Efficient Scoring of Multiple-Choice tests

Alexis Direer *

Abstract

This paper studies the optimal scoring of multiple choice tests in which the marks for wrong selections and omissions jointly minimize the mean square difference between score and examinees' abilities. Examinees are loss averse and, as a result, reluctant to risk answers on the basis of their knowledge. I find that it is efficient to incentivize the lowest able to omit, except when the test has a very large number of items. The mark for omission is positive when the test size is limited and negative when it is large. Loss aversion generally improves estimators efficiency by spontaneously inducing more omission and thereby reducing the need to bias the mark upward to encourage omission. The model sheds light on the statistical properties of two widely used scoring methods, Number right scoring and Formula scoring.

Keywords : estimation theory; multiple choice tests; decision making; loss aversion

J.E.L. codes: A200, C930, D800

*

Alexis Direr: Univ. Orléans, LEO. Address: rue de Blois - BP 26739, 45067 Orléans Cedex 02. E-mail: alexis.direr@univ-orleans.fr. ORCID: 0000-0002-4459-7780.

I thank for useful comments Marcel Voia, Christoph Heinzel, the two referees of the Revue and participants of the 2019 AFSE Conference and 2024 Niort International Conference on Economic and Financial Risks.

1 Introduction

Multiple-choice tests are a popular type of assessment in education. They have several advantages like fast and easy scoring, wide sampling of the content and grading exempt from rater bias. A major drawback is the difficulty of dealing with guessing. Examinees who have no clues about which answer is right may still select one at random and reap a point if lucky. More generally, examinees have often partial knowledge and select answers which they judge more likely. While an incorrect selection is always the result of a lack of knowledge, a correct one may result either from knowing, supposing or guessing, without possibly telling the three apart.

Guessing adds an error component to scores. Suppose that a test-taker has a probability 0.5 of selecting the right option. She may be lucky and gets an average score of 60%, or unlucky and gets a score of 40%. In both cases, her success score is mismeasured. If the test consists of many items, the law of large numbers ensures that the measurement error converges to zero. But for practical reasons, most tests have a limited number of items. Insofar as the scoring rule is not intended to reward chance, efficient marks should adequately correct for it.

The scoring method should also take into account the possibility given to examinees to leave some items blank if they are unsure about the right option. The mark for omission is an estimator of average omitters' ability. Omission suppresses the uncertainty due to the chance factor but introduces another type of measurement error which stems from the inability of sorting examinees with different levels of partial knowledge. The problem is especially acute if a significant fraction of examinees omit.

How do the marks affect incentives also depends on the extent to which examinees are reluctant to risk answers on the basis of their knowledge. Several studies have shown that examinees do not answer all items even when expected mark from

guessing is greater than for omitting (Sheriffs and Boomer, 1954, Ebel, 1968, Cross and Frary, 1977, Bliss, 1980, Pekkarinen, 2015). Those observations are not consistent with examinees being risk neutral score maximizers. A departure from risk neutrality is introduced by assuming that examinees are loss averse: they dislike receiving a bad mark by a larger extent than they like getting a full mark when they are right. This creates a bias toward omission, which consequences for the design of the scoring rule are investigated.

To this end, a statistically efficient scoring model is studied which marks minimize a well specified measurement error function. The problem differs from a standard mean estimation procedure as the marks serve two purposes at once. They provide an estimation of ability through the computation of a score for every examinee, but they also influence examinees in their choice between selection and omission, which in turn changes the conditions under which abilities are estimated. To study to what extent those two objectives interact, the grading rule minimizes the mean square difference between examinees' scores and abilities. By estimating unobservable characteristics from observable decisions, the model shares conceptual similarities with McFadden's (1981) econometric framework for inferring latent variables from discrete choices.

I find that the efficient scoring rule is highly sensitive to the size of the test. When a limited number of items is proposed to examinees, answers by the less able are too noisy to allow accurate estimation of their ability. The efficient mark for omission is positive to induce the less able to omit and reveal their type. The fewer items, the more omitters there should be and the higher the mark for omission. Loss aversion generally improves estimators efficiency by spontaneously inducing more omission and thereby reducing the need to bias the omission mark upward. When the test has a large sample of questions, ability of low able examinees is estimated with more accuracy, eliminating the need to induce them to omit. The mark for omission drops to negative values so that all examinees answer.

Multiple choice tests as an assessment tool have a long history. They were first administered on a large scale during the World War I by the US Army to quickly identify the abilities of hundred of thousands of recruits (Ebel, 1979). Its adoption spread rapidly in various domains, like intelligence testing (Pintner, 1923) or in education. Kelly (1916) is the first researcher to report and investigate the use of multiple choice tests in measuring children reading skills. The standardization of the evaluation process proved to be particularly adapted to large scale and high stake exams, like the Scholastic Aptitude Test (SAT) and Graduate Record Examination (GRE), to take two prominent examples in the USA.

To what extent tests provide accurate and valid measures of ability, skills or educational achievement has been studied for more than a century by psychometrics, a research domain at the intersection of psychology and statistics. Many of its results have been incorporated into what is regarded today as classical test theory (see e.g., McDonald, 1999). It is based on the central assumption that a person's score on a test is the sum of a true score and an error score (Harvill, 1991). The research program has developed around two key concepts: reliability and validity. A measure is reliable if it produces similar results under consistent conditions. Reliable scores are reproducible from one test to another (Traub and Rowley, 1991). A valid measure is one that measures what it is intended to measure. A voluminous theoretical and empirical literature has applied those concepts to the properties of different scoring rules (e.g., Diamond and Evans, 1973; Burton, 2001; Lesage et al., 2013).

The model departs from psychometric studies in two ways. First, a special attention is paid to the interplay between the scoring rule, risk preferences and ability estimation. In most existing studies, risk preferences are not modeled or when they are, examinees are risk neutral. By posing the realistic joint assumption of loss aversion and narrow framing, examinees display a bias toward omission in accordance with the empirical literature (e.g., Akyol et al., 2016). Second, the literature has focused on ad hoc scoring rules in which the marks for wrong answers and omission

are not derived from first principles. The two marks are made endogenous here by explicitly modeling the deviation of actual score from true score, which is the score that examinees would obtain if their ability were observed.

A few articles have also made the marks endogenous. Espinosa and Gardeazabal (2010) simulate a model of optimal scoring with heterogeneous risk aversion and varying item difficulty. They find a relatively high penalty to dissuade guessing. Budescu and Bo (2015) simulate a model of optimal scoring with different assumptions (heterogeneous loss aversion and miscalibration of probabilities). They find that a negative penalty aggravates the score bias and standard deviation, and decreases the correlation between simulated and true scores. Akyol, Key and Krishna (2016) model the test-taking behavior of students in the field, and use the model to estimate their risk preferences. They then simulate counterfactual scoring rules and find that increasing the penalty for wrong answers has a significant impact on omission, which in turn improves estimation of examinees' abilities. Risk aversion heterogeneity has little influence on simulated scores, which makes the case for negative penalty. In those articles, only the penalty for wrong answers is optimized, whereas both the marks for wrong answers and omission are endogenous in the present model. Another major difference is the use of the widely adopted mean squared error to measure the quality of the estimators, which allows analytical results and simple interpretations. By assuming that examinees only differ by their knowledge, and not personality traits like risk aversion, the present model does not address the issue of the impact of heterogeneous preferences on measures' validity.

The remainder of the paper is organized as follows. Section 2 presents the scoring model and its basic ingredients: true score, loss aversion and mean squared error. Section 3 put forth several analytical properties of the efficient scoring model. Section 4 calibrates a stylized model and presents simulation results. Section 5 relates and compare the model to the two most used scoring rules, Number right scoring and Formula scoring. Section 6 concludes.

2 Scoring model

2.1 Scoring rule

A test composed of n items is taken by examinees. Each item has m possible answers, one correct and $m-1$ incorrect. Items are supposed to be well written, without obvious answers, traps, or ambiguous formulations. Options are correctly randomized within each item. There is enough time for all questions to be answered. I assume further that all items are equally difficult and that examinees have a constant probability p of correctly answering any of them. The probability varies across examinees and is a proxy of their knowledge of the content area covered by the test.

The test-maker's objective is to design a scoring rule so that examinees receive a score as close as possible to their ability. Every item has three possible outcomes to which are assigned specific marks. The mark given to a correct selection is normalized to 1. The mark assigned to wrong selections is denoted θ and the one to omissions γ. Minimal restrictions are imposed on the marks:

$$\theta \leq \gamma < 1$$

The final score is the summation of marks obtained for all items divided by the number n of items. Let $z \in [0, n]$ be the number of omitted items and $\tilde{x} \in [0, n-z]$ the number of right selections among the $n-z$ answered items. $\tilde{x}$ follows a binomial distribution $B(n-z, p)$. Examinees' score is the sum of right answers, wrong answers and omitted items weighted by 1, θ and γ respectively, and divided by the number of items:

$$\tilde{s} = \frac{\tilde{x} + \gamma z + \theta(n - z - \tilde{x})}{n}$$

2.2 True score

True score $s(p)$ is defined as the expected score obtained in a test with marks 1 for right selections and θ^* for wrong ones, assuming examinees do not omit:

$$s(p) = p + (1-p)\theta^*$$

It is the observed score's component uninfluenced by random events (Harvill, 1991). True score will in general be lower than p ($\theta^* < 0$) to penalize guessing. An examinee selecting options at random would otherwise obtain a strictly positive score: $s(p) = p = 1/m$, with m the number of options per item. One way to eliminate the chance factor consists in setting $\theta^* = -1/(m-1)$, as in Formula scoring (see Section 5). Examinee's expected score is zero in the case they select options at random in all items:

$$E\Big(s\big(\frac{1}{m}\big)\Big) = \frac{1}{m} - \frac{m-1}{m}\frac{1}{m-1} = 0 \tag{1}$$

Other corrections are possible. If some examinees have false knowledge, they could perform worse than selecting options at random. In many situations, being aware of one's ignorance about a topic is preferable to having false knowledge about it. The first case is likely to encourage individuals to search for information, whereas the second case may lead individuals to make wrong decisions. In this case, pure guessing (or omission) reflects a minimal ability which could be rewarded by setting θ^* above $-1/(m-1)$. The case of misinformation is ruled out in this paper by assuming that examinees' lowest ability, denoted p_0, is equal to $1/m$.

2.3 Risk Preferences

Omission delivers a sure mark compared to selection, unless examinees are certain about which option is right. The choice between a sure outcome and a risky one is modeled through three assumptions. First, examinees get utility $u(x)$ from mark x

of every item, and not from average or aggregate score. Narrow framing (Tversky and Kahnemman, 1981), the assumption that people do not pool all sources of risk before deciding, has proven relevant in various contexts of decision involving multiple risks (Tversky and Kahnemman, 1981, Read, Loewenstein and Rabin, 1999).

Second, examinees focus on losses and gains and overweight losses. They are more affected by negative outcomes than by positive ones of same magnitude. Loss aversion is a central feature of Kahneman and Tversky's (1979) prospect theory of how people evaluate risks. Its validity is based on extensive experimental evidence, particularly when associated with narrow framing. Bereby-Meyer, Meyer and Flascher (2002) provide evidence of narrow framing and loss aversion in the context of exam taking.[1]

Third, the utility derived from a positive or negative mark is linear: $u(1) = 1$, $u(\gamma) = \gamma$. Applied to the context of exam taking, the utility loss associated with a wrong selection is proportional to the mark: $u(\theta) = \lambda\theta$, with λ the coefficient of loss aversion. A wrong selection is edited as a loss by examinees whatever the mark's sign: $\lambda > 1$ if $\theta \leq 0$ and $\lambda < 1$ if $\theta > 0$. Loss aversion is synthetically defined by the sign condition

$$\theta(\lambda - 1) \leq 0$$

Loss neutrality is equivalent to risk neutrality if $\lambda = 1$. Loss averse examinees do not like risk. They always prefer a sure mark to a random one with the same expectation.

Given a scoring rule (γ, θ), omission is preferred to selection if the mark for

[1] See also Budescu and Bo (2015). The joint assumption that people tend to focus on individual gains and losses rather than on average outcomes is sometimes labeled myopic loss aversion (Barberis, Huang, and Thaler, 2006; Barberis and Huang, 2008). Narrow framing is also in accordance with observations showing that individuals do not become risk neutral when they take large tests involving many independent items, which risk vanishes once aggregated (Pekkarinen, 2015; Akyol, Key and Krishna, 2022; Iriberri and Rey-Biel, 2021).

omission is greater than the loss-weighted expected mark of a response:

$$\gamma > p + (1-p)\lambda\theta$$

$\bar{p}$ is defined as the success probability of test-takers indifferent between choosing and omitting:

$$\gamma = \bar{p} + (1-\bar{p})\lambda\theta$$

Examinees omit when they are not confident enough in their selection: $p \leq \bar{p}$, and answer in the contrary case. $\bar{p}$ positively depends on mark γ and negatively on penalty θ. Compared to the case of risk neutrality ($\lambda = 1$), loss aversion raises the threshold probability $\bar{p}$:

$$\bar{p} = \frac{\gamma - \lambda\theta}{1 - \lambda\theta} > \frac{\gamma - \theta}{1 - \theta} \qquad \text{if } \lambda > 1 \tag{2}$$

2.4 Mean squared error

Examinee's true score is estimated through respondent's success rate in case of answer, or by assigning a constant mark in case of omission, which signals a low ability on average. Both methods produce measurement errors.

Consider first an examinee whose success probability is $p > \bar{p}$. Since p does not vary across items and all items have the same difficulty, she answers all of them and gets the score:

$$\tilde{s} = \frac{\tilde{x} + (n - \tilde{x})\theta}{n} \tag{3}$$

which is interpreted as a point linear estimator of true score $s(p)$. Its quality can be measured by common statistical methods and optimized by the adequate choice of θ and γ. The mean squared error (MSE) of observed score $\tilde{s}$ taken by examinee with success probability p is the average squared difference between $\tilde{s}$ and true score $s(p)$:

$$\text{mse}(\theta; p) = E\big(\big(\tilde{s} - s(p)\big)^2\big)$$

The MSE is a commonly used measure of estimator's performance. It is analytically tractable and lends itself to the intuitive decomposition:

$$E\big((\tilde{s}-s(p))^2\big) = V(\tilde{s};p) + \big(E(\tilde{s};p)-s(p)\big)^2$$

The first component is observed score's variance. The second one is squared bias, which measures by how far the expected score deviates from its theoretical mean. The MSE criterion controls both for sample fluctuations and estimator's accuracy.

Consider now a test-taker whose success probability is $p \leq \bar{p}$. Since p is constant across items, she omits all of them and gets the score γ. She would obtain the true score $s(p)$ if her ability was perfectly measured. Examinee's quadratic error is the squared bias:

$$sb(\gamma;p) = \big(s(p)-\gamma\big)^2$$

While individual success probabilities are not observed by the test-maker, their distribution is assumed to be known. Let $f(p)$ denote the success probability density function. The test-maker chooses the marks θ and γ so as to minimize the MSE averaged over examinees:

$$\begin{aligned}\min_{\gamma,\theta} \mathrm{MSE}(\gamma,\theta) &= \int_{p_0}^{\bar{p}} \mathrm{sb}(\gamma;p)f(p)dp + \int_{\bar{p}}^{1} \mathrm{mse}(\theta;p)f(p)dp \\ &= \int_{p_0}^{\bar{p}} \big(s(p)-\gamma\big)^2 f(p)dp + \int_{\bar{p}}^{1} E\Big(\big(\tilde{s}-s(p)\big)^2\Big)f(p)dp \qquad (4)\end{aligned}$$

Like the MSE component from selection, the MSE component from omission, normalized by their proportion $F(\bar{p})$ in the population, lends itself to a decomposition. Let us first rewrite the average squared error:

$$\frac{1}{F(\bar{p})}\int_{p_0}^{\bar{p}} \big(s(p)-\gamma\big)^2 f(p)dp = E_{|\mathrm{omit}}\big((s(p)-\gamma)^2\big)$$

where $E_{|\mathrm{omit}}$ is expectation conditional on examinees being omitters. Let us denote $\bar{s}(p)$ as omitters' average ability:

$$\bar{s}(p) = E_{|\mathrm{omit}}\big(s(p)\big) = \frac{1}{F(\bar{p})}\int_{p_0}^{\bar{p}} s(p)f(p)dp \qquad (5)$$

and omitters' ability conditional variance as:

$$V_{|\text{omit}}\big(s(p)\big) = E_{|\text{omit}}\Big(\big(s(p) - \bar{s}(p)\big)^2\Big) \tag{6}$$

The MSE component from omissions writes:

$$\frac{1}{F(\bar{p})}\int_{p_0}^{\bar{p}} \big(s(p) - \gamma\big)^2 f(p)dp = V_{|\text{omit}}\big(s(p)\big) + \big(\bar{s}(p) - \gamma\big)^2$$

Total omitters' measurement error has two components. The variance term measures how far omitters' ability deviates from its mean. The more omitters (the higher $\bar{p}$), the larger the dispersion and the higher the MSE. The second term is squared bias which measures by how far the mark deviates from omitters' average ability. It follows that, as long as some examinees omit, the variance term in (6) is a lower bound whatever the number of items in the test. This is a major difference with the MSE component from answers where the average error can always be brought to zero with n large enough.

3 Efficient scoring

When the number of items grows larger, ability is estimated from answers with unbounded accuracy. To the contrary, since omission signals low ability only on average, omitters create measurement errors which do not vanish with test length. It follows that when the number of items in the test is arbitrarily large, scores are perfect estimators of ability provided no examinees omit. The efficient score is such that the MSE tends to zero, statistical score converges to true score ($\hat{\theta} \to \theta^*$) and even the least able answer $\hat{\gamma} < p_0 + (1 - p_0)\lambda\theta^*$.

When the number of items is finite, examinees' abilities are estimated with errors due to finite-sample fluctuations. First order conditions of the minimization

program (4) are:

$$\frac{\partial \text{MSE}}{\partial \gamma}(\gamma, \theta) = \big(sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p})\big)\frac{d\bar{p}}{d\gamma}f(\bar{p}) + \int_{p_0}^{\bar{p}} \frac{\partial sb}{\partial \gamma}(\hat{\gamma}; p)f(p)dp = 0 \tag{7}$$

$$\frac{\partial \text{MSE}}{\partial \theta}(\gamma, \theta) = \big(sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p})\big)\frac{d\bar{p}}{d\theta}f(\bar{p}) + \int_{\bar{p}}^{1} \frac{\partial mse}{\partial \theta}(\hat{\theta}; p)f(p)dp = 0 \tag{8}$$

The common term in both equations

$$sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p}) = \big(\hat{\gamma} - s(\bar{p})\big)^2 - E\big((\tilde{s} - s(\bar{p}))^2\big) \tag{9}$$

is the net effect on the MSE of marginal examinees with ability $\bar{p}$ changing their choice from selection to omission. The terms $d\bar{p}/d\gamma$ and $d\bar{p}/d\theta$ are the effects of the marks on threshold probability $\bar{p}$ (see (2)). Raising γ or lowering θ both encourage omission and expand the group of omitters:

$$\begin{aligned} \frac{d\bar{p}}{d\gamma} &= \frac{1}{1 - \lambda\hat{\theta}} > 0 \\ -\frac{d\bar{p}}{d\theta} &= \frac{(1 - \bar{p})\lambda}{1 - \lambda\hat{\theta}} > 0 \end{aligned} \tag{10}$$

4 Simulated scoring

Since this problem has no analytical solutions in the general case, simulated scoring rules are presented in the next section.

4.1 Simulation strategy

Tversky and Kahneman (1992) estimate a loss aversion coefficient $\lambda = 2.25$ in Cumulative prospect theory. Since it not entirely clear how a parameter estimated from choices involving monetary outcomes translates to the context of grades, three plausible levels of loss aversion are considered: loss neutrality ($\lambda = 1$), moderate loss aversion ($\lambda = 1.5$) and strong loss aversion ($\lambda = 2.5$).[2]

[2]Mistakes could be positively marked ($\theta > 0$) in theory. Wrong answers would still be edited as a loss by examinees, i.e. $\lambda \in (0, 1)$. This situation never happens in simulations.

Actual ability distributions are expected to vary with test's difficulty relative to examinees' proficiency. Some distribution may be U-shaped with two modes close to the bounds (absence of knowledge and perfect ability). Others may be bell-shaped with a higher proportion of examinees around mean ability. Estimating the ability distribution from real tests is beyond the scope of this article. Without population and exam-specific information, I choose a simple uniform distribution over the space of ability $[p_0, 1]$.

The MSE (4) is computed over a double grid of values for parameters $\theta \in [\underline{\theta}, \theta^*]$ and $\bar{p} \in [1/m, 1]$. The mark γ is retrieved for each couple $(\theta, \bar{p})$ by the condition $\gamma = \bar{p} + (1 - \bar{p})\lambda\theta$. The two grids are composed of 2500 points each, so that $2500^2 = 6,250,000$ different values of MSE are computed. The efficient marks correspond to the lowest value calculated.

I use as a metric of fitness the root mean square error (RMSE), the geometric mean of measurement errors for all examinees:

$$\mathrm{RMSE}(\gamma, \theta) = \sqrt{\int_{p_0}^{\bar{p}} \big(\gamma - s(p)\big)^2 f(p)dp + \int_{\bar{p}}^{1} E\big((\tilde{s} - s(p))^2\big) f(p)dp}$$

I also compute the bias on omitters' score $\hat{\gamma} - \bar{s}(p)$ (see its expression (5)), which informs about to what extent omitters' ability estimator is distorted to encourage (if positive) or dissuade (if negative) omission. The incentives to omit are measured by the mark differential $\gamma - \theta$.

4.2 Baseline results

I first study a baseline model in which various numbers of items ($n = 1$, 5, 10, 20, 40, 80, 200, ∞) are considered. Each item has $m = 3$ options.[3] True scores $s(p)$ are computed given a notional mark correcting for pure guessing: $\theta^* = -1/(m-1)$.

[3]Score formulas with $m = 2$ to 5 options have also been simulated. Results, not included in the paper, show that error measures do not vary much, provided that the total number of options $m \times n$ in the test remains constant.

The loss aversion coefficient is set to 1.5.

Table 4 in Appendix A presents the efficient marks and main statistics in function of test's size n for the baseline calibration. Figure 1 in Appendix B shows how efficient mark $\hat{\gamma}$ varies with n.

Two distinct scoring strategies emerge depending on the test's size. When the number of items is limited (less than 170 in Figure 1), omission of the less able individuals is encouraged to palliate the estimation inaccuracy from answers of this group. The mark for omission is positive and set above average omitters' ability (see omission bias in table 4).

When the number of items is large enough, omission is discouraged altogether. Figure 2 shows that the proportion of omitters drops to zero. The mark differential $\hat{\gamma} - \hat{\theta}$, which measures the incentives to omit, decreases from 0.6 to 0.33. Therefore, it is efficient to force selection when the test has a sufficient number of items.

The intuition is as follows. Since the less knowledgeable individuals select options with little knowledge, the estimation error of their ability from their answers is high. It is more efficient to induce them to omit and thereby reveal their low ability. As more items are included in the test, abilities are estimated with increasing precision, even for the less able.

Note that, although omission can be an efficient way of reducing estimators variance, it is still subject to a trade-off between two types of measurement errors. Pooling too many omitters would not signal much information about their true ability (the variance term (6) would be large).

Except in the extreme case $n = 1$ in which more than 80% of examinees omit, the efficient penalty $\hat{\theta}$ is greater than the notional mark θ^* (Table 4), i.e. the penalty is milder than what prescribes a mere correction for guessing. It lies nevertheless in the close neighborhood of the notional mark, suggesting that a fixed penalty equal to the notional mark might prove a good approximation of the efficient rule. The mark

for omission is more sensitive to test's size n than the penalty for wrong answers. The behavior of the low able is indeed better targeted by the mark for omission than by the penalty which impacts all examinees, including the most proficient who never omit.

4.3 Efficient scoring and risk preferences

To what extent do risk preferences interact with the scoring rule and the efficiency of estimators? Loss-averse examinees overweight the utility loss when they get an answer wrong and tend to abstain more often. The consequences for omission, however, are ambivalent when the scoring rule is efficient (see Table 1). The proportion of omitters increases with loss aversion when omission is allowed, but the switch to a no-omission regime occurs sooner as the number of items increases.

Table 1: Efficient proportion of omitters (%) and loss aversion

Number of items (n)	1	5	10	20	40	80	200	∞
Risk neutrality ($\lambda = 1$)	43.4	26.0	19.6	14.4	10.5	7.6	4.9	0.00
Moderate loss aversion ($\lambda = 1.5$)	83.5	39.4	30.6	24.9	21.1	18.6	0.00	0.00
Strong loss aversion ($\lambda = 2.5$)	85.7	46.8	38.0	32.8	29.9	0.00	0.00	0.00

Model: notional mark corrects for pure guessing ($\theta^* = -0.50$). See Tables 3, 4 and 5 for detailed statistics. Reading: 26% of risk neutral examinees omit in an efficient test with 5 items.

To understand why, recall that the proportion of omitters depends on the incentives to omit given by the marks. Reducing the proportion of omitters with n requires to bias the mark for omission down, which comes with its own costs in terms of measurement errors. The more loss-averse examinees are, the lower the mark for omission needed to achieve a desired share of omitters. Therefore, it is

less costly to transition to a no-omission regime when n is large enough.

Loss aversion promotes efficiency to some extent, as shown by root mean squared errors reported in Table 2. Errors are decreasing with loss aversion for tests with a limited number of items $n \leq 40$. There are no visible differences for tests with larger n.

Table 2: Root mean squared error (RMSE) and loss aversion

Number of items (n)	1	5	10	20	40	80	200	∞
Risk neutrality ($\lambda = 1$)	0.406	0.241	0.181	0.133	0.097	0.070	0.045	0.00
Moderate loss aversion ($\lambda = 1.5$)	0.386	0.221	0.166	0.122	0.089	0.066	0.046	0.00
Strong loss aversion ($\lambda = 2.5$)	0.324	0.196	0.151	0.119	0.097	0.072	0.046	0.00

Model: notional mark corrects for pure guessing ($\theta^* = -0.50$). See Tables 3, 4 and 5 for detailed statistics.

When the number of items is finite, it is generally efficient to induce the less able to omit. If examinees are loss averse, the less able spontaneously omit without the need to distort the marks.

5 Relation to existing scoring methods

The model sheds lights on the efficiency of the two most used scoring methods, number right scoring (NRS) and formula scoring (FS). NRS simply counts the number of right selections and divides the sum by the total number of items. Omitted items and wrong selections count for zero (θ, $\gamma = 0$). A critic often made to the method is that examinees selecting options at random obtain a positive expected score in expectation equal to $1/m$. FS also sets $\gamma = 0$ but imposes a penalty for incorrect selection equal to $-1/(m-1)$. The formula equalizes the expected scores of pure

guessing and omission as shown in (1) (Thurstone, 1919, Holzinger, 1924).

The superiority of one of those rules to the other is still debated in the psychometric literature. By implementing negative marking, FS encourages omission, which increases reliability (Lord, 1975, Mattson, 1975, Burton, 2001). Some authors have argued that FS not only measures the mastery of domain knowledge but also reflects examinees' answering strategies and risk-taking behavior (e.g., Votaw, 1936; Frary, 1988; Budescu and Bar-Hillel, 1993). NRS provides more incentives to answer all questions, which minimizes this type of bias.

The main shortcomings of NRS and FS compared to the present model is the way they treat omission. The marks are not adjusted for finite sample to induce sufficient omissions when the number of items is not large. They both set the mark for omission to zero, which is not efficient whatever test's size (see Figure 1). By setting the mark differential $\gamma - \theta$ to zero, NRS dissuades omission, which is only efficient for large tests. FS provides more incentives to omit but only by raising the penalty for incorrect answer. To the contrary, the model shows that for a broad range of test's size, the mark for omission is strictly positive as it fulfills two functions: partial knowledge is credited and omission is encouraged.

Whatever the method, the marks in FS and NRS are not derived from an explicit estimation model. For instance, the correction for guessing made in FS starts from the assumption that ignorant examinees choose to answer all items at random. The assumption is not consistent with the present model according to which examinees with insufficient knowledge should be induced to omit, not to answer. This implies that the targeted mark assigned to fully ignorant examinees is not zero but is strictly positive.

6 Conclusion

Three main lessons can be drawn from the scoring model. First, a test-maker should include a large number of items when feasible to exploit the law of large numbers. Including additional items proves to be an effective way to enhance score efficiency, especially for tests with a limited number of items. Numerical simulations suggest a number greater than 40 and as much as 100. However, writing large-sized multiple-choice tests requires skill and time.

Second, efficiency dictates to target a decreasing proportion of omitters with test length. If the number of items is large, ability is generally better estimated by selection than omission. Selection may be forced by setting a negative mark for omission. If the number of items is limited, omission may be encouraged by a positive mark. The fewer items, the more omission needed and the higher the mark. The resulting proportion of omitters is quite significant in small tests. The instructions given to examinees should be consistent with the scoring strategy. If the number of items is small, examinees should be encouraged to omit. In the contrary case, they should be instructed to answer all questions even if they are unsure about the correct answers.

Third, the behavior of low able examinees is better targeted by the mark for omission than by the penalty for wrong answers. The finding reformulates the longstanding debate about the relative advantages of Formula scoring and Number right scoring which exclusively focuses on the best value that the mark for wrong answers should take.

The model has made a number of simplifying assumptions which implications for the estimation strategy could be interesting to investigate in the future. First, experimental studies in psychology suggest that people are generally overconfident about their own knowledge (e.g., Keren, 1991; Yates, 1990). Overconfidence reduces the omission rate and may impact estimation efficiency, especially if the tendency

correlates with ability (Lichtenstein and Bishhoff, 1977; Heath and Tversky, 1991). A related issue is how to score misinformation, which arises when examinees have erroneous knowledge (Burton, 2004). Second, the tests could be more realistically modeled by considering items with varying difficulty, for instance if the difficulty of the test increases as examinees progress through the multiple choice questions. Examinees' probability of being right and their incentives to omit would fluctuate across items. It could then be interesting to adapt the marks for mistakes and omissions with item difficulty. Third, the model assumes that examinees only differ by their knowledge, and not by personality traits like risk or loss aversion or the degree of self-confidence. Recent research by Akyol et al. (2022) suggests that heterogeneity in risk preferences may affect how examinees respond to penalty incentives. Part of the randomness would come from unobserved heterogeneity, which would require more complex estimation procedures.

References

Akyol P., Key J. and K. Krishna (2022) "Hit or miss? Test taking behavior in multiple choice exams", *Annals of Economics and Statistics==*, 147, 3-50.

Barberis N., M. Huang, and R. Thaler (2006) "Individual preferences, monetary gambles, and stock market participation: A case for narrow framing", *American Economic Review* 96, 1069-1090.

Barberis N. and M. Huang (2008) "The loss aversion/narrow framing approach to the equity premium puzzle", Mehra R. (ed.) Handbook of the Equity Risk Premium. Elsevier Science, NBER version.

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

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.V. and M. Bar-Hillel (1993) "To guess or not to guess: A decision-theoretic view of formula scoring", *Journal of Educational Measurement*, 30 (4), 277-291.

Budescu D.V. and Y. Bo (2015) "Analyzing test-taking behavior: Decision theory meets psychometric theory", *Psychometrika* 80 (4), 1105-1122.

Burton R.F. (2001) "Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers", *Assessment and Evaluation in Higher Education*, 26 (1), 41-50.

Burton R.F. (2004) "Multiple choice and true/false tests: reliability measures and some implications of negative rking", *Assessment and Evaluation in Higher Education*, 29, 585-595.

Ebel R.L. (1968) "Blind guessing on objective achievement tests", *Journal of Educational Measurement* 5, 321-325.

Ebel R.L. (1979) Essentials of educational measurement (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Cross L.H. and R.B. Frary (1977) "An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests", *Journal of Educational Measurement* 14, 313-321.

Diamond J. and W. Evans (1973) "The correction for guessing", *Review of Educational Research*, 43, 181-191.

Espinosa M.P. and J. Gardeazabal (2010) "Optimal correction for guessing in multiple-choice tests", *Journal of Mathematical Psychology* 54 (5), 415-425.

Frary R.B. (1988) "Formula scoring of multiple-choice tests (correction for guess-

ing)", *Educational Measurement: Issues and practice*, 7, 33-38.

Harvill L.M. (1991) "Standard error of measurement", *Educational Measurement: Issues and Practice*, 10, 33-41.

Heath C. and A. Tversky (1991) "Preference and belief: Ambiguity and competence in choice under uncertainty", *Journal of Risk and Uncertainty*, 4 (1), 5-28.

Holzinger K.J. (1924) "On scoring multiple-response tests", *Journal of Educational Measurement*, 15, 445-447.

Iriberri N. and P. Rey-Biel (2021) "Brave boys and play-it-safe girls: gender differences in willingness to guess in a large scale natural field experiment", *European Economic Review*, 131, 103603.

Kahneman D. and A. Tversky (1979) "Prospect theory: An analysis of decision under risk", *Econometrica*, 47(2), 263-92.

Kelly F.J. (1916) "The Kansas silent reading tests", *Journal of Educational Psychology*, 7(2), 63-80.

Keren G. (1991) "Calibration and probability judgments: conceptual and methodological issues", *Acta Psychologica* 77, 217-273.

Lesage E., Valcke M. and A. Sabbe (2013) "Scoring methods for multiple choice assessment in higher education - Is it still a matter of number right scoring or negative marking?", *Studies in Educational Evaluation*, 39, 118-193.

Lichtenstein S. and B. Fischhoff (1977) "Do those who know more also know more about how much they know?", *Organizational Behavior and Human Performance* 20, 159-183.

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

Mattson D. (1975) "The effects of guessing on the standard error of measurement and the reliability of test scores", *Educational and Psychological Measurement*, 25,

727-730.

McDonald R.P. (1999) Test theory: A unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates.

McFadden, D. (1981) "Econometric models of probabilistic choice" In C.F. Manski & D. McFadden (Eds.), Structural Analysis of Discrete Data with Econometric Applications (pp. 198-272). MIT Press.

Pekkarinen T. (2015) "Gender differences in behaviour under competitive pressure: Evidence on omission patterns in university entrance examinations", *Journal of Economic Behavior and Organization*, 115, 94-110.

Pintner R. (1923) Intelligence testing. New York: Holt, Rinehart and Winston.

Read D., Loewenstein G. and M. Rabin (1999) "Choice bracketing", *Journal of Risk and Uncertainty*, 19 (13), 171-197.

Sheriffs A.C. and D.S. Boomer (1954) "Who is penalized by the penalty for guessing?", *Journal of Educational Psychology*, 45, 81-9.

Thurstone L.L. (1919) "A method for scoring tests", *Psychological Bulletin*, 16, 235-240.

Traub R.E. and Rowley G.L. (1991) "Understanding reliability", *Educational measurement: Issues and practice*, 10(1), 37-45.

Tversky A. and D. Kahneman (1981) "The framing of decisions and the psychology of choice", *Science*, 211, 453-458.

Tversky, A. and D. Kahneman (1992) "Advances in prospect theory: Cumulative representation of uncertainty", *Journal of Risk and Uncertainty*, 5, 297-323.

Votaw D.F. (1936) "The effect of do-not-guess directions on the validity of true-false or multiple-choice tests", *Journal of Educational Psychology*, 27, 698-703.

Yates J.F. (1990) Judgment and decision making, Englewood Cliffs, NJ: Prentice

Hall.

Appendix A Tables

Efficient scoring and risk preferences

Table 3: Scoring properties with risk neutrality ($\lambda = 1$)

Number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\theta}$	0.00	−0.36	−0.43	−0.46	−0.48	−0.49	−0.50	−0.50
$\hat{\gamma}$	0.62	0.33	0.23	0.17	0.12	0.08	0.05	0.00
$\hat{\gamma} - \hat{\theta}$	0.62	0.7	0.66	0.63	0.60	0.57	0.55	0.50
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	40.6	19.7	13.6	9.31	6.36	4.36	2.67	0.00
Omitters (%)	43.4	26.0	19.6	14.4	10.5	7.6	4.9	0.00
RMSE	0.406	0.241	0.181	0.133	0.097	0.070	0.045	0.000

Model: notional penalty corrects for pure guessing ($\theta^* = -0.50$), $\hat{\theta}$: efficient mark for wrong selections. $\hat{\gamma}$: efficient mark for omission. $\hat{\gamma} - \hat{\theta}$ measures the incentives to omit. $100\left(\hat{\gamma} - \bar{s}(p)\right)$ is omission bias with $\bar{s}(p)$ average omitters' ability. Omitters (%): share of examinees who omit. RMSE: root mean squared error. For $n = \infty$, $\hat{\gamma}$ is the highest mark inducing all examinees to answer (any lower value would also be efficient).

Table 4: Scoring properties with moderate loss aversion ($\lambda = 1.5$)

Number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\theta}$	-1.79	-0.48	-0.46	-0.45	-0.45	-0.46	-0.49	-0.50
$\hat{\gamma}$	0.59	0.30	0.22	0.16	0.11	0.08	-0.16	-0.17
$\hat{\gamma} - \hat{\theta}$	2.39	0.78	0.68	0.61	0.57	0.54	0.33	0.33
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	17.7	10.81	6.6	3.3	1.0	-0.7	0.00	0.00
Omitters (%)	83.5	39.4	30.6	24.9	21.1	18.6	0.00	0.00
RMSE	0.386	0.221	0.166	0.122	0.089	0.066	0.046	0.000

Model: notional penalty corrects for pure guessing ($\theta^* = -0.50$), $\hat{\theta}$: efficient mark for wrong selection. $\hat{\gamma}$: efficient mark for omission. $\hat{\gamma} - \hat{\theta}$ measures the incentives to omit. $100\left(\hat{\gamma} - \bar{s}(p)\right)$ is omission bias with $\bar{s}(p)$ average omitters' ability. Omitters (%): share of examinees who omit. RMSE: root mean squared error. For $n \geq 200$, $\hat{\gamma}$ is the highest mark inducing all examinees to answer (any lower value would also be efficient).

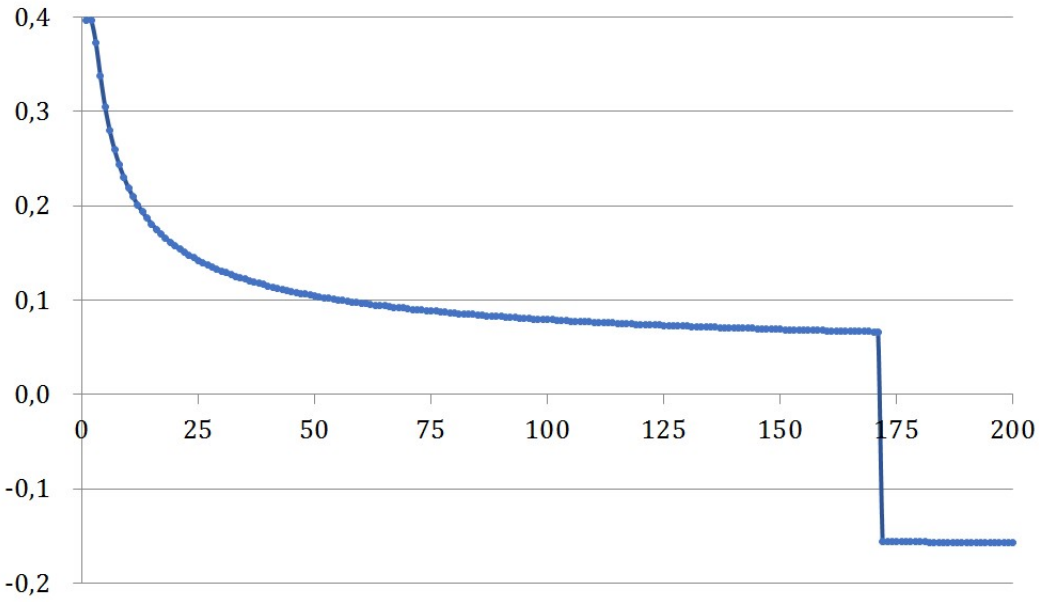
Table 5: Scoring properties with strong loss aversion ($\lambda = 2.5$)

Number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\theta}$	-1.64	-0.41	-0.36	-0.35	-0.34	-0.48	-0.49	-0.50
$\hat{\gamma}$	0.51	0.28	0.21	0.16	0.14	-0.46	-0.49	-0.50
$\hat{\gamma} - \hat{\theta}$	2.15	0.70	0.57	0.51	0.48	0.02	0.00	0.00
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	8.49	4.44	2.01	0.05	-1.27	0.00	0.00	0.00
Omitters (%)	85.7	46.8	38.0	32.8	29.9	0.00	0.00	0.00
RMSE	0.324	0.196	0.151	0.119	0.097	0.072	0.046	0.000

Model: notional penalty corrects for pure guessing ($\theta^* = -0.50$), $\hat{\theta}$: efficient mark for wrong selection. $\hat{\gamma} - \hat{\theta}$ measures the incentives to omit. $100\left(\hat{\gamma} - \bar{s}(p)\right)$ is omission bias with $\bar{s}(p)$ average omitters' ability. Omitters (%): share of examinees who omit. RMSE: root mean squarer error. For $n \geq 80$, $\hat{\gamma}$ is the highest mark inducing all examinees to answer (any lower value would also be efficient).

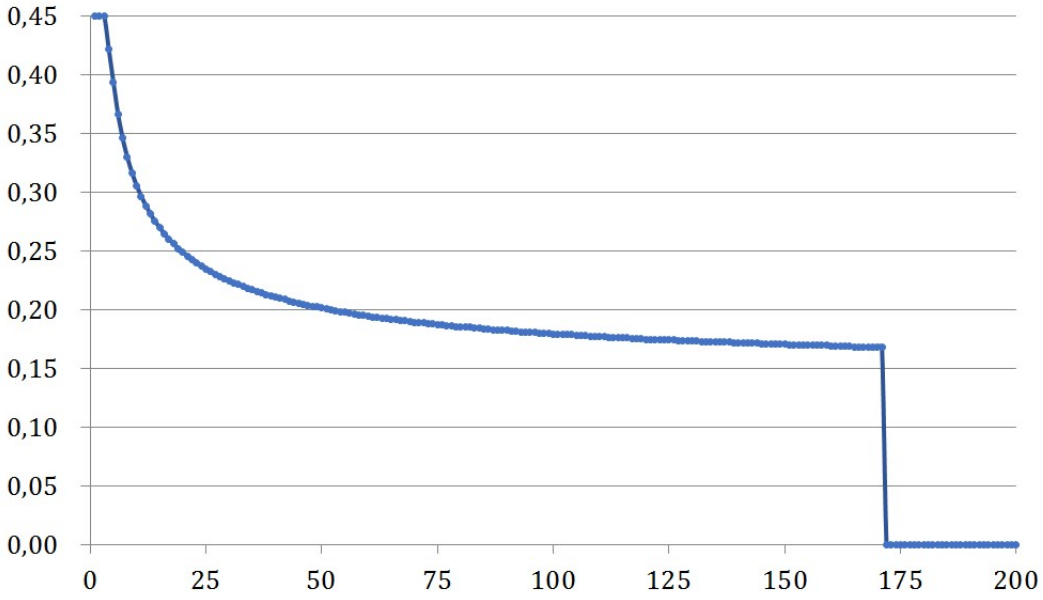
Appendix B Figures

Figure 1: Efficient mark for omission and number of items (horizontal line)



Model: notional mark corrects for pure guessing ($\theta^* = -0.50$), moderate loss aversion ($\lambda = 1.5$).

Figure 2: Proportion of omitters and number of items (horizontal line)



Model: notional mark corrects for pure guessing ($\theta^* = -0.50$), moderate loss aversion ($\lambda = 1.5$).