eliciting the students’ attitudes towards risk the results can be interpreted as prima facie evidence of risk aversion.

There exist other possible explanations for this evidence: 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). However, in this paper we will develop an explanation based on the idea that under formula scoring rational examinees may try to avoid the risk of being penalized by omitting some items in which they have partial knowledge.

When students omit items with positive expected score, it is reasonable to argue that correction for guessing may discriminate against them and, therefore, to recommend not to use formula scoring but number right scoring. Formula scoring may reduce the measurement error due to guessing only at the expense of introducing some bias in the scores of those who omit. However, number-right scoring induces students to answer all questions and the bias disappears. Moreover, the measurement error can be reduced by administering tests with a large number of questions. In practice, time constrains may preclude the use of a large test. Increasing the number of distractors would also be a means to reduce guessing but sometimes the question does not admit a large number of alternatives.

The debate whether to use formula scoring or number right scoring is an old and hot one. Early studies by Edgington (1965) and Little (1966) recommended number right scoring while Cureton (1966) and Davis (1967) argued in favor of formula scoring. Sabers and Feldt (1968), on the other hand, found that validity and reliability did not exhibit significant differences for correction or non-correction. Face validity however would favor correction
for guessing.\textsuperscript{1} The debate was surveyed by Diamond and Evans (1973), who examined the debate in terms of validity, reliability, test-taker behavior and other dimensions. Frary (1988) views formula scoring as "inappropriate for most classroom testing but desirable for speeded tests and for difficult test with low passing scores." Bar-Hillel, Budescu, and Attali (2005) suggest that using formula scoring is a form of irrationality on the part of the examiner, as they claim number right is a superior scoring method.

In parallel with the theoretical debate, real tests are administered with and without correction for guessing. For example, in the multiple choice component of the Medical Council of Canada Qualifying Examination (MCCQE) or in the United States Medical Licensing Examination (USMLE) there is no penalty, while in the Combined Medical Services Examination in India (CMS) or the MIR (Médico Interno Residente) exam in Spain wrong answers are penalized (in the first case $p = 1/3$ with four alternatives and in the second $p = 1/3$ with five alternatives). Not only there are differences across countries in this respect, but within the same country there is no standard. College admissions in the USA are based on the SAT and ACT test scores. SAT is administered by the Educational Testing Service and uses a penalty of one fourth with 5 alternatives. ACT is administered by ACT Inc. and does not penalize for incorrect answers.

In this paper we develop a model of students' behavior combining IRT and DT. The model provides a theoretical framework to analyze the optimal correction for guessing. The model can be solved by simulation and yields some support to the use of formula scoring. Depending on the students' attitudes towards risk, the optimal penalty need not be the typically used $1/M - 1$ value nor zero. In fact, our simulations show that the optimal penalty is above $1/M - 1$. The reason can be found in 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 & Delgado, 1999). Although the theoretical framework developed in this paper allows us to make some claims about the desirability of correcting for guessing, the actual endorsement of these claims should await for field experimentation.\textsuperscript{2}

We find instructive to begin the analysis with the simple but unrealistic setting of binary knowledge. Under binary knowledge a student knows the correct answer with either probability zero or one. This deterministic setting was first studied by Guttman (1950). Unrealistic as it is, this model is used as a benchmark and it is helpful to understand the partial knowledge case. The binary knowledge assumption allows us to illustrate the trade-off between bias and measurement error and how the choice of the penalty level affects that trade-off. In this framework we introduce the basic concepts borrowed from DT and are

---

\textsuperscript{1}Consider an examinee who does not know the answer to an item. Asking the examinee to guess has less face validity than instructing the examinee to omit.

\textsuperscript{2}See Espinosa and Gardeazabal (2005) for a field experiment which provides support to the hypothesis of students’ rationality when they face different scoring rules.
able to state and prove three propositions regarding the optimal scoring in MCT. Educational measurement has long recognized that the binary knowledge hypothesis is inappropriate and developed IRT and other tools to deal with the more satisfying assumption of partial knowledge. Thus, we amend our setting to combine our DT framework with IRT. In this case, the model becomes more complex and an analytical solution cannot be found. We then proceed by simulation to obtain a numerical solution.³

**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 establishing a benchmark.

**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 \( \theta_t \). We assume that students have an objective function that can be interpreted as an index of satisfaction with the outcome of the exam. In the tradition of DT we call this function the utility function, \( u_t(s_t) \) and assume it is continuous on the score, strictly increasing (the higher the score the better) \( u'_t(s_t) > 0 \) and concave \( u''_t(s_t) \leq 0 \). Therefore, a student is completely characterized by \( \{\theta_t, u_t(s_t)\} \).

Note that this is more general than the usual objective function \( u_t(s_t) = s_t \) and includes it as a special case. When the utility function is an affine transformation of the score, \( u''_t(s_t) = 0 \), we say that the student is risk neutral. When the utility function is strictly concave \( u''_t(s_t) < 0 \) the student is risk averse.

**Binary knowledge.** For each question \( i \), student \( t \) 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 \( b_i \).

**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 and guesses, the probability of getting the right answer is \( \frac{1}{M} \).

**True score.** The Classical Model (e.g., Lord & Novick, 1974) postulates that the score measurement in item \( i \) by student \( t \), \( x_{it} \) contains an error of measurement \( e_{it} \) and the difference \( x_{it} - e_{it} \) is the true score. We show that in the binary knowledge model the true score is the number of questions the student knows the answer to. In the binary knowledge model, the true score, \( k_{it} = x_{it} - e_{it} \) is either 0 or 1. When a student knows the answer to an item, the score is \( x_{it} = 1 \) and the true score \( x_{it} - e_{it} = 1 \). Thus, the error of measurement is \( e_{it} = 0 \) and this occurs with probability one, so the expected error is \( 0 \times 1 = 0 \). When a student does not know the answer, the true score is \( x_{it} - e_{it} = 0 \) and the score is \( x_{it} = 1 \)

³Simulation exercises similar to the one performed in this paper have been successfully used (e.g., Frary, 1982, 1989; García-Pérez & Frary, 1989).
with probability \( \frac{1}{M} \) and \( x_i = -\frac{1}{M-1} \) with probability \( \frac{M-1}{M} \). Thus, the expected error of measurement is zero,

\[
E(e_i) = \left( 1 \times \frac{1}{M} \right) - \left( \frac{1}{M-1} \times \frac{M-1}{M} \right) = 0,
\]
as in the Classical Model. The score obtained in a test of \( N \) questions is \( s_t = \sum_{i=1}^{N} x_{it} \). Under these conditions the expected score measurement is

\[
E(s_t) = \sum_{i=1}^{N} E(x_{it}) = \sum_{i=1}^{N} E(k_{it} + e_{it}) = \sum_{i=1}^{N} 1(\theta_t > b_i)
\]

where \( 1(A) \) is equal to one when \( A \) is true and zero otherwise. Therefore, the expected score is equal to the number of questions the student knows the answer to, the true score.

**Score.** The score is written as \( s_t = r_t - pw_t \). Assume student \( t \) answers \( n_t \) questions out or \( N \). 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 (risk neutral students) as a special case, with \( u(s_t) = s_t \), but it also includes risk averse students with \( u'(\cdot) > 0 \) and \( u''(\cdot) < 0 \), which are usually left out of the analysis.

**Objective.** The examiner would like the grading procedure to elicit the true score or at least order students according to it. The examiner sets the penalty \( p \), and the distribution of difficulties \( b_t \) with the objective of maximizing Pearson’s correlation between knowledge and score

\[
\rho_P(\theta, s) = \frac{T \sum_{t=1}^{T} (s_t - \bar{s})(\theta_t - \bar{\theta})}{\sqrt{\sum_{t=1}^{T} (s_t - \bar{s})^2} \sqrt{\sum_{t=1}^{T} (\theta_t - \bar{\theta})^2}}
\]

where \( \bar{s} \) and \( \bar{\theta} \) 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

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

would be the appropriate objective.

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, \( \frac{T}{4} \) students a score of 1 and \( \frac{T}{2} \) 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 partial knowledge.

Our main result in this section is that in the binary knowledge model 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:
\[
E(s(k + 1)) = \frac{1}{M} - \frac{M - 1}{M}p.
\]
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 \( E(u(s(k + 1))) < u(E(s(k + 1))) \) (see Jensen, 1906).

\( ^{4} \)The same result may obtain for exams with a larger number of questions. For example, if the exam had three questions and all the students knew the answer to the first two questions, for the risk neutral students guessing at the last question is better than omitting since \( \frac{1}{2}(4 + 1) + \frac{1}{2}(4 - p) > 4 \) while for risk averse students the opposite is true, \( \frac{1}{2}\sqrt{(4 + 1)} + \frac{1}{2}\sqrt{(4 - p)} < 4 \).
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 \( s \). Our following two 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 \( \theta_1 < \theta_2 < \ldots < \theta_T \) and the exam has at least \( T - 1 \) questions with difficulties \( b_1 < b_2 < \ldots < b_{T-1} \) such that \( \theta_1 < b_1 < \theta_2 < b_2 < \ldots < b_{T-1} < \theta_T \), 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 \( \text{rank}(s_t) = t = \text{rank}(\theta_t) \) 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 \( b_{N,i}^{\text{max}} \) be the maximal element of the set \( \{ b_i | b_j < \theta_j, j = 1, \ldots, N \} \). Since the support of \( b_i \) and \( \theta_i \) are the same, \( \ldots, \theta_i - b_{N,2,i}^{\text{max}} \geq \ldots \geq \theta_i - b_{N,1,i}^{\text{max}} \geq \theta_i - b_{N,i}^{\text{max}} \geq \theta_i \geq 0 \). Therefore, as \( N \) tends to infinity, \( b_{N,i}^{\text{max}} - \theta_i \) 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 that student $t$ answers item $i$ correctly 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. 

\[ P(y_{i1} \geq 0) \geq P(y_{i2} \geq 0) \geq \ldots \geq P(y_{iN} \geq 0). \]

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 $s_t = (1 + p) \sum_{i=1}^{n_t} 1(z_{ti} = 1) - p n_t$. Since $E(1(z_{ti} = 1)) = P(z_{ti} = 1)1 + (1 - P(z_{ti} = 1))0 = P(z_{ti} = 1)$, therefore, the expected score is $E(s_t) = (1 + p) \sum_{i=1}^{n_t} P(z_{ti} = 1) - p n_t$.

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

\[ P(z_{ti} = 1) > \frac{p}{1 + p}. \]

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

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

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 and this prevents examiners from getting a perfect ordering of the students according to their true score. With partial knowledge it is no longer possible to characterize the optimal test so easily and 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 \( \phi = 1 \).

The extended power utility function (EPU) is appropriate for several reasons. First, we want the utility function to be flexible enough to represent attitudes towards risk of both risk averse and risk neutral subjects. This requirement rules out other functions such as the exponential utility, which does not nest the risk neutrality case, and the quadratic utility, which sets an upper bound on the degree of risk aversion. The EPU captures risk neutrality when \( \phi = 0 \) and risk aversion when \( \phi > 0 \). Second, another frequently used utility function, the power utility function, obtains when \( a = 0 \). However, in our setting a strictly positive value of \( a \) is needed so that \( a + s \) is non-negative and the utility function is well defined. Third, this utility function exhibits increasing relative risk aversion. The Arrow-Pratt measure of relative risk aversion (Arrow, 1965; Pratt, 1964) is

\[ R(s) = -\frac{s u''(s)}{u'(s)} = \frac{s \phi}{a + s} > 0. \]

This seems a reasonable assumption: it amounts to saying that the student exhibits more risk aversion for high values of the score than for low values of the score. Although there is no evidence on this in the field of psychometrics, in other decision settings agents behave differently in their attitudes towards risk when uncertainty affects small or large quantities, (e.g., Pratt, 1964).

**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 \( (\mu_1, \sigma^2_1) \) and independent of risk aversion.

**Normal difficulty.** The examiner can set exams whose difficulties are random draws from a normal distribution with parameters \( (\mu_2, \sigma^2_2) \).

**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

\[ P(z_{ti} = 1) = c_i + \frac{1 - c_i}{1 + \exp(-(\theta_t - b_i)/g_i)} \]  

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 \( \{s(0, 0), s(0, 1), s(1, 1), s(0, 2), s(1, 2), s(2, 2)\} = \{\frac{p}{1+p}, \frac{p}{1+p}, \frac{1+p}{2(1+p)}, 0, \frac{1}{2}, 1\} \), 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 \( i = 1, \ldots, N \), 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 \( r = 1, \ldots, n, n = 1, 2, \ldots, N \) 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 utility at all possible values \( n = 1, \ldots, N \).

**Numerical results**

This section presents the results of the simulation. All calculations were coded in GAUSS. First, 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 \( \frac{3}{\pi} \), hence \( v_{it} \) has unit variance. This is repeated for penalty values in the interval \([0, 2]\).

Figure 1 graphs the optimal number of omissions computed following the numerical procedure described in the previous section. In particular, the number of omissions maximizes the expected utility for three individuals with different levels of knowledge. 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

\[^5\]These programs as well as all results stated in this section and not shown in order to save space are available from the authors upon request.
risk averse but start omitting at different penalties. The two least knowledgeable ones start omitting for penalties below the critical value \( M^{-1} = 0.33 \) 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.

Next we simulate 30 exams using the parameters of Table 1. The number of students and items used in the simulations are high enough to get sensible estimates of correlation.
Figure 2. Omissions as a function of penalty for three different values of risk aversion and knowledge fixed at the median value.

coefficients. The parameter values of the distributions of knowledge and risk aversion have been chosen so that the number of questions students omit has enough dispersion.

TABLE 1
Parameter values used in the simulations

<table>
<thead>
<tr>
<th>T</th>
<th>N</th>
<th>M</th>
<th>(a)</th>
<th>(c_i)</th>
<th>(g_i)</th>
<th>(\phi_t)</th>
<th>(\theta_t)</th>
<th>(b_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>30</td>
<td>4</td>
<td>1</td>
<td>0.25</td>
<td>(\frac{3}{\pi})</td>
<td>3</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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
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 $g_i = \frac{1}{\pi^2}$ so that the variance of the disturbance term $v_{ti}$ is 1. Figure 4 plots the iso-correlation curves. As it was the case with the binary knowledge model, maximum correlation between knowledge and score is

---

6In simulations not reported here we have verified that when the discrimination parameter $g_i$ is allowed to vary across items the results are very similar.
contour lines of the correlation between score and knowledge as a function of mean difficulty and penalty in the partial knowledge model. 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. Although the maximum value of the correlation surface is attained for positive penalty values, a penalty value of zero reduces correlation by a small amount. Therefore, there is no much improvement in term of correlation by using a penalty with respect to number-right scoring. Summing up, according to our simulation results shown in Figures 3 and 4 we can state the following:

**Result 1** A relatively high penalty is not worse 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 et al., 2005).

Notice that the correlation between score and knowledge is maximized for values of the mean difficulty equal to the value of the mean knowledge.  

---

We have verified that this result also obtains for different values of mean knowledge. This result is shown in an appendix available from the authors upon request.
**Result 2** The maximum correlation between knowledge and score is attained when the mean value of difficulty is near the mean value of knowledge.

This result is consistent with previous results on the optimal difficulty. Felt (1993) obtained an item optimal difficulty of 0.5 (so that half the students would answer it correctly). In our setting this is equivalent to a difficulty parameter equal the mean level of knowledge since knowledge is assumed to have a symmetric distribution around the mean.

So far we have assumed that risk aversion and knowledge are independent. We have verified that 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.\(^8\)

**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, \(\sigma^2 = 2\), and Figure 6 shows the iso-correlation lines when the variance of difficulty is low, \(\sigma^2 = 0.5\). When the variance of the distribution of difficulty is high,\(^8\) these simulation results are shown in an appendix available from the authors upon request.
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.

**Does correction for guessing discriminate against risk averse students?**

Some researchers and practitioners (e.g. Frary, 1988) 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 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 discrimination is quantitatively small, peaking at 3.1 per cent of the score when penalty is 0.3 (at this point the risk averse individual omits 69.3 percent more items than the risk neutral). We have repeated this simulation with more knowledgable students, but then the difference in omissions between the risk averse and the risk neutral individual decreases and so does the difference in the score. Therefore, it is fair to say that formula scoring does not discriminate against knowledgable, risk averse examinees and only discriminates against risk averse examinees with little knowledge. More importantly, the difference in the score almost disappears 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 and penalty values in
the neighborhood of the typical value \( \frac{1}{M - 1} \), but it is almost negligible for students with high or average levels of knowledge. This discrimination disappears for penalty values above or below \( \frac{1}{M - 1} \).

**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:

\[
s_i = \max\left\{ \frac{1}{N} (r_t - pw_t), 0 \right\}.
\]

Truncating the score at zero has the potential to affect the correlation surface. However, 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

\[
u(s) = \begin{cases} \frac{a^1 - \phi}{1 - \phi} & \text{if } s < \bar{s} \\ \frac{(a+s)^1 - \phi}{1 - \phi} & \text{if } s \geq \bar{s} \end{cases}
\]

where \( \bar{s} \) is the passing score. Under this assumption, students value any score below \( \bar{s} \) at \( u(0) \) and utility is increasing in the score only if \( s \geq \bar{s} \).

Figure 8 shows the contour lines of the correlation surface when exams have a passing score of \( \bar{s} = 0.75 \). 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:

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

where $u_f < u_p$. Therefore expected utility is

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

Since $u_f < u_p$ and both are fixed values, maximizing expected utility is equivalent to maximizing the probability of passing the exam, $P(s \geq s_c)$. 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 seeking behavior on the part of examinees. To compensate for this behavior, the optimal penalty is higher.
Conclusions

In this paper we formalize the idea that under formula scoring students’ behavior in terms of the number of omissions depends on the penalty: the higher the penalty the larger the number of omissions. Rational but cautious students may leave questions blank even though the expected score is positive. This theoretical result is consistent with the experimental evidence (Bliss, 1980; Cross & Frary, 1977; Ebel, 1968; Harden et al., 1976; Highham, 2007; Muijtjens et al., 1999; Sanderson, 1973; Sax & Collet, 1968; Sherriffs & Boomer, 1954; Slakter, 1968a, 1968b). To introduce this idea in a formal model we use the theory of decision under uncertainty (e.g., Arrow, 1971; Pratt, 1964; von Neumann & Morgenstern, 1944) which explains why individuals may behave cautiously and for example be willing to buy insurance. According to this theory, subjects may not behave in order to maximize expected payoffs; rather, individuals may be willing to pay a risk premium to avoid bearing too much risk.

Combining IRT and the theory of decision under uncertainty, we have developed a model of students’ optimal behavior. First, we introduce this idea in a binary knowledge model. Our main conclusions are the following: (i) in a binary knowledge model where examinees know the answer to each question with probability either one or zero, the examiner can set the exam to get a perfect correlation between knowledge and scores, (2) in a partial knowledge model the examiner cannot set the exam to guarantee that the correlation will be one. The examiner can, however, set the exam to maximize the correlation. The partial knowledge model cannot be solved analytically so we simulate it. Our simulations indicate that (3) the optimal value of penalty is relatively high, well above \( \frac{1}{M-1} \), and (4) the mean difficulty coincides with the mean level of knowledge. This last result is consistent with Felt (1993) and Lord and Novick (1974) who have dealt with the question of optimal item difficulty and estimated it to maximize test score variance. These results are robust to a number of changes of the benchmark model (different normalizations of the score, positive or negative correlation between the student’s knowledge and cautiousness and to a pass-fail break point) and also to different assumptions on the discrimination parameter, the variance of the error term and the mean and variance of the distribution of knowledge and difficulties.

The main implication of our analysis is that number-right scoring is not superior to formula scoring when the examiner wants to maximize the correlation between scores and knowledge, and students have a rational but cautious behavior. Field experiments should be performed to verify whether or not our results have any empirical relevance.

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 Martín, del 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


