Efficient Scoring of Multiple-Choice Tests

Alexis Direr

Efficient Scoring of Multiple-Choice Tests

Alexis Direr *

June 2, 2020

Abstract

This paper studies the optimal scoring of multiple choice tests in which the marks for wrong selection and omission jointly minimize the mean square difference between score and examinees' ability. 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 in the first case and negative in the second case. The model sheds lights on the statistical properties of two widely used scoring methods, Number right scoring and Formula scoring.

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

Keywords : estimation theory, multiple choice tests, scoring rule, loss aversion

*Univ. Orléans, CNRS, LEO, FRE 2014. E-mail : alexis.direr@univ-orleans.fr. Phone: 33 (0)1 49 55 68 64. Address: rue de Blois - BP 26739, 45067 Orléans Cedex 02. I thank for useful comments Marcel Voia and participants of the 2019 AFSE Conference. This article has a companion website.

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 by examinees. 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 to be 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 proposes 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. The chance factor also depends on examinees' ability (proficient examinees rely little on luck) and the number of options per item (the more options the harder to select by chance the right option). 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 impossibility to distinguish examinees with different levels of partial knowledge. The problem is especially acute if a significant fraction of examinees omit.

The aim of the paper is to investigate a statistically efficient scoring method 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, I pose an estimation model in which the marks jointly minimize the mean square difference between examinees' scores and abilities. By using the mean square error as a fitness criterion, the error term can be usefully decomposed into a variance and a bias components. The model's finite and large sample statistical properties can also be conveniently distinguished.

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.

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 them to omit and reveal their low ability. The fewer items, the more omitters and the higher the mark for omission. Loss aversion generally improves estimators efficiency by inducing spontaneously more omission and thereby reducing the need to bias the mark upward to favor omission. When the

test has a large sample of questions, ability of low able examinees is estimated with accuracy, eliminating the need to induce them to omit. The mark for omission drops to negative values so that all examinees answer. The penalty for wrong answers is approximately insensitive to the number of items and the scoring strategy.

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 then 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 obtained 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 between true score (the score that examinees would obtain if their ability were observed) and actual score. The use of a well specified estimation model helps to clarify the objectives and criteria used in choosing a scoring method.

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 and find a relatively high penalty to dissuade guessing. Budescu and Bo (2015) also simulate a model of optimal scoring but 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' ability. 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 to obtain 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 answering correctly any of them. The probability varies across examinees and is a proxy of their ability in 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:

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

2.2 True score

True score is the score examinees would be credited if probabilities p of selecting the right option were perfectly observed. The score is linear in p:

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

and is interpreted as expected score obtained in a test with marks 1 for right selections and θ^* for wrong ones, assuming examinees do not omit. 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 extreme 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 an option at random. In many situations, being aware of one's ignorance about a topic is preferable to have wrong 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 $-1/(m-1)$.

2.3 Risk Preferences

Omission delivers a sure mark compared to selection, unless examinees are sure 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 useful 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

[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, 2016; Iriberri and Rey-Biel, 2018).

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 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 selection and omission:

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

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

Examinees' true score is estimated through respondents' 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 average squared difference between $\tilde{p}$ and true score $s(p)$:

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

MSE is a commonly used measure of estimators 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 therefore 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 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:

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

Like the MSE component from selections, the MSE component from omissions, 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:

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

and omitters' ability conditional variance as:

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

The MSE component from omissions writes:

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

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 (5) 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

While the full model cannot be studied analytically, three simple cases are of interest. In the first case, examinees' ability is estimated under the assumption that their choice to select or omit is insensitive to the marks (Subsection 3.1). In the second case, omission is endogenous, but the number of items is arbitrarily large (Subsection 3.2). In the last configuration, examinees are risk neutral (Subsection 3.3). The general case with a finite sample of items, omission, and loss aversion is studied in next section by way of simulations.

3.1 Exogenous omission

Let us assume as a first approximation that the proportion of omitters is exogenous and equal to $F(\bar{p})$. The scoring problem can be decomposed into two separate and simpler problems. The efficient mark $\hat{\theta}$ minimizes the part of the MSE restricted to the most proficient examinees who choose to select:

$$\int_{\bar{p}}^{1} \text{mse}(\theta; p) f(p) dp = \int_{\bar{p}}^{1} V(\tilde{s}; p) + \big(E(\tilde{s}; p) - s(p)\big)^2 f(p) dp$$

The efficient penalty $\hat{\theta}$ is smaller than the notional mark θ^*, and the discrepancy closes with test's size n:

Proposition 1 *(i)* $\hat{\theta} \in (\theta^*, 1)$*, (ii)* $\hat{\theta}$ *decreases with* n*, (iii)* $\hat{\theta} \to \theta^*$ *if* $n \to \infty$*.*

Proof Score's formula is given by (3) where $\tilde{x}$ follows a binomial distribution $B(n, p)$. Its first two moments are $E(\tilde{s}; p) = p + (1-p)\theta$ and $V(\tilde{s}; p) = (1-\theta)^2 p(1-p)/n$. It follows that:

$$\int_{\bar{p}}^{1} \text{mse}(\theta; p) f(p) dp = \int_{\bar{p}}^{1} \Big(\frac{1}{n}(1-\theta)^2 p(1-p) + (1-p)^2 \theta^2 \Big) f(p) dp$$

The efficient mark θ^* satisfies:

$$\frac{\hat{\theta} - \theta^*}{1 - \hat{\theta}} = \frac{1}{n} \frac{\int_{\bar{p}}^{1} p(1-p) f(p) dp}{\int_{\bar{p}}^{1} (1-p)^2 f(p) dp} \tag{7}$$

(i): the right hand term of (7) is positive. The case $\hat{\theta} > 1$ is ruled out by $\theta^* < 1$. (ii) and (iii) are straightforward from (7). □

A reduced penalty (θ closer to 1) lowers the score's variance but biases the estimator. As more items are included in the test, abilities are estimated with increasing precision, making less necessary to bias the mark to reduce statistical fluctuations.

The efficient mark $\hat{\gamma}$ minimizes the part of the MSE restricted to omitters:

$$\int_{p_0}^{\bar{p}} \text{mse}(\gamma; p) f(p) dp = \int_{p_0}^{\bar{p}} \big(s(p) - \gamma\big)^2 f(p) dp$$

The first order equation is:

$$2 \int_{p_0}^{\bar{p}} \big(s(p) - \hat{\gamma}\big) f(p) dp = F(\bar{p}) \big(\hat{\gamma} - E_{|\text{omit}}\big(s(p)\big)\big) = 0$$

Given an exogenous proportion $F(\bar{p})$ of omitters, the efficient estimator is an unbiased measure of omitters' average ability: $\hat{\gamma} = E_{|\text{omit}}\big(s(p)\big)$. When the proportion $F(\bar{p})$ is endogenous and responds to marks' variations, we will see that it may be efficient to bias the mark to induce a desired proportion of omission.

3.2 Large sample properties

When the number of items in the test is arbitrarily large, scores are perfect estimators of ability. Omission should be discouraged as a result, except possibly for the least able.

Proposition 2 *$\hat{\theta} \to \theta^*$ if $n \to \infty$; $\hat{\gamma} < p_0 + (1 - p_0)\lambda\theta^*$ if $\lambda > 1$, and $\hat{\gamma} \leq p_0 + (1 - p_0)\theta^*$ if $\lambda = 1$.*

Proof The MSE minimization program with a finite n is:

$$\min_{\gamma,\theta} \text{MSE}(\gamma,\theta) = \min_{\gamma,\theta} \int_{p_0}^{\bar{p}} \Big(\gamma - (p + (1-p)\theta^*)\Big)^2 f(p)dp + \int_{\bar{p}}^{1} \Big(\frac{1}{n}(1-\theta)^2 p(1-p) + (1-p)^2(\theta-\theta^*)^2\Big) f(p)dp$$

The variance term asymptotically tends to zero with n:

$$\lim_{n\to\infty} \text{MSE} = \int_{p_0}^{\bar{p}} \Big(\gamma - (p + (1-p)\theta^*)\Big)^2 f(p)dp + \int_{\bar{p}}^{1} \Big((1-p)^2(\theta-\theta^*)^2\Big) f(p)dp$$

Hence the MSE tends to zero when $\hat{\theta} \to \theta^*$ and $\hat{\gamma} < p_0 + (1-p_0)\lambda\theta^*$ such that all examinees answer, implying that the first integral is zero. If $\lambda = 1$, the omission condition $\hat{\gamma} = p_0 + (1-p_0)\lambda\theta^*$ induces the least knowledgeable to omit without deteriorating the MSE. □

When the number of items grows larger, ability is estimated with increasing accuracy. To the contrary, since omission signals low ability only on average, omitters create measurement errors which do not vanish with test length. Under risk neutrality ($\lambda = 1$), efficient marks may indifferently induce the least able to answer or to omit, as the unbiasedness condition coincides with the incentives given to them to omit.[2]

3.3 Finite sample properties

When the number of items is finite, examinees' ability is estimated with errors due to finite-sample fluctuations. First order conditions of the minimization program (4) are:

$$\frac{\partial \text{MSE}}{\partial \gamma}(\gamma,\theta) = \big(sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p})\big)\frac{d\bar{p}}{d\gamma} f(\bar{p}) + \int_{p_0}^{\bar{p}} \frac{\partial sb}{\partial \gamma}(\hat{\gamma};p) f(p)dp = 0 \tag{8}$$

$$\frac{\partial \text{MSE}}{\partial \theta}(\gamma,\theta) = \big(sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p})\big)\frac{d\bar{p}}{d\theta} f(\bar{p}) + \int_{\bar{p}}^{1} \frac{\partial mse}{\partial \theta}(\hat{\theta};p) f(p)dp = 0 \tag{9}$$

[2] If $\hat{\gamma} = p_0 + (1-p_0)\theta^*$, the least able are indifferent between selection and omission. If they omit, they get the unbiased mark $\hat{\gamma} = s(p_0)$. If they answer, they asymptotically obtain the same score.

The common term in both equations

$$sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p}) = \big(\hat{\gamma} - s(\bar{p})\big)^2 - E\big((\tilde{s} - s(\bar{p}))^2\big) \tag{10}$$

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:

$$\begin{aligned} \frac{d\bar{p}}{d\gamma} &= \frac{1}{1-\lambda\hat{\theta}} > 0 \\ -\frac{d\bar{p}}{d\theta} &= \frac{(1-\bar{p})\lambda}{1-\lambda\hat{\theta}} > 0 \end{aligned} \tag{11}$$

Suppose now that examinees are loss neutral ($\lambda = 1$). By definition, marginal examinees expect to obtain the same score whether they answer or omit:[3]

$$E(\tilde{s};\bar{p}) = \gamma \tag{12}$$

The replacement effect, a key element in conditions (8) and (9), simplifies to:[4]

$$sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p}) = -V(\tilde{s};\bar{p}) \tag{13}$$

The replacement effect is negative (the MSE decreases when marginal examinees change from selection to omission) and asymptotically tends to zero when the number of items in the test expands. The benefit of inducing more examinees to omit at the margin leads to the following proposition:

Proposition 3 *(i)* $\hat{\gamma} > p_0 + (1-p_0)\hat{\theta}$ *and (ii)* $\hat{\gamma} \to p_0 + (1-p_0)\theta^*$ *if* $n \to \infty$.

Proof First order condition (8) with (11) and Lemma 13 becomes:

$$\int_{p_0}^{\bar{p}} \frac{\partial sb}{\partial \gamma}(\hat{\gamma};p) f(p) dp = \frac{V(\tilde{s};\bar{p})}{1-\hat{\theta}} \geq 0$$

[3] $E(\tilde{s};\bar{p}) = \bar{p} + (1-\bar{p})\theta$. Using Definition (2) of $\bar{p}$ with $\lambda = 1$: $E(\tilde{s};\bar{p}) = (1-\theta)\dfrac{\gamma-\theta}{1-\theta} + \theta = \gamma$.

[4] Develop (10): $sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p}) = \big(\hat{\gamma} - s(\bar{p})\big)^2 - \big(E(\tilde{s};\bar{p}) - s(\bar{p})\big)^2 - V(\tilde{s};\bar{p})$, then use (12).

The left-hand term simplifies to:

$$\int_{p_0}^{\bar{p}} \frac{\partial}{\partial \gamma}\big(\hat{\gamma} - s(p)\big)^2 f(p)dp = 2\int_{p_0}^{\bar{p}} \big(\hat{\gamma} - s(p)\big) f(p)dp = F(\bar{p})\big(\hat{\gamma} - \bar{s}(p)\big)$$

with $\bar{s}(p) = E_{|\text{omit}}\big(s(p)\big)$ defined in (5). The left-hand side term $V(\tilde{s};\bar{p})/(1-\hat{\theta}) > 0$, which implies both $F(\bar{p}) > 0$ and $\hat{\gamma} > \bar{s}(p)$. The first inequality means that the less able omit ($\bar{p} > p_0$), which implies $\hat{\gamma} > p_0 + (1-p_0)\hat{\theta}$. If $n \to \infty$, $V(\tilde{s};\bar{p}) \to 0$ and as a result $F(\bar{p})\big(\hat{\gamma} - \bar{s}(p)\big) \to 0$. From Proposition 2, $\hat{\theta} \to \theta^*$, $F(\bar{p}) \to 0$ and only the least able answer asymptotically: $\hat{\gamma} \to p_0 + (1-p_0)\theta^*$. □

Since the less knowledgeable select options with little knowledge, it is efficient to induce them to omit and thereby to reveal their low ability. In addition, the mark for omission is biased above omitters' average ability: $\hat{\gamma} > \bar{s}(p)$ (See Proof of Proposition 3). This illustrates the double role of the marks in presence of endogenous choice between selection and omission. The mark for omission not only provides an estimation of omitters' ability but also gives extra incentives to omit. As the number of items in the test increases, the score becomes more accurate. The optimal subset of omitters shrinks and the bias applied to the mark for omission vanishes. At the limit, only the least able omit.

Whether the penalty for wrong answer is below or above the notional mark θ^* is ambiguous. On the one hand, $\theta > \theta^*$ lowers the error variance for an exogenous proportion of omitters, as explained in subsection 3.1. On the other hand, $\theta < \theta^*$ encourages omission which reduces measurement errors for low able examinees. Simulations in next section shows that the variance minimization argument will generally prevail ($\hat{\theta} > \theta^*$).

Note that, although omission is 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).

4 Simulated scoring

This section presents numerical results from the statistical model of scoring with omission, loss aversion and tests of finite size.

4.1 Simulation strategy

Regarding risk preferences, 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 ($\lambda = 2.5$).[5]

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

[5] 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.

mean of measurement errors for all examinees:

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

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. True scores $s(p)$ are computed given a notional mark correcting for pure guessing: $\theta^*=-1/(m-1)$. The loss aversion coefficient is set to 1.5.

Table 7 in Appendix A presents the efficient marks and main statistics in function of test's size n for the baseline calibration. In addition, Figure 1 in Appendix B shows how $\hat{\gamma}$ varies with n. Two distinct scoring strategies emerge in function of test's size. When the number of items is below a threshold (less than 170 in Figure 1) omission is encouraged to palliate estimation inaccuracy of less able's ability. The mark for omission is positive and above average omitters' ability (see omission bias in table 7). It gradually decreases to around 0.1 before jumping to negative values. Figure 2 shows that the proportion of omitters is also decreasing with n with a sudden fall to zero. The mark differential $\hat{\gamma}-\hat{\theta}$, which measures the incentives to omit, drops from 0.6 to 0.33. It is therefore efficient to force selection when the test has a sufficient number of items.

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 7), 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 (more in Subsection 4.5). 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 risk preferences interact with the scoring rule and estimators' efficiency? Loss averse examinees overweight utility loss when they get wrong and tend to abstain more often. The consequences for omission are however ambivalent when the scoring rule is efficient (see Table 1). On the one hand, the proportion of omitters increases with loss aversion. On the other hand, the stronger they are loss averse, the smaller the number of items above which omission is dissuaded.[6]

Table 1: 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: $m = 3$ options per item,the notional mark corrects for pure guessing ($\theta^* = -0.50$). See Tables 6, 7 and 8 for the detailed statistics. Reading: 26% of risk neutral examinees omit in a test with 5 items.

To understand why, recall that the proportion of omitters depends on the incentives to omit given by the marks. The more examinees are loss averse, the lower the

[6]Simulations show that omission is discouraged for $n > 171$ if examinees are moderately loss averse, and $n > 57$ if they are strongly loss averse. Some examinees still omit for $n = 200$ in case of risk neutrality.

mark for omission needed to achieve a desired share of omitters. For a sufficiently strong loss aversion, γ is below most true scores $s(p)$. Lowering it further raises measurement errors $sb(\gamma; p)$. Hence when loss aversion strengthens, the threshold test's size for which selection by low able examinees is efficient, is lowered.

Loss aversion promotes to some extent efficiency, 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: $m = 3$ options per item, notional mark corrects for pure guessing ($\theta^* = -0.50$). See Tables 6, 7 and 8 for the detailed statistics.

To get the intuition, recall that if examinees are risk neutral and the number of items is finite, it is efficient to induce the less able to omit (see Subsection 3.3). If examinees are loss averse, the less able spontaneously omit without the need to distort the marks.

4.4 Efficient scoring and test length

How many items should be included in a test? How many options should be proposed in every item? Is there a trade-off between the two margins? While the first

question has been rarely investigated in the psychometric literature,[7] the optimal number of options per item has been discussed at length (see Rodriguez (2005) for a survey).

Increasing the number of options generally increases the difficulty of the item (assuming all alternatives are plausible), which increases the likelihood that a test-taker will select a distractor item. Pure guessing becomes more hazardous. At the other end of the distribution, perfectly informed examinees select the right option whatever the number of distractors. This suggests that examinees with partial knowledge are expected to be confused by more distractors, but less so they are more able.

Varying the number of options from m to $m' > m$ changes the success probability of pure guessing and therefore minimal ability from $p_0 = 1/m$ to $p'_0 = 1/m' < p_0$. Let us consider an examinee whose ability is $p < 1$ with m options and $p' < p$ with $m' > m$ options. Assuming that examinees relative standings remain the same whatever the number of distractors: $F(p') = F(p)$, stretching the interval of probability from $[p_0, 1]$ to $[p'_0, 1]$ mechanically reduces the probability of a correct answer whatever the actual distribution of ability.

In the baseline model with a uniform ability distribution, the assumption $F(p') = F(p)$ gives p' in function of p, given m and m', or p_0 and p'_0:

$$p' = p'_0 + \frac{1 - p'_0}{1 - p_0}(p - p_0)$$

Figure 3 plots examinees ability in function of their relative rank for tests with two and five options per item. In accordance with intuition, the more able an examinee, the less affected by the inclusion of additional options per item. For instance, low able examinees whose rank is $F(p) = 0.1$ have 55% chance of correctly picking the right answer with two options, and only 28% with five options. At the other end, examinees whose rank $F(p)$ is 0.9 have 95% chance of success with two

[7] Burton and Miller (1999) is an exception.

options, and still 92% with five options.

Figure 4 shows how fast the root mean squared error (RMSE) declines with the number of items for $m = 2$ and 5 options per item. Efficiency gains from additional items are large for tests with few items, less than 25, whatever the number of options per item. The gain then decelerates rapidly and slowly converges to zero. It is around 0.05 for $n = 200$ and $m = 3$, and still 0.03 for $n = 1000$. The figure suggests that tests with more than 100 items do not seem to be worth devising, considering the time spent to construct and administer them.

Since the inclusion of additional distractors reduces the influence of blind or educated guessing, the RMSE in Figure 4 are logically decreasing with the number of options for a given number of items. Increasing the number of options from 2 to 5 significantly reduces the RMSE, even for large n where it becomes hard to reduce measurement errors by adding new items. The gains from increasing the number of options from 3 to 4 are smaller, and even so from 4 to 5 (see Table 3).[8]

Table 3: Root mean squared errors (RMSE) and number of options per item

Number of items (n)	1	5	10	20	40	80	200	∞
RMSE with 2 options	0.407	0.244	0.186	0.140	0.106	0.080	0.057	0.00
RMSE with 3 options	0.386	0.221	0.166	0.122	0.089	0.066	0.046	0.00
RMSE with 4 options	0.371	0.211	0.156	0.114	0.083	0.061	0.041	0.00
RMSE with 5 options	0.365	0.205	0.151	0.110	0.080	0.058	0.038	0.00

Model: notional mark corrects for pure guessing ($\theta^* = -1/(m-1)$), moderate loss aversion ($\lambda = 1.5$). RMSE are extracted from Tables 9 (2 options), 7 (3 options), 10 (4 options) and 11 (5 options).

One may wonder whether creating new items might be preferable to devising additional options, given a fixed number of options summed over all items. This

[8] See Burton (2001) for similar conclusions.

issue has practical relevance insofar as the total testing time is not extensible and is increasing with the number of options reviewed.[9] To check this point, we compare tests with varying number of items and options, but a constant total number of options, equal to 100.

Table 4: Root mean squared errors (RMSE) with 100 options

Number of options per item (m)	2	3	4	5
Number of items (n)	50	33	25	20
RMSE	0.096	0.098	0.103	0.110

Model: notional mark corrects for pure guessing ($\theta^* = -1/(m-1)$), moderate loss aversion ($\lambda = 1.5$). The total number of of options (number of items $\times$ number of options per item) is constant and equal to 100. See Table 12 for detailed statistics.

Table 4 shows that the RMSE hardly varies with the test composition. It is almost equivalent to administer a test with 50 items and two options or a test with 20 items and 5 options. The result rests however on the assumption that the test-maker is in capacity to find as many as four plausible distractors or up to 50 different items. The consequences of decreasingly effective distractors with the number of options per item are not explored here. Likewise, including more items has the potential to cover more content, a benefit not investigated here.

4.5 Quasi-efficient scoring

Quantitative analyses have shown that efficient penalty $\hat{\theta}$ does not deviate much from notional mark θ^* for $n > 5$ items (see Tables 6 to 11). In the baseline model, the efficient penalty is close to the notional mark (about 0.10 points below for $n = 10$ to 40 and around 0.01 or 0.02 below for $n \geq 80$ (Table 7). This suggests that a simplified scoring rule with a fixed penalty could provide a satisfactory estimation of

[9] See Budescu and Nevo (1985) for a discussion.

examinees' ability. To check this possibility, scoring rules with $\theta = \theta^*$ are compared to fully efficient scoring rules in the baseline calibration.

Table 5: Root mean squared errors (RMSE) with quasi-efficient scoring

Number of items (n)	1	5	10	20	40	80	200	∞
Efficient scoring	0.386	0.221	0.166	0.122	0.089	0.066	0.046	0.000
Quasi-efficient scoring	0.383	0.221	0.166	0.123	0.091	0.069	0.046	0.000

Model: 3 options per item and moderate loss aversion ($\lambda = 1.5$). The notional mark corrects for pure guessing ($\theta^* = -0.50$) in the efficient scoring model. See Table 7 for detailed statistics. Quasi-efficient scoring: the penalty is fixed ($\theta = -0.50$). See Table 13 for detailed statistics.

The two scoring rules produce similar result. The penalty θ is moderately higher than the efficient penalty. The mark for omission is also slightly higher so that the incentives to omit are globally preserved. The differential marks $\gamma - \theta$ are similar, and so are the proportion of omitters. Overall, the RMSE are very close. The simplified scoring rule is a reasonable approximation of the fully efficient rule.

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 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 the 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 when feasible a large number of items to exploit the law of large numbers.

Additional items proves 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. Raising the number of options per item is another way to improve estimation, especially from 2 options (true/false type items), to 3 options. Proposing more than 3 options reduces measurement errors to a lesser extent. Moreover, the literature points to the difficulty of writing more than two plausible distractors (Rodriguez, 2005).

Second, the targeted proportion of omitters should vary with test length. If the number of items is large, ability is generally better estimated by selection than omission. Selection should be forced by setting a negative mark for omission. If the number of items is limited, omission should 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 answer.

Third, the behavior of low able examinees is better targeted by the mark for omission than by the penalty for wrong answers. A fixed penalty is a satisfactory and easy-to-implement second best rule. 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 some 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). Last, the tests could be more realistically modeled by considering items with varying difficulty. 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.

References

Akyol S. P., Key J. and K. Krishna (2016) "Hit or miss? Test taking behavior in multiple choice exams", NBER Working Paper 22401.

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 B. Nevo (1985) "Optimal number of options: An investigation of the assumption of proportionality", *Journal of Educational Measurement*, 22, 183-196.

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 marking", *Assessment and Evaluation in Higher Education*, 29, 585-595.

Burton R.F. and D.J. Miller (1999) "Statistical modelling of multiple-choice and true/false tests: ways of considering, and of reducing, the uncertainties attributable to guessing", *Assessment and Evaluation in Higher Education*, 24 (4), 399-411.

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 guessing)", *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 (2018) "Brave boys and play-it-safe girls: gender differences in willingness to guess in a large scale natural field experiment", working paper.

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.

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.

Rodriguez M.C. (2005) "Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research", *Educational Measurement: Issues and Practice*, 24, 3-13.

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 6: 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 ($\theta^* = -0.50$), 3 options per item. $\hat{\theta}$: 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 7: 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 ($\theta^* = -0.50$), 3 options per item. $\hat{\theta}$: 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 8: 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 ($\theta^* = -0.50$), 3 options per item. $\hat{\theta}$: 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).

Number of options

Table 9: Scoring properties with 2 options per item

Number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\theta}$	-3.2	-0.96	-0.91	-0.90	-0.90	-0.90	-0.98	-1
$\hat{\gamma}$	0.64	0.35	0.26	0.19	0.15	0.11	-0.24	-0.25
$\hat{\gamma} - \hat{\theta}$	3.90	1.31	1.18	1.10	1.04	1.00	0.74	0.75
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	19.7	11.9	7.35	3.70	0.98	-0.87	0.00	0.00
Omitters (%)	87.6	47.0	37.8	31.5	27.3	24.5	0.00	0.000
RMSE	0.407	0.244	0.186	0.140	0.106	0.080	0.057	0.00

Model: notional mark corrects for pure guessing ($\theta^* = -1$), moderate loss aversion ($\lambda = 1.5$). $\hat{\theta}$: efficient mark for incorrect 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 square error. For $n \geq 200$, $\hat{\gamma}$ is the highest mark inducing all examinees to answer (any lower value would also be efficient).

Scoring properties with 3 options per item: See Table 7.

Table 10: Scoring properties with 4 options per item

Number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\theta}$	-1.38	-0.31	-0.30	-0.30	-0.30	-0.31	-0.33	-0.33
$\hat{\gamma}$	0.58	0.28	0.20	0.14	0.10	0.07	-0.12	-0.12
$\hat{\gamma} - \hat{\theta}$	1.96	0.59	0.50	0.44	0.40	0.38	0.21	0.21
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	16.9	10.7	6.58	3.42	1.17	-0.36	0.00	0.00
Omitters (%)	81.6	34.8	26.4	21.1	17.7	15.3	0.00	0.00
RMSE	0.371	0.211	0.156	0.114	0.083	0.061	0.041	0.000

Model: notional mark corrects for pure guessing ($\theta^* = -0.33$), moderate loss aversion ($\lambda = 1.5$). $\hat{\theta}$: efficient mark for incorrect 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 square error. For $n \geq 200$, $\hat{\gamma}$ is the highest mark inducing all examinees to answer (any lower value would also be efficient).

Table 11: Scoring properties with 5 options per item

Number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\theta}$	-1.18	-0.23	-0.22	-0.23	-0.23	-0.23	-0.23	-0.25
$\hat{\gamma}$	0.57	0.26	0.18	0.13	0.09	0.06	0.04	-0.1
$\hat{\gamma} - \hat{\theta}$	1.75	0.49	0.41	0.35	0.32	0.38	0.28	0.15
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	16.5	10.7	6.63	3.52	1.35	-0.11	-1.30	0.00
Omitters (%)	80.5	31.4	23.4	18.5	15.3	13.2	11.42	0.00
RMSE	0.365	0.205	0.151	0.110	0.080	0.058	0.038	0.000

Model: notional mark corrects for pure guessing ($\theta^* = -0.25$), moderate loss aversion ($\lambda = 1.5$). $\hat{\theta}$: efficient mark for incorrect selection. $\hat{\gamma}$: efficient mark for omission. $\hat{\gamma} - \hat{\theta}$ measures the incentives to omit. Omitters (%): share of examinees who omit. $100\left(\hat{\gamma} - \bar{s}(p)\right)$ is omission bias with $\bar{s}(p)$ average omitters' ability. RMSE: root mean square deviation. For $n > 200$, $\hat{\gamma}$ is the highest mark inducing all examinees to answer (any lower value would also be efficient).

Table 12: Efficiency and number of options for a test with a total of 100 options

Number of options per item (m)	2	3	4	5
Number of items (n)	50	33	25	20
$\hat{\theta}$	-0.90	-0.45	-0.30	-0.23
$\hat{\gamma}$	0.13	0.13	0.13	0.13
$\hat{\gamma} - \hat{\theta}$	1.03	0.58	0.43	0.36
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	0.31	1.54	2.60	3.52
Omitters (%)	26.1	22.0	19.9	18.5
RMSE	0.096	0.098	0.103	0.110

Model: moderate loss aversion ($\lambda = 1.5$). All tests have approximately a total of 100 options. $\hat{\gamma} - \hat{\theta}$ measures 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 square error.

Quasi-efficient scoring

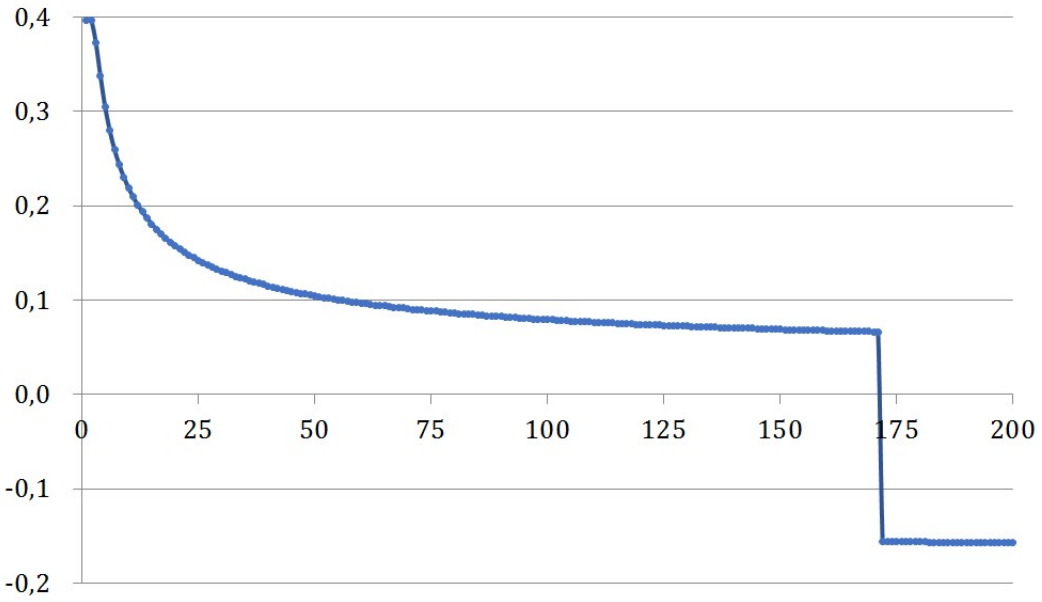
Table 13: Scoring properties with a fixed penalty

Number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\gamma}$	0.59	0.31	0.22	0.16	0.12	0.09	-0.17	-0.17
$\hat{\gamma} - \theta$	1.09	0.81	0.72	0.66	0.62	0.59	0.33	0.33
$100\left(\hat{\gamma} - \bar{s}(p)\right)$	26.5	10.3	5.56	2.09	-0.31	-1.90	0.00	0.00
Omitters (%)	64.8	40.6	33.3	28.1	24.5	22.2	0.00	0.00
RMSE	0.383	0.221	0.166	0.123	0.091	0.069	0.046	0.000

Model: moderate loss aversion ($\lambda = 1.5$), 3 options per item. Quasi-efficient scoring: the penalty corrects for pure guessing ($\theta = \theta^* = -0.50$), but is not adjusted for finite sample; $\hat{\gamma}$: 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).

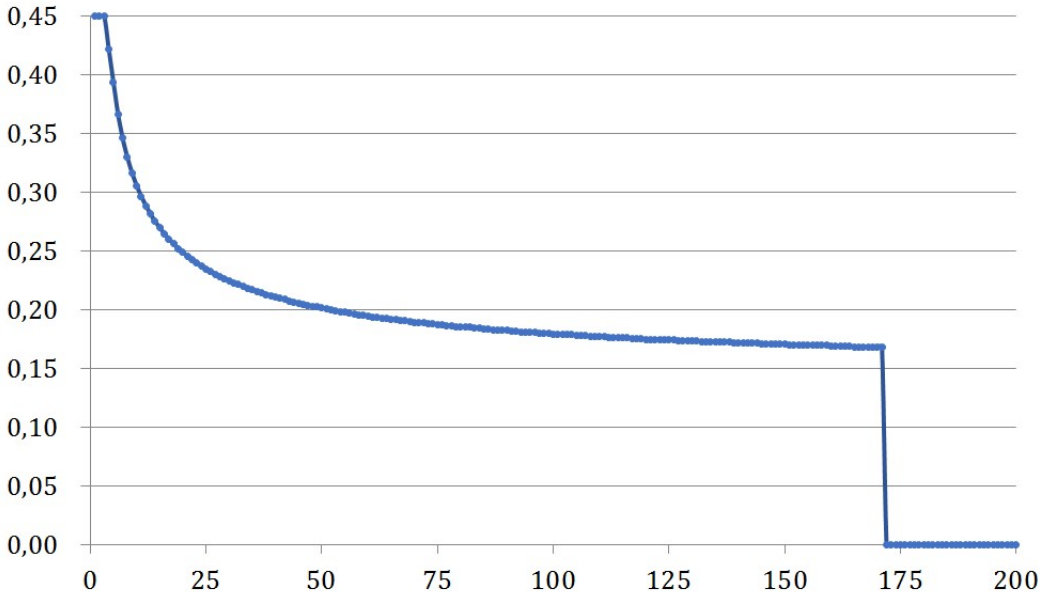
Appendix B Figures

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



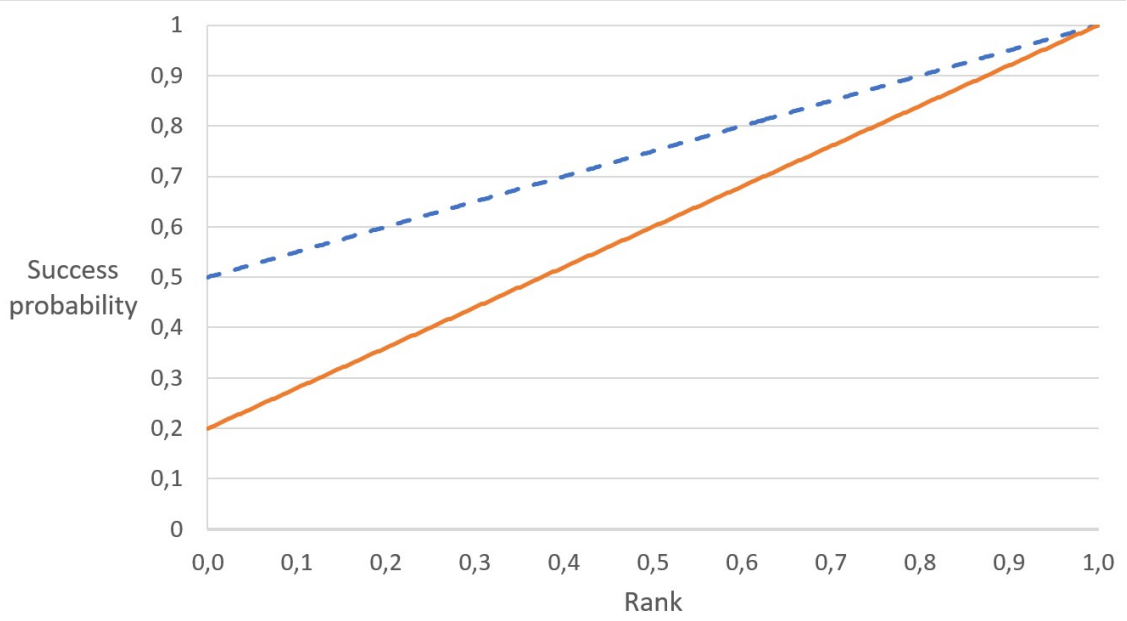
Model: 3 options per item, notional mark corrects for pure guessing ($\theta^* = -0.50$), moderate loss aversion ($\lambda = 1.5$).

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



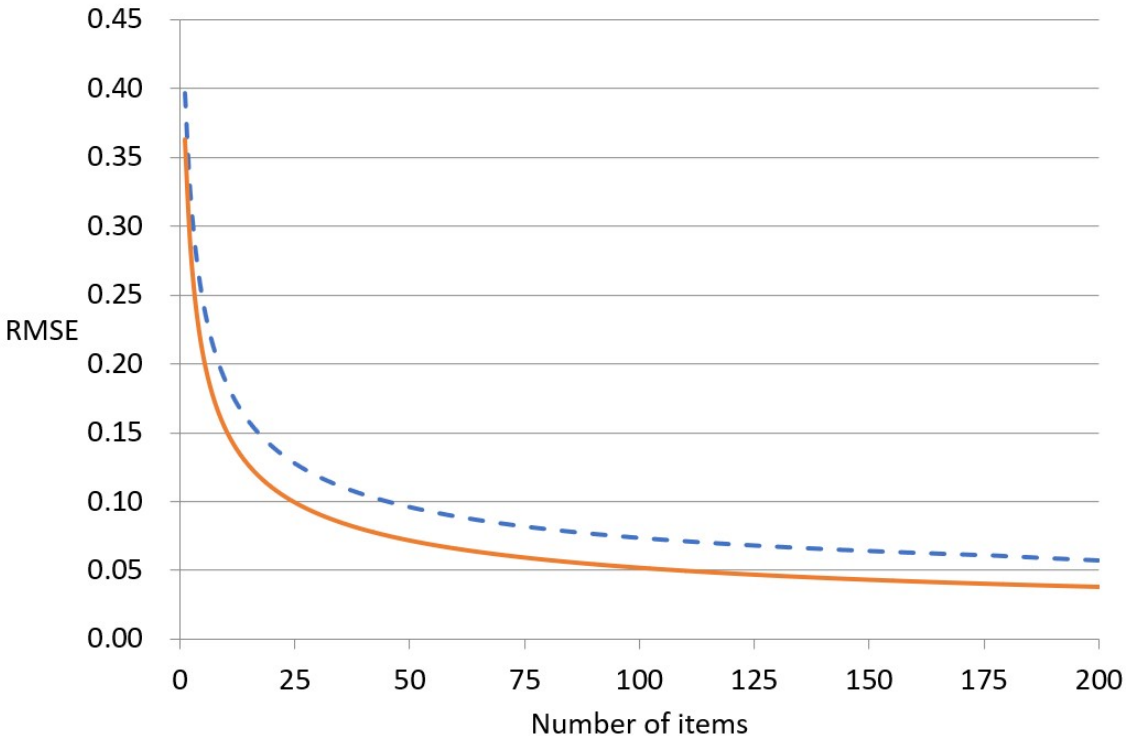
Model: 3 options per item, notional mark corrects for pure guessing ($\theta^* = -0.50$), moderate loss aversion ($\lambda = 1.5$).

Figure 3: Examinees' rank according to their success probability p



Notes. p is the probability of selecting the right option. Solid line: success probabilities for $m = 5$ options. Dotted line: success probabilities for $m = 2$ options. Rank is examinees' relative standings as measured by $F(p)$. Reading: with five options (solid line), an examinee with median ability ($F(p) = 0.5$) has 0.6% chance of selecting the right option.

Figure 4: Root mean squared errors (RMSE) for 2 and 5 options per item



Model: moderate loss aversion ($\lambda = 1.5$). Dotted line: RMSE for $m = 2$ options and $\theta^* = -1$. Solid line: RMSE for $m = 5$ options and $\theta^* = -0.25$.