Departament d'Estadística i I.O., Universitat de València.
Facultat de Matemàtiques, 46100–Burjassot, València, Spain.
Tel. 34.6.386.4314 (direct), 34.6.386.4362 (office); Fax 34.6.386.4735,
Internet: jose.m.bernardo@uv.es, Web: http://www.uv.es/~bernardo/

Printed on November 25, 1997
Invited paper presented at the *International Workshop on Decision Analysis Applications*, held at the Royal Academy of Sciences, in Madrid, Spain, on July 11–12, 1997.

A Decision Analysis Approach to Multiple-Choice Examinations

JOSÉ M. BERNARDO
Universitat de València, Spain

SUMMARY

We present a decision analysis approach to the problems faced by people subject to multiple-choice examinations, as often encountered in their education, in looking for a job, or in getting a driving permit.

From the candidate's viewpoint, each question in this form of examination is a *decision problem*, where the decision space depends on the examination rules and the expected utility is some function of the expected score. We analyse this problem for the two basic situations which occur in practice, namely when the candidate wants to maximize his or her expected score, and when he or she wants to maximize the probability of obtaining the minimum grade required to pass, and we derive the corresponding optimal strategies.

We argue that for multiple-choice examinations to be *fair*, candidates should be required to provide a *probability distribution* over the possible answers to each question, rather than merely marking the answers judged to be more likely; we then discuss the appropriate scoring rules and the corresponding optimal strategies. As an interesting byproduct, we deduce some illuminating consequences on the scoring procedures of multiple-choice examinations, as they are currently performed.

Keywords: BAYESIAN EDUCATION; DECISION ANALYSIS; MULTIPLE-CHOICE QUESTIONNAIRES; PROBABILITY ASSESSMENT; SCORING RULES; UTILITY ASSESSMENT.

José M. Bernardo is Professor of Statistics at the Universitat de València, Spain. This research was partially funded with grant PB96-0754 of the DGICYT, Madrid, Spain.

1. INTRODUCTION

Young citizens are increasingly subject to a number of formal examinations whose results are often very important to their future. Many of these examinations take the form of a multiple-choice questionnaire; important examples include educational examinations, obtaining a driving licence, and interviewing for a job. In this paper, we present a decision analysis approach to the problems faced by people subject to multiple-choice questionnaires.

From the candidate's viewpoint, each question in this form of examination is a *decision problem*, where the decision space depends on the examination rules, and where the expected utility is some function of the expected score. We analyse this problem for the two basic situations which occur in practice, namely when the candidate wants to maximize his or her expected score, and when he or she wants to maximize the probability of obtaining the minimum grade required to pass. We avoid tiresome repetition, 'he or she' will systematically be replaced by a generic 'she' in the rest of the paper.

The literature on non decision-based analysis of conventional multiple-choice tests is huge, specially is psychology-oriented journals. Relevant signposts include see Lierly (1951), Solomon (1955), Chernoff (1962), Nogaki (1984), Thissen and Steinberg (1984), Pollard, (1985), Martín and Luna (1989, 1990), Hsu, Leonard and Tsui (1991), Hutchinson (1991, 1993), Klein (1992), Post (1994), and references therein.

The problem is sociologically important, and a knowledge of the optimal strategy will certainly be of interest to people taking this type of examinations. Moreover, the analysis presented has strong *pedagogical* motivations, for it may successfully be used to introduce the basic elements of decision theory to undergraduate candidates of *any* discipline. Indeed, whether or not it is considered appropriate to include some axiomatic foundations in the course, we argue that it is pedagogically best to begin with a *detailed* example of a decision problem where the basic concepts may be illustrated; ideally, this should be a simple but *realistic* problem, within a context which students understand and care for. We claim that the decision problems posed by multiple-choice examinations are ideally suited for this purpose.

In Section 2 we analyze the conventional decision problem where, for each question, the candidate is required to mark that answer which she judges to be more likely, or to leave that question blank, and we derive the candidate's optimal strategy if she wants to maximize her expected score. In Section 3, we argue that for multiple-choice examinations to be *fair*, candidates should be requested to provide a complete *probability distribution* over the answers proposed, rather than to merely mark that answer which they judge to be more likely; we analyze the corresponding candidate's optimal strategy if she wants to maximize her expected score, and we make use those results to derive some possible improvements on the scoring procedure, if conventional multiple-choice examinations are nevertheless to be performed. In Section 4, we analyze the more complex problem posed when the candidate's preferences are to maximize her probability of obtaining a minimum grade, rather than maximizing her total score. In Section 5 we make a specific proposal for an appropriate implementation of multiple-choice examinations, and suggest areas for additional research.

2. THE CONVENTIONAL DECISION PROBLEM

The structure of any finite decision problem may formally be described in terms of the set $A = \{a_1, \ldots, a_k\}$ of possible *alternatives*, the set $\Theta = \{\theta_1, \ldots, \theta_m\}$ of exhaustive, mutually exclusive *relevant uncertain events*, and the set of *consequences*, $c_{ij} = c(a_i, \theta_j)$ which may possibly result; it is well known that sensible set of axiom requirements on coherent decision making (see Bernardo and Smith, 1994, Ch. 2, and references therein, for some options) logically

imply (i) that preferences among consequences must be quantified with a real-valued *utility function* $u(c_{ij}) = u(a_i, \theta_j)$, (ii) that the plausibility of the uncertain events must be described with a *probability distribution* $\{p(\theta_1), \ldots, p(\theta_m)\}$ ($p(\theta_j) \geq 0$, $\sum_j p(\theta_j) = 1$) over the relevant uncertain events, and (iii) that the optimal alternative a^* is that which *maximizes the expected utility*

$$\overline{u}(a_i) = \sum\nolimits_{j=1}^{m} u(a_i, \theta_j)\, p(\theta_j); \tag{1}$$

It should be stressed that the word *probability* is to be understood in its common *semantic* use, that is, as a *degree of belief*.

Let $\{\delta_1, \ldots, \delta_k\}$ be the k possible answers to a multiple-choice question, which are assumed to be mutually exclusive and to contain the true answer δ^*. Following conventional practice, we suppose that (i) the candidate is required to either mark the answer which she considers to be correct or to leave the question blank, and that (ii) the scoring rule is such that one point is awarded if the correct answer is marked, zero points if left blank, and $c \geq 0$ points are subtracted if an incorrect answer is marked.

We will first consider the simplest situation, where the candidate wants to maximize her expected total grade. In that case, since her total grade is just the sum of the scores obtained in each question, she clearly has to maximize her expected score in each individual question.

For each question, the set of alternatives is $\{a_0, a_1, \ldots, a_k\}$, where a_0 denotes the option of leaving the question blank, a_i denotes the option of marking δ_i as the correct answer, and the score function may be written as

$$u(a_0, \delta_j) = 0, \qquad u(a_j, \delta_j) = 1, \quad u(a_j, \delta_l) = -c, \text{ if } l \neq j. \tag{2}$$

Thus, if $p_j = \Pr[\delta^* = \delta_j]$ denotes the probability which the candidate actually associates to the event that δ_j is the true answer, then it follows from (1) that her expected utility for each possible alternative is given by

$$\overline{u}(a_0) = 0, \qquad \overline{u}(a_j) = (1+c)p_j - c, \quad j = 1, \ldots, k, \tag{3}$$

and, therefore,

$$\begin{array}{l} \overline{u}(a_j) > \overline{u}(a_l) \iff p_j > p_l \\ \overline{u}(a_j) > \overline{u}(a_0) \iff p_j > c/(1+c). \end{array} \tag{4}$$

Thus, to maximize her expected score in one particular question, the candidate must determine her most likely answer (or anyone of them if there are several equally likely), that is one *mode* of her belief distribution $\{p_1, \ldots, p_k\}$, and to mark such a mode in the examination paper if, and only if, its associated probability $p^* = \max_j p_j$ is larger or equal than $c/(1+c)$. It follows from (4) that the expected score in one question for a candidate which acts optimally is

$$\overline{u}(a^*) = \max_{j=0,\ldots,k} \overline{u}(a_j) = \max\{(1+c)p^* - c,\, 0\}. \tag{5}$$

In particular, the score that a candidate may *expect* in one question if she is *convinced* that a particular answer is true is one, while the score she may expect if she has no idea of what the true answer might be, so that her belief distribution ($p_j = 1/k$) is uniform over the k possible answers, is given by $\max\{(1+c)/k - c,\, 0\}$.

From the instructor's point of view, it may seem natural to set the penalty *cost* c of marking a wrong answer such that the expected score associated to random guessing would be zero, and this is achieved if, and only if, $(1+c)/k - c = 0$, that is if, and only if, $c = 1/(k-1)$, in

which case the optimal strategy of the candidate is to mark a mode if its associated probability p^* is such that $p^* \geq c/(1+c) = 1/k$. Since this condition is satisfied by the mode of *any* belief distribution over the k possible answers, the optimal strategy of the candidate would then *always* be to mark what she believes to be the more likely answer, no matter how small her associated probability might be, a less than satisfactory solution for an instructor who is interested in learning about the candidate's level of knowledge!

Example 1. In Spain, candidates to postgraduate medical schools are selected in the order established by the total score they achieve in a multiple-choice questionnaire (the *MIR* examination) which consists of $n = 250$ questions, each of which has $k = 5$ possible answers, of which one, and only one, is correct. Candidates are requested to mark one answer from each question or to leave it blank. A point is awarded for each correct answer, zero points are awarded for blank answers and $c = 1/3$ points are subtracted for each incorrect answer. If more than one possibility is marked, the answer is considered to be incorrect. Since, in this case, $c/(1+c) = 1/4$, the argument above shows that the candidate's optimal strategy is *to mark all questions such that*, $p^* \geq 1/4$, that is to mark all questions for which the probability of the more likely answer is, at least, $1/4$. In particular, this is automatically achieved if the candidate may rule out at least one of the answers. It also follows that random guessing with no knowledge (so that $p^* = 1/5$) diminishes the candidate's expected score. $\triangleleft$

3. THE RELEVANT DECISION PROBLEM

The conventional practice in multiple-choice examinations described above, where for each question candidates are supposed to mark as correct one of the possible alternative answers, is clearly inappropriate: the candidate is often encouraged to always mark an answer, no matter what her level of knowledge is, and the instructor has no way of discriminating between a correct answer from a knowledgeable candidate, and a correct answer obtained from random guessing. The axiomatic arguments mentioned in Section 1 prescribe that *all* uncertainties *must* be described by probabilities; hence, candidates should be required to specify a complete *probability distribution* over the possible answers, and not merely to quote a mode. Thus, for instance, an answer of the form $(0, 1, 0, \ldots, 0)$ would indicate that the candidate is *absolutely convinced* that the true answer is δ_2, while $(1/2, 1/2, 0, \ldots, 0)$ would indicate that she does not know whether the true answer is δ_1 or δ_2 but has ruled out all other possibilities, and $(1/k, 1/k, \ldots, 1/k)$ would indicate she has no relevant information about what is the true answer to the question posed.

Naturally, an appropriate evaluation function, or *scoring rule*, is needed to score each question as a function of the probability distribution written and the correct answer. We now turn to discuss the crucial question of how the scoring rule should be chosen. As before, we will still assume that the candidate wants to maximize her expected total grade, so that she has to maximize her expected score in each individual question.

From the candidate point of view, each question in this form of examination is a decision problem, where the set of alternatives is the class $Q = \{\boldsymbol{q} = (q_1, \ldots, q_m)\}, q_j \geq 0, \sum_j q_j = 1\}$, of probability distributions over the set $(\delta_1, \ldots, \delta_m)$ of possible answers, and the expected utility is the *expected score*

$$\overline{u}(\boldsymbol{q}) = \sum\nolimits_{j=1}^{m} u(\boldsymbol{q}, \delta_j)\, p_j, \tag{6}$$

where $u(\boldsymbol{q}, \delta_j)$ is the score awarded to a candidate who quotes the probability distribution $\boldsymbol{q}$ when the correct answer is δ_j and where, as before, $p_j = \Pr\{\delta^* = \delta_j\}$ is the personal probability

which the candidate associates to the event that the correct answer is δ_j. Note that, in principle, there is no reason to expect that the probability distribution $\boldsymbol{q}$ written by the candidate on the examination form will be the same as the distribution $\boldsymbol{p} = (p_1, \ldots, p_m)$ which describes her actual beliefs.

The instructor is naturally interested in encouraging honesty and, therefore, she should choose a scoring function $u(\boldsymbol{q}, \delta_j)$ whose expected value (6) is maximized if, and only if, $\boldsymbol{q} = \boldsymbol{p}$, that is if, and only if, the candidate writes in the examination form her true beliefs. A scoring rule with this property is called a *proper* scoring rule. The quadratic scoring function

$$\begin{aligned} u(\boldsymbol{q}, \delta_j) &= a\Big[1 - \sum\nolimits_{j=1}^{k} (q_j - \mathbf{1}_{\delta_j})^2\Big] + b_j, \quad a > 0, \quad b_j \in \Re, \\ &= a\Big[2q_j - \|\boldsymbol{q}\|\Big] + b_j, \quad \|\boldsymbol{q}\| = \sum\nolimits_{j=1}^{k} q_j^2, \quad a > 0, \quad b_j \in \Re, \end{aligned} \tag{7}$$

where $\mathbf{1}_{\delta_j}$ is the indicator function of δ_j, is perhaps the simplest *bounded* proper scoring rule. This imposes as a penalty some linear transformation of the squared euclidean distance between a perfect response (a degenerate distribution which gives probability one to the true answer) and the stated answer, $(q_1, q_2, \ldots, q_k)$; it also has the interesting additional property of remaining proper, that is encouraging honesty, even if the q_j's are *not* a priori required to be a probability distribution. The quadratic scoring function was introduced by Brier (1950); major developments were due de Finetti (1962, 1965); further discussion was provided by Savage (1971) and by Lindley (1982).

The arbitrary constants in (7) are fixed with side conditions; thus, if one point is awarded to a perfect reply giving probability one to the right answer, and zero points are awarded to the lack of knowledge described by a uniform distribution, the quadratic scoring rule (7) becomes standardized to

$$u(\boldsymbol{q}, \delta_j) = \frac{k}{k-1}\Big[2q_j - \|\boldsymbol{q}\|\Big] - \frac{1}{k-1}\,, \quad \|\boldsymbol{q}\| = \sum\nolimits_{j=1}^{k} q_j^2, \tag{8}$$

which, as one would anticipate, yields *negative* values to distributions which associate high probabilities to wrong answers; in particular, a candidate who gives probability one to a wrong answer would get, using (8), the negative score $-(k+1)/(k-1)$, that is, she would be penalized by $c = (k+1)/(k-1)$ points. The condition of zero points awarded to a uniform distribution automatically implies that blank questions may be treated as questions for which a uniform distribution is stated as the answer; thus, without loss of generality, it may be assumed that probability distributions are specified for all questions.

The result described above indicates that, in conventional multiple-choice examinations, a far more serious penalty should be used for wrong answers than the often suggested value $c = 1/(k-1)$. Indeed, if a conventional multiple-choice questionnaire was scored using the penalty value indicated by the quadratic scoring rule, the analysis in Section 2 establishes that, to maximize her expected score, the candidate should only mark the mode of those questions for which the probability p^* associated to that mode verifies the inequality

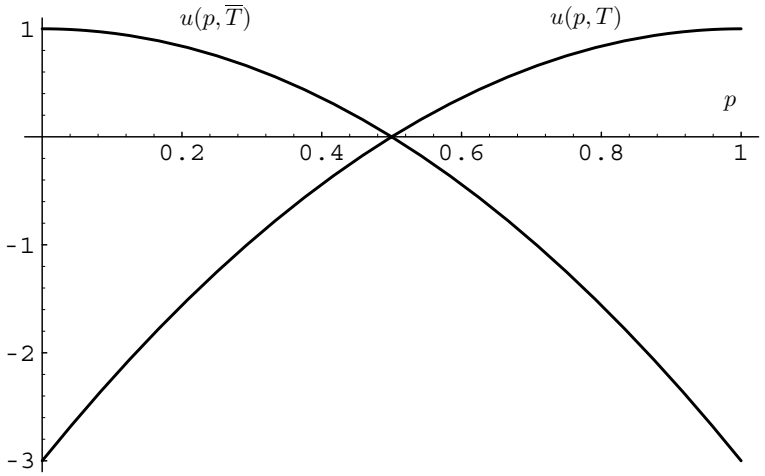
$$p^* \geq \frac{c}{1+c} = \frac{k+1}{2k} > \frac{1}{2}\,, \tag{9}$$

so that an 'absolute (probability) majority' is required, and random guessing in *not* encouraged.

Example 1. (continued). In the *MIR* example discussed before, where future medical post-graduates face a multiple-choice questionnaire with $k = 5$ possible answers associated to each question, one has

$$c = \frac{k+1}{k-1} = \frac{3}{2}\,, \qquad p^* \geq \frac{k+1}{2k} = \frac{3}{5}\,, \tag{10}$$

so that the quadratic scoring rule suggests a penalty of $3/2$ points for each incorrect answer (rather than the penalty $c = 1/3$ currently used!); if this were implemented, the optimal candidate's strategy would be to mark the mode of a question if, and only if, its associated probability were larger than $3/5$ (a kind of 'absolute majority'). This would immediately lead to a much better discrimination of good candidates against merely lucky candidates, with the corresponding obvious social benefits. $\triangleleft$



p	$u(p,T)$	$u(p,\overline{T})$
0.00	-3.00	1.00
0.10	-2.24	0.96
0.25	-1.25	0.75
0.50	0.00	0.00
0.75	0.75	-1.25
0.90	0.96	-2.24
1.00	1.00	-3.00

Figure 1. *standardized quadratic scoring rule when $k = 2$, as a function of $p = Pr(T)$*

The case $k = 2$ is specially interesting both by its simplicity and by its pedagogic value: it simply requires to propose a statement and to ask the candidate to quote her probability, say p, of this statement being true, with no need for the instructor to provide alternative answers, which are often artificially confusing. If T denotes the event that the statement is true, then the scoring rule (8) simply reduces to

$$u(p,T) = 1 - 4(1-p)^2, \quad u(p,\overline{T}) = 1 - 4p^2, \tag{11}$$

plotted and tabulated in Figure 1.

Proper scoring rules heavily penalize false *presumed* knowledge. Thus, using the quadratic scoring rule described above, a totally ignorant candidate would have a total score of zero, which would *also* be the total score of a candidate with 75% perfectly correct answers ($p^* = 1$) and 25% total blunders ($p^* = 0$), while a candidate with 50% perfect answers and openly ignorant on the rest would get 50% of the maximum score. Thus, a candidate with strong *misconceptions* on 25% of the questions is considered far worst than one which only shows her knowledge on half the questions but *realizes* her ignorance on the rest. We believe this is appropriate: people should be taught that assigning high probability to false statements may be very damaging in real life (if treated by a less than knowledgeable doctor, most of us would rather prefer him to realize his ignorance than to be convinced of a false diagnosis!). Thus, proclaiming false statements with high probability should be seriously penalized.

A similar analysis may be carried out with any other *proper* evaluation function, that is one which encourages honesty by guaranteeing that the maximum expected score is obtained if, and only if, the stated probability distribution $\boldsymbol{q}$ is that describing the candidates actual beliefs, $\boldsymbol{p}$. An alternative candidate to (8) is the *logarithmic* scoring rule

$$u(\boldsymbol{q}, \delta_j) = a \log(q_j) + b_j, \quad a > 0, \quad b_j \in \Re, \tag{12}$$

which reduces to

$$u(\boldsymbol{q}, \delta_j) = 1 - \frac{\log(q_j)}{\log(k^{-1})} \,, \tag{13}$$

if, as before, it is required that one point is awarded to a perfect reply and zero to a uniform distribution. Introduced by Good (1950), the logarithmic scoring rule is the *only* proper scoring rule which *exclusively* depends on the probability q_j associated to the true answer δ_j (Savage, 1971), and also has a rather nice decision-theoretical interpretation (Bernardo, 1979). However, its use in the context of multiple-choice examinations is problematic because it is *not* bounded; thus, the score associated to a response which gives probability zero to the correct answer is $-\infty$. Although this is mathematically sound (the worst possible mistake is to dogmatically declare false the correct conclusion, and this has the worst possible score, namely $-\infty$), it would be rather rash in practice: it would mean to automatically exclude from the examination any candidate which commits such a mistake. Besides, to be proper, the logarithmic scoring rule requires the stated q_j's to be probability values, ($0 \leq q_j \leq 1$) while the quadratic scoring rule encourages honesty *even* if the candidate is allowed to write any numerical values for the q_j's.

4. MAXIMIZING THE PROBABILITY OF A MINIMUM GRADE

We will now turn to consider the situation where the candidate wants to maximize the probability of getting a minimum grade, such as that needed to simply pass the examination, or that required to qualify for a grant. This is a far more complex decision problem, for it involves a *joint* analysis of the answers given to *all* the questions posed.

Let us consider a multiple-choice questionnaire consisting of n questions with k alternative answers, one of which is assumed to be true, associated to each question. We denote by $\delta_i^* \in \{\delta_{i1}, \ldots, \delta_{ik}\}$ the correct answer to the i-th question, so that $\delta_i^* = \delta_{ij}$ means that δ_{ij} is the correct answer to the i-th question, and we denote by $\boldsymbol{\delta} = \{\delta_1^*, \ldots \delta_n^*\}$ the set of all correct answers. Furthermore, we represent the candidate's true beliefs on the questions posed by the $n \times k$ matrix $\boldsymbol{P}$, with rows

$$\{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_n\}, \quad \boldsymbol{p}_i = \{p_{i1}, \ldots, p_{ik}\}, \quad p_{ij} \geq 0, \quad \sum\nolimits_{j=1}^{k} p_{ij} = 1, \tag{14}$$

where p_{ij} is the probability which the candidate associates to the event that δ_{ij} is the correct answer to the i-th question.

4.1. *Conventional Multiple-Choice Questionnaires*

Suppose that a conventional multiple-choice examination is performed, so that, for each question, the candidate is required to mark one, and only one of its associated possible answers, or to leave the question blank. We denote by $a_{(i)} \in \{a_{i0}, a_{i1}, \ldots, a_{ik}\}$ the answer provided to the i-th question, so that $a_{(i)} = a_{i0}$ means that the candidate has left the i-th question blank, and $a_{(i)} = a_{ij}$ means that the candidate has marked the j-th answer within the i-th question, and we denote by $\boldsymbol{a} = \{a_{(1)}, \ldots, a_{(n)}\}$ the set of all answers, that is, the complete answer to the examination. As before, we assume that the scoring rule is such that

$$u(a_{i0}, \delta_{ij}) = 0, \qquad u(a_{ij}, \delta_{ij}) = 1, \quad u(a_{ij}, \delta_{il}) = -c, \text{ if } l \neq j. \tag{15}$$

It follows that, if r and s respectively denote the number of correct and incorrect answers contained in $\boldsymbol{a}$, therefore leaving $n - r - s \geq 0$ questions left blank, then the *total score* t associated to the set of answers $\boldsymbol{a}$ is given by

$$t = t(\boldsymbol{a}, \boldsymbol{\delta}) = r(\boldsymbol{a}, \boldsymbol{\delta}) - c\, s(\boldsymbol{a}, \boldsymbol{\delta}), \quad c \geq 0, \tag{16}$$

where, again, c is the penalty associated to one incorrect answer. The preferences of a candidate who wants to maximize her probability of getting an score equal or larger than t_0 are described by the dichotomous utility function

$$\begin{aligned} u(\boldsymbol{a}, \boldsymbol{\delta}) = u(t(\boldsymbol{a}, \boldsymbol{\delta})) &= 1, \quad \text{if } t \geq t_0 \\ &= 0, \quad \text{otherwise.} \end{aligned} \tag{17}$$

It follows that the expected utility $\overline{u}(\boldsymbol{a})$ associated to a particular answer $\boldsymbol{a}$ will therefore be

$$\overline{u}(\boldsymbol{a}) = \Pr[t \geq t_0 \,|\, \boldsymbol{a}, \boldsymbol{P}]. \tag{18}$$

It is easily established that the expected utility $\overline{u}(\boldsymbol{a})$ is maximized by *marking the m more likely modes*, where $t_0 \geq m \geq n$, and where the optimal value of $m = m(t_0, n, \boldsymbol{P}, c)$ depends both on the probability matrix $\boldsymbol{P}$ and on the penalty constant c. Computationally, this is a simple exercise, for it only involves $n - t_0 + 1$ evaluations of $\overline{u}(\boldsymbol{a})$; however this means that, in order to be able to implement her optimal strategy, the candidate would need access to a preprogrammed calculator. As one would expect, $m(t_0, n, \boldsymbol{P}, 0) = n$ so that, if there is no penalty for a wrong answer ($c = 0$), then the candidate's optimal strategy is to mark *all* the modes. Thus, *a null penalty strongly encourages blind guessing*.

Example 2. The theoretical knowledge required to obtain a driving permit is usually controlled with multiple-choice examinations, where a minimum score is required to pass and where there is no penalty for incorrect answers. For instance, in Spain the test consists of $n = 40$ questions, each of which has $k = 3$ possible answers, of which one, and only one, is correct. Candidates are requested to mark one answer from each question or to leave it blank. A point is awarded for each correct answer, zero points are awarded for blank answers and nothing is subtracted for incorrect answers. Candidates pass the exam if they get a minimum of $t_0 = 36$ correct answers. Since, in this case, $c = 0$, the argument above shows that the candidate's optimal strategy is *to mark all questions* using a relevant mode if available, or randomly guessing the correct answer if no knowledge is held. $\triangleleft$

4.2. *Probabilistic Multiple-Choice Questionnaires*

Let us now suppose that candidates are requested to specify their probability distributions over the sets of alternative answers to each question. We represent the candidate's statements on the questions posed by $\boldsymbol{a} = \{a_{(1)}, \ldots, a_{(n)}\}$ where $a_{(i)} \in \{q_{i0}, \boldsymbol{q}_i\}$ is the answer provided to the i-th question, so that $a(i) = q_{i0}$ means that the candidate has left the i-th question blank, and $a_{(i)} = \boldsymbol{q}_i = \{q_{i1}, \ldots, q_{ik}\}$ means that the candidate has marked $\{q_{i1}, \ldots, q_{ik}\}$ as the probability distribution which she associates with the answer to the i-th question.

The total score will typically be of the form

$$t(\boldsymbol{a}, \boldsymbol{\delta}) = \sum_{i=1}^{n} f(a_{(i)}, \delta_{ij}), \tag{19}$$

where $f(q_{i0}, \delta_{ij}) = 0$ and $f(\boldsymbol{q}_i, \delta_{ij})$ is an evaluation function possibly, but not necessarily, of the form specified by (8), and where the candidate preferences are again described by the utility function (17). If the evaluation function f is chosen such that the score associated to a uniform distribution is zero, than blank answers may be treated as those which specify a uniform

probability distribution and hence, without loss of generality, the set $\boldsymbol{a}$ of all the answers in the examination may be represented as the $n \times k$ matrix $\boldsymbol{Q}$, whose rows are

$$\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n\}, \quad \boldsymbol{q}_i = \{q_{i1}, \ldots, q_{ik}\}, \quad q_{ij} \geq 0, \quad \sum_{j=1}^{k} q_{ij} = 1, \tag{20}$$

where q_{ij} is the probability which the candidate *states* as her belief on the event that δ_{ij} is the correct answer to the i-th question.

The candidate's optimal strategy would then be described by that matrix $\boldsymbol{Q}$ which maximizes the expected utility

$$\overline{u}(\boldsymbol{Q}) = \Pr[t \geq t_0 \,|\, \boldsymbol{Q}, \boldsymbol{P}]. \tag{21}$$

The solution will depend on the evaluation function f which is used. Note that, in general, the optimal strategy for the candidate will *not* be to be honest, that is to write $\boldsymbol{Q} = \boldsymbol{P}$, therefore revealing her true beliefs, *even* if f is chosen to be a proper scoring rule. Intuitively, if she is convinced that she 'must' get a minimum of t_0 points, then she would tend to pretend to have enough knowledge to have a chance to get those t_0 points, even if this is bound to decrease her *expected* total score.

5. DISCUSSION

In our many years of teaching experience, a detailed analysis of the decision problems associated to multiple-choice examinations has proved to be a very valuable introduction to the main ideas in decision analysis. Moreover, when candidates are required to provide probabilities, and are evaluated using some proper scoring rule, multiple-choice examinations serve both to improve the general ability of the population to assess probabilities, and to faithfully gauge their level of knowledge. Thus, probabilistic multiple-choice examinations serve society far better than their conventional counterpart, for they enhance the candidates abilities to assess probabilities, and they help to discriminate really good candidates from those merely lucky.

We have found that, even if the conventional form of examination is used, setting the penalty value to $c = (k+1)/(k-1)$ as suggested by the standardized quadratic score (in which case the optimal strategy for the candidate is to mark a mode if, and only if, its associated probability p^* is such that $p^* \geq (k+1)/(2k) > 1/2$ and guessing is not encouraged), does result in an appreciable improvement on the value of the multiple-choice examinations as a measure of the candidates' level of knowledge.

We have also found that, in practice, best results are obtained with $k = 2$, that is with 'multiple' choice examimations which simply consist of a sequence of statements, for each of which candidates are required to state their probability of the statement being true. Indeed, the use of a sequence of statements which are either true or false, rather than a sequence of questions with k alternative answers, has at least two obvious advantages: (i) it disposes of the need to provide alternative, often confusing answers, and (ii) it avoids a possible criticism of evaluation functions on the ground that, within each question, some wrong answers may be more nearly correct than others. In any case candidates should have immediate access to the behaviour of the evaluation function which will be used, that is to the information provided in Figure 1 if, as we are suggesting, the standardized quadratic score with $k = 2$ is actually used.

We have mentioned in Section 4 that the use of proper scoring rules is only guaranteed to encourage honesty if the candidate's preferences are to maximize her expected score. Thus, if the instructor is interested in faithful results, she should use a proper evaluation function *and* she should try to motivate the candidates into maximizing their score, rather than merely getting

some minimum grade. Nevertheless, an interesting further development consists of determining whether there are conditions under which, using a proper evaluation function (so that the optimal strategy of a candidate interested in maximizing her score is to be honest), the optimal strategy of a candidate who wants to maximize the probability that her total score is larger or equal to some known minimum level,would *also* be to be honest.

It may also be interesting to study sequential decision problems within this context. Thus, depending on her preferences, a candidate may find it optimal to abandon an examination if, say, her expected score is not large, and she thus ensure another opportunity to take the examination.

ACKNOWLEDGEMENTS

The author is grateful to Jim Berger, Dennis Lindley and Tony O'Hagan for their comments on an earlier draft of this paper.

REFERENCES

Bernardo, J. M. (1979). Expected information as expected utility. *Ann. Statist.* **7**, 686–690.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Chichester: Wiley.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Month. Weather Rev.* **78**, 1–3.

Chernoff, H. (1962). The scoring of multiple-choice questionnaires. *Ann. Math. Statist.* **33**, 375-393.

de Finetti, B. (1962). Does it make sense to speak of 'Good Probability Appraisers'? *The Scientist Speculates: An Anthology of Partly-Baked Ideas* (I. J. Good, ed.). New York: Wiley, 257–364. Reprinted in 1972, *Probability, Induction and Statistics* New York: Wiley, 19–23.

de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. *British J. Math. Statist. Psychol.* **18**, 87–123. Reprinted in 1972, *Probability, Induction and Statistics* New York: Wiley, 25–63.

Good, I. J. (1950). *Probability and the Weighing of Evidence*. London : Griffin; New York: Hafner Press.

Hsu, J. S. J., Leonard, T. and Tsui, K.-W. (1991). Statistical inference for multiple-choice tests. *Psychometrika* **56**, 327–348.

Hutchinson, T. P. (1991). *Ability, Partial Information, Guessing: Statistical Modelling Applied to Multiple-Choice Tests*. Adelaide, Australia: Rumsby.

Hutchinson, T. P. (1993). Second attempts at multiple-choice test items. *J. Statist. Computation and Simulation* **47**, 108–112.

Klein, S. P. (1992). Statistical evidence of cheating in multpiple-choice tests. *Chance* **5**, 23–27.

Lierly, S. B. (1951). A note for correcting for chance success in objective tests. *Psychometrika* **22**, 63–73.

Lindley, D. V. (1982). Scoring rules and the inevitability of probability. *Internat. Statist. Rev.* **50**, 1–26 (with discussion).

Martín, A. and Luna, J. D. (1989). Test and intervals in multiple-choice tests: A modification of the simplest classical model. *British J. Math. Statist. Philosophy* **42**, 251–264.

Martín, A. and Luna, J. D. (1990). Multiple-choice tests: Power, lengh and optimal number of choices per item. *British J. Math. Statist. Philosophy* **43**, 57–72.

Nogaki, A. (1984). Some remarks on multiple-choice questions in competitive examinations. *Behaviormetrika* **16**, 13–19.

Pollard, G. H. (1985). Scoring in multiple-choice examinations. *Math. Scientist* **10**, 93–97.

Post, G. V. (1994). A quantal choice model for the detection of copying on multiple-choice examinations. *Decision Sciences* **25**, 123–142.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *J. Amer. Statist. Assoc.* **66**, 781–801. Reprinted in 1974 in *Studies in Bayesian Econometrics and Statistics: in Honor of Leonard J. Savage* (S. E. Fienberg and A. Zellner, eds.). Amsterdam: North-Holland, 111–156.

Solomon, H. (1955). Item analysis and classification techniques. *Proc. Third Berkeley Symp.* **5** (J. Neyman and E. L. Scott, eds.). Berkeley: Univ. California Press, 169–184.

Thissen, D. and Steinberg, L. (1984). A response model for multiple choice items. *Psychometrika* **49**, 501–519.