Efficient Scoring of Multiple-Choice Tests

Alexis Direr

Efficient Scoring of Multiple-Choice Tests

Alexis Direr *

June 16, 2019

Abstract

This paper studies the optimal scoring of multiple choice tests by using standard estimation theory where obtained scores are efficient estimators of examinees' ability. The marks for wrong selections and omissions jointly minimize the mean square difference between obtained score and ability. Examinees are loss averse, ie. disproportionately weight the penalty for wrong selection in their utility function, which entails a preference for omission. With a limited number of items, it is efficient to incentivize the lowest able to omit as their answers essentially reflect noise. The shorter the test, the stronger the incentives to omit. Loss aversion improves estimators efficiency by inducing more omission, which reduces the need to bias the marks to foster omission. The model also sheds new lights on the statistical properties of two widely used scoring methods: number right and formula scoring.

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

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

*Univ. Orléans, CNRS, LEO, UMR 7322, F45067, and Paris School of Economics. E-mail : alexis.direr@univ-orleans.fr.

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 often have 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 the possibility to tell the three apart.

Guessing adds an error component to scores. Suppose that a test-taker has a probability 0.8 of selecting the right option. She may be lucky and gets an average score of 90%, or unlucky and gets a score of 70%. In both cases, her true ability is mismeasured. If the number of items is large enough, the law of large numbers applies and ensures that the measurement error converges to zero. But for practical reasons, most tests have a limited number of items.

The aim of this paper is to design a scoring rule, a mark which penalizes wrong selections, so that the measurement error is as low as possible. The task is complicated by the possibility given to examinees to leave some items blank if they are unsure about the right option. 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. An efficient scoring rule should also include a mark for omission which gives the best estimates of omitters' ability.

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 each examinee, but they also influence examinees in their choice between answering and omitting, which in turn changes the conditions under which abilities are estimated. 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. To study to

what extent those two objectives interact and may possibly conflict, I pose an estimation model in which the marks jointly minimize the mean square difference between examinees' scores and abilities.

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 efficient scoring are investigated.

I find that the efficient scoring rule is fundamentally 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. 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 when they answer, 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 essentially insensitive to the number of items and the scoring strategy.

The model sheds new lights on statistical properties of the two most used scoring methods, number right and formula scoring. Number right scoring (NRS) counts the number of right selections and divides the sum by the total number of items. Omitted items and wrong selections count for zero. Formula scoring (FS) imposes a penalty for incorrect selection equal to $-1/(m-1)$, where m is the number of options in items. The formula equalizes the expected scores of pure guessing and omission (Thurstone, 1919, Holzinger, 1924). I find that the two scoring rules estimate examinees ability with similar degree of accuracy. On the one hand, FS induces more omission by penalizing mistakes,

which reduces the error component in a fully efficient model. On the other hand, omitters ability is poorly estimated with a zero mark for omission. NRS rules out omission, which is only efficient with a large number of items, but avoids any estimation bias problem arising from omissions.

While the two scoring methods produce similar measurement errors, they both underperform compared to an efficient scoring rule. I find in a calibrated model that a test-maker using FS or NRS would have to increase the number of items by on average 30% to obtain the same estimation accuracy than an efficient scoring rule. NRS and FS share two shortcomings. First they do not adjust the marks for finite sample, that is they do not induce more omissions when the number of items is smaller. Second they both set the mark for omission to zero, which induce too much or too few omission, depending of the length of the test.

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 competencies 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). It 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 superiority of one of those rules to the other is still debated. By implementing negative marking and correcting for guessing behavior, 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 students' answering strategies and risk-taking behavior (e.g. Votaw, 1936; Frary, 1988; Budescu and Bar-Hillel, 1993). NRS provides strong incentives to answer all questions, which minimizes the bias.

By assuming that examinees only differ by their knowledge, and not personality traits like risk aversion, the present model does not address this issue. Its general aim is to recast the issue of evaluating ability through multiple choice tests into as standard the framework of estimation theory as possible. By using as a fitness criterion the mean square error, the error term can usefully be decomposed into a variance and a bias components. The model finite and large sample statistical properties can be contrasted. A major finding to this regard is that the efficient scoring rule takes two different forms with a limited number of items and a large set of items.

The model departs from psychometric studies in two other 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 empirical literature (e.g. Akyol et al., 2016). Second, whereas the literature has essentially focused on existing scoring rules, mostly FS and NRS, they do not derive the marks for wrong answers and omission from first principles. They are made endogenous here by making a distinction between a notional mark for wrong answer, essentially a scaling parameter which pins the true score down, and actual marks which minimize measurement errors defined as deviations from true score.

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 answer 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 may vary, whereas both the marks for wrong answers and omission are endogenous in the present model.

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 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 of equal difficulty, so that examinees have a constant probability p of answering correctly any of them. The probability varies across examinees and is a measure 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 in all item divided by the number n of items. Let us consider an examinee who never omits. The number of right selections is the random variable $\tilde{x}$ which follows a binomial distribution $B(n, p)$ with p examinee's probability of a correct selection. Examinees' score is the sum $\tilde{x} \in [0, n]$ of right answers, plus the sum of wrong answers $n - \tilde{x}$ weighted by the penalty θ, divided by the number of items:

$$\tilde{s} = \frac{\tilde{x} + (n - \tilde{x})\theta}{n} \tag{1}$$

The score's first two moments, given success rate p, are $E(\tilde{s}; p) = p + (1 - p)\theta$ and $V(\tilde{s}; p) = \frac{1}{n}(1 - \theta)^2 p(1 - p)$.

2.2 True score

True score depends on examinee's ability p. It is the observed score's component uninfluenced by random events (Harvill, 1991) or the score an examinee would get if p were observable:

$$s(p) = p + (1 - p)\theta^* \tag{2}$$

with $\theta^* < 1$ a notional mark which would prevail in absence of measurement errors. The notional mark is free here from normative justification. It is essentially a scaling parameter which does not affect the way examinees are ranked relative to each other. What will matter for estimation efficiency will be how actual marks for wrong answers and omissions relate to the notional mark.

To fix ideas, the notional mark may take a reference value borrowed from one of the two most used scoring rules, number right scoring (NRS) or formula scoring (FS). In NRS, the final mark is the number of right answers, implying $\theta^* = 0$ and $s(p) = p$. If examinees without any knowledge select an option at random, their expected score is positive and equal to the probability of picking the right option among m ones: $E(s(1/m)) = 1/m$.

FS aims at removing in expectation the reward from pure guessing. It imposes the penalty $\theta^* = -1/(m-1)$ whenever an incorrect answer is selected, so that examinee's expected score is zero with pure guessing:

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

If some examinees are misinformed or have false knowledge, they could perform worse than selecting an option at random. The minimal ability p would lie between 0 and $1/m$ in this case. NRS also rewards misinformation, albeit to a lesser extent than pure guessing. Misinformed examinees would obtain a negative mark in expectation under FS. Misinformation is ruled out in the following 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 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. Applied to the context of exam taking, the utility loss associated with being wrong is larger than the utility gain of being right or omitting: $u(1) = 1$, $u(\gamma) = \gamma$ and $u(\theta) = \lambda\theta$, 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, a limit case of risk preferences with $\lambda = 1$. 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 $\{\gamma, \theta\}$, omitting is preferred to responding if its mark is greater than the loss-weighted expected mark of a response:

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

Marginal examinees are test-takers whose ability $\bar{p}$ makes them indifferent between selecting and omitting:

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

Compared to the case of risk neutrality, loss aversion raises the threshold probability $\bar{p}$:

$$\bar{p} = \frac{\gamma - \lambda\theta}{1 - \lambda\theta} > \frac{\gamma - \theta}{1 - \theta} \tag{3}$$

Examinees omit when they are not confident enough in their selection: $p \leq \bar{p}$, which depends positively on the mark γ and negatively on penalty θ.

[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).

2.4 Mean squared error

Examinees' true score is either estimated thanks to respondents' success rate, or by assigning a constant mark to omissions, which exploits the fact that omitting reveals a low ability on average. Both methods produce error measurements.

Consider first an examinee whose ability is $p > \bar{p}$. Because her ability does not vary across items and all items have the same difficulty, she answers all of them and gets the score $\tilde{s}$ defined in (1). The score 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 the marks θ and γ. The mean squared error (MSE) of observed score $\tilde{s}$ taken by examinee with ability p is the average squared difference between $\tilde{s}$ and true score $s(p)$:

$$\mathrm{mse}(\theta; p) = E\Big(\big(\tilde{s} - s(p)\big)^2\Big) \tag{4}$$

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

$$E\big((\tilde{s} - s(p))^2\big) = V(\tilde{s}; p) + \big(E(\tilde{s}; p) - s(p)\big)^2 \tag{5}$$

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 this way both for sample fluctuations and estimator's accuracy. MSE of an unbiased score ($E(\tilde{s}) = s(p)$) is equal to score's variance.

Consider now a test-taker whose ability is $p \leq \bar{p}$. Because her ability 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. Hence examinee's quadratic error is the squared deviation of γ from true score $s(p)$, or squared bias:

$$sb(\gamma; p) = \big(s(p) - \gamma\big)^2 \tag{6}$$

While individual abilities are not observed by the test-maker, their distribution is assumed to be known. Let $f(p)$ denote the ability probability density function. The test-

maker chooses the marks θ and γ so as to minimize the MSE averaged over examinees:

$$\begin{aligned}\min_{\gamma,\theta} \text{MSE}(\gamma,\theta) &= \int_{p_0}^{\bar{p}} \text{sb}(\gamma;p)f(p)dp + \int_{\bar{p}}^{1} \text{mse}(\theta;p)f(p)dp \\ &= \int_{p_0}^{\bar{p}} \big(s(p)-\gamma\big)^2 f(p)dp + \int_{\bar{p}}^{1} E\Big(\big(\tilde{s}-s(p)\big)^2\Big)f(p)dp \end{aligned} \tag{7}$$

Like the MSE component from answers, the MSE component from omissions, normalized by their proportion $F(\bar{p})$ in the population, lends itself to a decomposition:

$$\begin{aligned}\frac{1}{F(\bar{p})}\int_{p_0}^{\bar{p}} \big(s(p)-\gamma\big)^2 f(p)dp &= E_{|\text{omit}}\big((s(p)-\gamma)^2\big) \\ &= V_{|\text{omit}}\big(s(p)\big) + \big(\bar{s}(p)-\gamma\big)^2 \end{aligned} \tag{8}$$

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

$$\bar{s}(p) = E_{|\text{omit}}\big(s(p)\big) = \frac{1}{F(\bar{p})}\int_{p_0}^{\bar{p}} s(p)f(p)dp \tag{9}$$

and $V_{|\text{omit}}\big(s(p)\big) = E_{|\text{omit}}\big(\big(s(p)-\bar{s}(p)\big)^2\big)$ is the conditional variance of omitters' ability.

Total omitters' measurement error has two components. The variance term classically 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.

Given a proportion $F(\bar{p})$ of omitters, the MSE is minimized for $\hat{\gamma} = \bar{s}(p)$. When the proportion $F(\bar{p})$ is endogenous and responds to variations of the marks, we will see that it may be efficient to bias $\hat{\gamma}$ to induce more or less omission.

It also follows that, as long as some examinees omit, the variance term in (8) is a lower bound whatever the number of items that compose the test. This is a major difference with the MSE component from answers where the average error can be brought to zero with n large enough.

3 Efficient scoring

3.1 Optimality conditions

Assume that a non-empty group of examinees with ability $p \in [p_0, \bar{p}]$ omit. First order conditions of the minimization program (7) are:

$$\frac{\partial \text{MSE}}{\partial \gamma}(\gamma, \theta) = \big(sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p})\big) \frac{d\bar{p}}{d\gamma} f(\bar{p}) + \int_{p_0}^{\bar{p}} \frac{\partial sb}{\partial \gamma}(\hat{\gamma}; p) f(p) dp = 0$$
$$\frac{\partial \text{MSE}}{\partial \theta}(\gamma, \theta) = \big(sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p})\big) \frac{d\bar{p}}{d\theta} f(\bar{p}) + \int_{\bar{p}}^{1} \frac{\partial mse}{\partial \theta}(\hat{\theta}; p) f(p) dp = 0$$

or

$$\frac{\partial \text{MSE}}{\partial \gamma} = \big(sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p})\big) \frac{1}{1 - \lambda\hat{\theta}} f(\bar{p}) + 2 \int_{p_0}^{\bar{p}} \big(\hat{\gamma} - s(p)\big) f(p) dp = 0 \tag{10}$$
$$\frac{\partial \text{MSE}}{\partial \theta} = -\big(sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p})\big) \frac{(1 - \bar{p})\lambda}{1 - \lambda\hat{\theta}} f(\bar{p})$$
$$+ 2 \int_{\bar{p}}^{1} \Big((1 - p)^2 (\hat{\theta} - \theta^*) - \frac{1}{n} p(1 - p)(1 - \hat{\theta})\Big) f(p) dp = 0 \tag{11}$$

The common term in the two equations

$$sb(\hat{\gamma}; \bar{p}) - mse(\hat{\theta}; \bar{p}) = \big(\hat{\gamma} - s(\bar{p})\big)^2 - E\big[(\tilde{s} - s(\bar{p}))^2\big]$$
$$= \big(\hat{\gamma} - s(\bar{p})\big)^2 - \Big(V(\tilde{s}; \bar{p}) + \big(E(\tilde{s}; \bar{p}) - s(\bar{p})\big)^2\Big)$$
$$= \Big(\hat{\gamma} - \big(\bar{p} + (1 - \bar{p})\theta^*\big)\Big)^2 - \Big(\frac{1}{n}(1 - \hat{\theta})^2 \bar{p}(1 - \bar{p}) + (1 - \bar{p})^2 (\hat{\theta} - \theta^*)^2\Big) \tag{12}$$

is a replacement effect caused by marginal examinees with ability $\bar{p}$ changing their choice from selection to omission. This impacts the MSE by substituting a measurement error from answering by one from omitting. $d\bar{p}/d\gamma$ and $d\bar{p}/d\theta$ are the effects of a variation of γ and θ on threshold probability $\bar{p}$ (see (3)). Raising γ or reducing θ both encourage omission and expand the group of omitters:

$$\frac{d\bar{p}}{d\gamma} = \frac{1}{1 - \lambda\hat{\theta}} > 0$$
$$-\frac{d\bar{p}}{d\theta} = \frac{(1 - \bar{p})\lambda}{1 - \lambda\hat{\theta}} > 0$$

Two model's particular cases are of interest. In the first case, examinees' ability is estimated without omission (Subsection 3.2). In the second case, omission is allowed, but the number of items included in the test is arbitrarily large (Subsection 3.3).

3.2 No omission

As a preliminary analysis, let us assume that the mark γ is set so that no examinees find preferable to omit, even the least knowledgeable ones whose ability is p_0: $\gamma \leq p_0 + (1-p_0)\lambda\theta$. A possible scoring rule which satisfies this property has the same mark for omission and wrong selection: $\gamma = \theta$. An example is NRS where the two marks equal zero and for which it is never optimal to omit. We are left with one endogenous parameter, the mark θ, which minimizes the MSE:

$$\begin{aligned}\min_{\theta} \text{MSE}(\theta) &= \int_{p_0}^{1} \text{mse}(\theta; p) f(p) dp = \int_{p_0}^{1} V(\tilde{s}; p) + \left[E(\tilde{s}; p) - s(p)\right]^2 f(p) dp \\ &= \int_{p_0}^{1} \left(\frac{1}{n}(1-\theta)^2 p(1-p) + (1-p)^2(\theta - \theta^*)^2\right) f(p) dp\end{aligned}$$

The variance term $V(\tilde{s}; p)$ is minimized by $\theta = 1$ and the squared bias by $\theta = \theta^*$. Hence, the efficient mark $\hat{\theta}$ lies somewhere between those two values. After some calculations, $\hat{\theta}$ satisfies:

$$\frac{\hat{\theta} - \theta^*}{1 - \hat{\theta}} = \frac{1}{n} \frac{\int_{p_0}^{1} p(1-p) f(p) dp}{\int_{p_0}^{1} (1-p)^2 f(p) dp} \tag{13}$$

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

Proof (i): the right hand term of (13) is positive. The case $\hat{\theta} > 1$ is ruled out by $\theta^* < 1$. (ii) and (ii) are straightforward from Condition (13). □

A reduced penalty (a higher θ) lowers the score's variance, which is traded off against accuracy. The resulting score lessens the penalty, compared to a scoring with the notional mark: $\hat{\theta} > \hat{\theta^*}$. Contrary to the variance, the bias is independent of n. Hence when n increases, its relative weight in the MSE also increases, which makes the bias more costly. In other words, 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.

3.3 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 in a borderline case.

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

Proof The MSE minization program is:

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

The variance term asymptotically tends to zero with n:

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

The MSE is minimized for $\hat{\theta} = \theta^*$ 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 condition $\hat{\gamma} = (p_0 + (1-p_0)\lambda\theta^*$ allows the least knowledgeable to omit since their MSE is also zero in this case. □

When the number of items is arbitrarily large, respondents' abilities are accurately estimated. To the contrary, omitters create measurement errors which do not vanish with test length, since omission signals low ability only on average.

Under risk neutrality ($\lambda = 1$), the unbiasedness condition for the least able coincides with the incentives given to them to omit. If $\hat{\gamma} = p_0 + (1-p_0)\theta^*$, they are indifferent between answering and omitting. If they omit, they get the unbiased mark $\hat{\gamma} = p_0 + (1-p_0)\theta^* = s(p_0)$. If they answer, they obtain the same score $p_0 + (1-p_0)\theta^*$. It results that efficient marks may indifferently induce the least able to answer or to omit.

3.4 Finite sample properties

When the number of items is finite, examinees' ability is estimated with errors due to finite-sample fluctuations. How does it affect efficient marks θ and γ? A major determinant is the sign of the replacement effect presented in Subsection 3.1. A negative replacement effect means that the MSE is reduced when marginal examinees switch from selection to omission. Lemma 1 indicates under which condition the replacement effect is negative.

Lemma 1 *$sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p}) < 0$ if (i) $\lambda = 1$, or (ii) $\hat{\theta} > \theta^*$ and $0 < -\hat{\theta}(\lambda - 1) \leq 2(\hat{\theta} - \theta^*)$.*

Proof $sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p}) < 0$ if $\big(\hat{\gamma} - s(\bar{p})\big)^2 < \big(E(\tilde{s};\bar{p}) - s(\bar{p})\big)^2 + V(\tilde{s};\bar{p})$ (see (12)). If $\lambda = 1$, $\hat{\gamma} = E(\tilde{s};\bar{p})$, the two biases cancel off exactly, $sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p}) < 0$. If λ is close enough to 1: $0 < \theta^* < \hat{\theta}$ implies $0 < \bar{p}+(1-\bar{p})\theta^* < \bar{p}+(1-\bar{p})\lambda\hat{\theta} < \bar{p}+(1-\bar{p})\hat{\theta}$ and therefore $0 < \hat{\gamma} - s(\bar{p}) < E(\tilde{s};\bar{p}) - s(\bar{p})$. Higher loss aversion reduces $\hat{\gamma}$ further and the bias sign may reverse: $\hat{\gamma} < s(\bar{p})$. The replacement effect is still negative if $s(\bar{p}) - \hat{\gamma} \leq E(\tilde{s};\bar{p}) - s(\bar{p})$, or, after some calculations, if $-\hat{\theta}(\lambda - 1) \leq 2(\hat{\theta} - \theta^*)$. □

Omission by marginal examinees entails an estimation bias as the corresponding mark $\hat{\gamma} = \bar{p} + (1 - \bar{p})\lambda\hat{\theta}$ typically differs from true score $s(\bar{p}) = \bar{p} + (1 - \bar{p})\theta^*$. But if examinees are risk neutral or if they are moderately loss averse (with $-\hat{\theta}(\lambda - 1)$ a measure of loss aversion), and the penalty is above the notional mark, the bias from omitting is lower than the measurement error from answering. It can be proved in this case that omission by the less able examinees is efficient:

Proposition 3 *$\hat{\gamma} > p_0 + (1 - p_0)\lambda\hat{\theta}$ if $-\hat{\theta}(\lambda - 1) < 2(\hat{\theta} - \theta^*)$.*

Proof If $\hat{\gamma} = p_0 + (1 - p_0)\lambda\hat{\theta}$, all examinees answer, except possibly the least able whose ability is p_0. Without omission, the efficient penalty, denoted $\hat{\theta}_A$ satisfies $\hat{\theta}_A > \theta^*$ (Condition 13). Given $\hat{\theta}_A$, omission by the least able is efficient if $-\hat{\theta}_A(\lambda - 1) < 2(\hat{\theta}_A - \theta^*)$ (Lemma 1). For $\hat{\gamma} = \bar{p} + (1 - \bar{p})\lambda\hat{\theta}_A > p_0 + (1 - p_0)\lambda\hat{\theta}_A$, the replacement effect remains

negative under the same condition. First order condition (10), with $\hat{\gamma}_A$ the efficient mark given $\hat{\theta}_A$, writes:

$$\big(sb(\hat{\gamma}_A;\bar{p}) - mse(\hat{\theta}_A;\bar{p})\big)\frac{\hat{\gamma}_A}{1-\lambda\hat{\theta}_A}f(\bar{p}) + 2\int_{p_0}^{\bar{p}}\big(\hat{\gamma}_A - s(p)\big)f(p)dp = 0$$

implying $\hat{\gamma}_A > \bar{s}(p)$ (see (9)) if the replacement effect is negative. Let $\hat{p}$ denote average omitters' ability. $\hat{\gamma}_A > \hat{p}+(1-\hat{p})\hat{\theta}_A \geq \hat{p}+(1-\hat{p})\lambda\hat{\theta}_A > p_0+(1-p_0)\lambda\hat{\theta}_A$. Hence, omission by the less able is efficient, given penalty $\hat{\theta}_A$. Now let $\{\hat{\gamma},\hat{\theta}\}$ be efficient marks, solution of optimality conditions (10) and (11). Suppose *ad absurdum* that $\hat{\gamma} \leq p_0+(1-p_0)\lambda\hat{\theta}$, hence $\hat{\theta} = \hat{\theta}_A$, but we have just proved that $\text{MSE}(\hat{\gamma}_A,\hat{\theta}_A) < \text{MSE}(\gamma,\hat{\theta}_A)\ \forall\gamma \leq p_0+(1-p_0)\lambda\hat{\theta}$ if $-\hat{\theta}(\lambda-1) < 2(\hat{\theta}-\theta^*)$, hence $\{\hat{\gamma},\hat{\theta}\}$ cannot be efficient. □

It is efficient that the less knowledgeable omit if they are not too loss averse. Since those examinees select options with no or little knowledge, their score essentially reflects noise. It is therefore efficient to induce them to omit and thereby reveal their low ability. The superiority of omission for the less able breaks if loss aversion exceeds a certain level.

Fig. 1 (in Appendix II) explains why. If examinees are risk neutral (diagram (a)), estimated ability of marginal omitters and respondents are equally biased: $\hat{\gamma} - s(\bar{p}) = E(\tilde{s};\bar{p}) - s(\bar{p})$. The mark $\hat{\gamma}$ is biased upward to induce examinees to omit, but so is respondents' score $\tilde{s}$. If examinees are loss averse (diagram (b)), the mark $\hat{\gamma}$ is moving to the left. The resulting omission bias is lower than the answer bias. Compared to loss neutrality, more examinees spontaneously omit, which limits the need for rewarding omission and distorting γ. Hence moderate loss aversion improves efficiency compared to loss neutrality. If examinees are "excessively" loss averse (diagram (c)), the mark $\hat{\gamma}$ is now moving away from true score and the omission bias may become larger than the answer bias. The mark is not intended to foster omission anymore, but to refrain too many examinees to omit. Its value is so low that it becomes a poor estimate of omitters' ability, which makes omission inefficient.

Lemma 1 shows that if examinees are not too loss averse, inducing more examinees to omit improves efficiency. A direct consequence is that the mark $\hat{\gamma}$ is greater than omitters' average ability $\bar{s}(p)$ (defined in (9)).

Proposition 4 *$\hat{\gamma} > \bar{s}(p)$ if $-\hat{\theta}(\lambda - 1) < 2(\hat{\theta} - \theta^*)$.*

Proof First order condition (10) is:

$$2\int_{p_0}^{\bar{p}} \big(\hat{\gamma} - s(p)\big) f(p)dp = -\big(sb(\hat{\gamma};\bar{p}) - mse(\hat{\theta};\bar{p})\big)\frac{\hat{\gamma}}{1-\lambda\hat{\theta}} f(\bar{p}) > 0$$

if the replacement effect is negative. It follows:

$$\hat{\gamma} > \frac{1}{F(\bar{p})}\int_{p_0}^{\bar{p}} s(p)f(p)dp = \bar{s}(p)$$

□

It is efficient to bias the mark upward to induce more omission.

4 Simulated properties

4.1 Simulation strategy

This section presents some numerical results from the statistical model of scoring. Regarding risk preferences, Tversky and Kahneman (1992) estimate a loss aversion coefficient $\lambda = 2.25$ in cumulative prospect theory. It is however not entirely clear how a parameter estimated from choices involving monetary outcomes translates to the context of grades. I assume three conservative and plausible levels of loss aversion: loss neutrality ($\lambda = 1$), moderate loss aversion ($\lambda = 1.5$) and stronger loss aversion ($\lambda = 2.5$). In the cases where a mistake is positively marked ($\theta > 0$), it is assumed to be still edited as a loss. The mark is reduced by the coefficient $1/\lambda$ in this case.

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 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 informations, I choose a simple uniform distribution over the space of ability $[p_0, 1]$.

The MSE (7) 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):

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

It is the geometric mean of measurement errors for all examinees. A RMSE of 0.10 for instance means that obtained scores deviate on average from true scores by this amount, which can be compared to the scales of a full point if a right answer and the notional mark θ^* (classically equal to $-1/(m-1)$ or 0) if a wrong one.

I also compute the bias on omitters' score $\hat{\gamma} - \bar{s}(p)$, which is estimated omitters' ability minus average omitters' ability $\bar{s}(p)$ (see its expression (9)). It informs about to what extent omission is fostered (if positive) or dissuaded (if negative). The bias depends on the incentives to omit, which is measured by the mark differential $\gamma - \theta$.

4.2 Efficient scoring

I first study a baseline model in which the test is composed of various numbers of items (n = 1, 5, 10, 20, 40, 80, 200, ∞). Each item has $m = 3$ options. True score $s(p)$ is computed for notional mark $\theta^* = -1/(m-1)$, which corrects for pure guessing, as in formula scoring. Loss aversion coefficient is set to 1.5.

Table 10 in Appendix I presents the efficient marks and main statistics in function of n for the baseline calibration. Fig. 2 and Fig. 3 display the full profile of mark $\hat{\gamma}$ and the proportion of omitters respectively in function of n.

Two distinct scoring strategies emerge. When the number of items is below a threshold (here less than 170) omission is encouraged to palliate inaccurate estimation of low able examinees ability. The mark for omission is positive and above average omitters' ability.

Except for very limited number of items, it slowly decreases around 0.1 (Fig. 2). The proportion of omitters is also decreasing with n (Fig. 3). When the test has a large number of items, omission is dissuaded altogether. The mark for omission drops to negative values, around -0.15. The mark differential $\hat{\gamma} - \hat{\theta}$ is reduced from around 0.6 to 0.33. The proportion of omitters follows logically the same profile with a sudden fall to zero, the efficient proportion for large sample. All examinees answer, including the lowest able.

Except for the extreme case $n = 1$ where the mark differential is 2.4 and 83.5% of examinees omit, the efficient penalty $\hat{\theta}$ is greater (milder) than notional mark θ^* (Table 10). It lies in the close neighborhood of the notional mark, which suggests that a scoring rule with a fixed penalty equal to the notional mark might prove a good approximation of the efficient rule (more in Subsection 4.6). The penalty for wrong answers varies little with n, compared to the mark for omission. The behavior of low able is indeed better targeted by the mark for omission than by the penalty which impacts all examinees, including the most proficient who will always answer.

Efficient scoring departs from actual scoring rules like number right scoring (NRS) or formula scoring (FS) in two ways. First, the marks are adjusted for finite sample, which seems particularly relevant for tests of small and medium sizes. In FS or NRS, the marks are fixed whatever the test length. Second, a mark for omission set to zero is not efficient. It is either positive and even biased upward to foster omission or conversely negative to dissuade omission (Fig. 2). For n not too large, the efficient value of γ is strictly positive for two reasons. First, insofar as a significant proportion of examinees omit, the mark should reflect omitters' average ability and credit partial knowledge. Second, it exceeds omitters' average ability in order to foster omission further, which has been shown to reduce measurement errors in the analytical section (Prop. 4).

The quantitative importance of adjusting for test length and setting γ above zero can be evaluated by comparing measurement errors of efficient scoring with notional penalty $\theta^* = -1/(m-1)$, and FS where actual penalty is $-1/(m-1)$ and the mark for omission is fixed and equal to zero. Table 1 extracts root mean squared errors (defined in (4.1)) from Tables 10 and 18 with $m = 3$ options.

Table 1: Efficient scoring vs formula scoring with moderate loss aversion and 3 options per item

number of items (n)	1	5	10	20	40	80	200	∞
Efficient scoring RMSE	0.386	0.221	0.166	0.122	0.089	0.066	0.046	0.00
Formula scoring RMSE	0.584	0.263	0.187	0.134	0.097	0.072	0.051	0.031
Compensated nb of items	3	7	13	25	48	99	285	-
Variation rate (%)	200	40	30	25	20	24	42	-

Notes. Examinees are moderately loss averse ($\lambda = 1.5$). Efficient scoring: notional mark corrects for pure guessing ($\theta^* = -0.50$). See Table 10 for detailed statistics. Formula scoring: the penalty corrects for pure guessing ($\theta = -0.50$), the mark for omission is set to zero ($\gamma = 0$), no adjustment is made for finite sample. See Table 18 for detailed statistics. RMSE: root mean squared error. Compensated nb of items: number of items which must be added to the test with FS to achieve the same level of accuracy than the efficient test. Variation rate: rate of increase of the number of supplementary items.

As expected, measurement errors are larger with formula scoring than with efficient scoring. The efficiency loss is significant for tests with a limited number of items due to a lack of omission in formula scoring. The proportion of omitters is constant and equal to 14.3% (Table 18), compared to 25% with efficient scoring and $n = 20$ (Table 10), and 18.6% with $n = 80$. Insufficient omission comes from a too low mark differential $\gamma - \theta$ equal to 0.50, compared to 0.61 for $n = 20$ and 0.57 for $n = 40$ with efficient marks.

The error differences between the two scoring rules are decreasing with n and become negligible in absolute terms for $n > 40$, except for very large n where a bias on omitters ability still remains with FS. Table 1's third and fourth lines show a persistent difference once expressed in additional items FS must include to perform as well as efficient scoring. The rate of increase is between 20% for $n = 40$ and 42% for $n = 200$.

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 from mistakes, which generates a preference for omission. Its consequences for omission are however ambivalent when the scoring rule is efficient. On the one hand, the proportion of omitters increases with loss aversion (Table 2). For $n = 20$, it is 14.4% if examinees are risk neutral ($\lambda = 1$), 24.9% if they are loss averse ($\lambda = 1.5$), and up to 32.8% if they are strongly loss averse ($\lambda = 2.5$). On the other hand, the stronger loss aversion, the smaller the number of items above which omission is dissuaded. Omission is discouraged for $n > 171$ if examinees are moderately loss averse, and as soon as $n > 57$ if they are strongly loss averse.

Table 2: Proportion of omitters and loss aversion

number of items (n)	1	5	10	20	40	80	200	∞
risk neutrality (%)	43.4	26.0	19.6	14.4	10.5	7.6	4.9	0.00
moderate loss aversion	83.5	39.4	30.6	24.9	21.1	18.6	0.00	0.00
strong loss aversion	85.7	46.8	38.0	32.8	29.9	0.00	0.00	0.00

Notes. Baseline model: $m = 3$ options per item, notional mark corrects for pure guessing ($\theta^* = -0.50$). Risk neutrality: $\lambda = 1$; moderate loss aversion: $\lambda = 1.5$; strong loss aversion: $\lambda = 2.5$. See Tables 9, 10 and 11 for detailed statistics. Reading: 26% of risk neutral examinees omit in a test with 5 items.

The reason is explained in Subsection 3.4 and Fig. 1. When examinees are loss averse, the mark which induces the less able to omit is below the unbiased mark. The more loss averse, the larger the discrepancy and the omission bias. The analytical part has also shown that, at least for moderate levels, loss aversion enhances efficiency, as omission by low able examinees is obtained by distorting less the mark for omission (see Prop. 3).

Root mean squared errors (RMSE) are reported in Table 3 for three loss aversion levels. They 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 3: Root mean squared error and loss aversion

number of items (n)	1	5	10	20	40	80	200	∞
risk neutrality	0.406	0.241	0.181	0.133	0.097	0.070	0.045	0.00
moderate loss aversion	0.386	0.221	0.166	0.122	0.089	0.066	0.046	0.00
strong loss aversion	0.324	0.196	0.151	0.119	0.097	0.072	0.046	0.00

Notes. Baseline model: $m = 3$ options per item, notional mark corrects for pure guessing ($\theta^* = -0.50$). Risk neutrality: $\lambda = 1$; moderate loss aversion: $\lambda = 1.5$; strong loss aversion: $\lambda = 2.5$. RMSE: root mean squared error. See Tables 9, 10 and 11 for detailed statistics.

4.4 Number right and formula scoring

The model allows a comparison of the two most used scoring methods, NRS ($\theta = \gamma = 0$) and formula scoring ($\theta = -1/(m-1)$ and $\gamma = 0$), in which the marks are not adjusted for test length. In NRS, omissions earn zero points, whereas a response can never earn less, while affording a positive probability of earning a point. Hence rational examinees should answer all items, whatever their level of loss aversion.[2] Omission is also sub-optimal under FS but only if examinees are risk neutral.

With no omission, the two scoring rules are equivalent. FS is a mere rescaling of NRS which does not affect examinees' relative standings. Certainly, the root mean square deviation (RMSE) with NRS is smaller than the one with FS (compare Tables 16 and 17), but the difference is entirely explained by FS spreading marks over a broader interval (between $-1/(m-1)$ and 1) than NRS (between 0 and 1).

The two scoring rules are not equivalent anymore when examinees are loss averse. Contrary to NRS, FS penalizes wrong answers, which discourages low able examinees to answer. In the numerical baseline, the proportion of omitters is 14.3% for moderate loss aversion (Table 18) and 25% for stronger loss aversion (Table 19).

[2] This is true if mistakes and omissions are treated the same manner by loss averse examinees, either as a loss or as a gain, which is plausible given that the two results receive the same mark. The alternative assumption, not investigated here, that only wrong selections are edited as a loss could explain why some examinees still omit despite the answers being not penalized (Grandy, 1987).

Although the two scoring rules cannot be compared *prima facie*, they differ by the way omission is treated relatively to answering. NRS dissuades omission by setting the mark differential $\gamma - \theta$ to zero, whereas FS gives some incentives to omit by raising the mark for omission above the one for incorrect answer. The effect of fostering omission on efficiency can be isolated by comparing two scoring rules which differ only by the way omission is rewarded compared to answering. To do so, FS with $m = 3$ options, $\theta = -0.50$ and $\gamma = 0$ is compared to a scoring method, called extended NRS, with the same mark for mistakes and omissions ($\theta = \gamma = -0.50$). The two methods having the same mark for mistakes, the scores are spread over comparable intervals.

Table 4: RMSE in formula scoring and extended number right scoring

number of items (n)	1	5	10	20	40	80	200	∞
FS, risk neutrality	0.645	0.289	0.204	0.144	0.102	0.072	0.046	0.00
FS, loss aversion	0.584	0.263	0.187	0.134	0.097	0.072	0.051	0.031
FS, strong loss aversion	0.497	0.243	0.189	0.155	0.135	0.124	0.117	0.111
Extended NRS	0.645	0.289	0.204	0.144	0.102	0.072	0.046	0.00

Notes. RMSE: root mean squared errors. 3 options per item. FS: the penalty corrects for pure guessing ($\theta = -0.50$), the mark for omission is set to zero ($\gamma = 0$). See Tables 17, 18 and 19 for detailed statistics. Extended NRS: penalty θ and mark for omission γ both set to $-1/(m-1) = -0.50$. No adjustment made for finite sample. It is never optimal to omit under extended NRS, whatever the level of loss aversion.

Table 4 does not show any difference between FS with risk neutrality and extended NRS. The two scoring methods dissuade omission and score mistakes the same way. With loss aversion, FS induces the less able to omit. Analytical results suggest that some extent of omission may reduce estimation errors, except here that the mark for omission is not set to its efficient value. The bias on omitters' estimated ability $\hat{\gamma} - \bar{s}(p)$ is -0.07 with moderate loss aversion ($\lambda = 1.5$) and -0.17 with strong loss aversion ($\lambda = 2.5$), whereas it is positive with an efficient scoring. The reverse bias offsets potential efficiency gains from omission and deteriorates RMSE. One may conclude that, once NRS is modified so that scores are spread over the same intervals as FS, FS performs better with a limited

number of items and worse with a large sample of items, a situation where omission should generally be discouraged.

4.5 Efficient scoring and test length

How many items should a test include? How many options per item? Is there a trade-off between the two margins? While the first question has been rarely investigated in the psychometric literature,[3] the optimal number of options per item has been discussed at length (see Rodriguez (2005) for a survey).

Increasing the number of response options generally increases the difficulty of the item (assuming all the alternatives are plausible), which increases the likelihood that a test-taker will select a distractor item. Pure guessing becomes more hazardous. The probability of picking the right option is 50% with two options, down to 20% with five options. At the other extremity, perfectly informed examinees retrieve the right option whatever the number of distractors. This suggests that examinees with partial knowledge are expected to be confused by a higher number of distractors, but to a lesser extent they are more able.

Varying the number of options from m to $m' > m$ changes the success rate 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.

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

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

Fig. 4 plots examinees ability in function of their relative rank for tests with two and

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

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 a probability of 55% of correctly answering with two options, and 28% with five options. At the other extremity, examinees whose rank $F(p)$ is 0.9 have 95% chance of success with two options, and still 92% with five options.

Fig. 5 shows how fast the root mean squared error (RMSE) declines with the number of items for $m = 2, 3, 4$ 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 gains then decelerate rapidly and reach a quasi-plateau. The RMSE eventually converges to zero but very slowly. It is around 0.05 for $n = 200$ and $m = 3$, and still 0.03 for $n = 1000$. 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 are logically decreasing with the number of options for a given number of items. We can see from Fig. 5 and Table 5 that increasing the number of options from 2 to 3 significantly reduces the RMSE, even for large n where it becomes hard to reduce it 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.[4]

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 issue has practical relevance insofar as the total testing time is not extensible and is increasing with the number of options reviewed.[5] To check this point, we compare tests with varying number of items and options, but constant total number of options, equal to 100.

Table 6 shows that the RMSE hardly varies with test composition. It is almost equivalent to administer a test with 50 items and two choices or a test with 20 items and 5 options.[6]

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

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

[6] The result rests on the assumption that the test-maker is in capacity to find as many as four plausible distractors (and incidentally up to 50 different items). The consequences of decreasingly effective

Table 5: Efficiency and number of options per item

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

Notes. Root mean squared errors (RMSE) are extracted from Tables 12 (2 options), 10 (3 options), 13 (4 options) and 14 (5 options). Baseline model: finite sample-adjusted formula scoring (the notional mark corrects for pure guessing: $\theta^* = -1/(m-1)$). Examinees are moderately loss averse: $\lambda = 1.5$.

Table 6: Number of items and number of options, tests 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

Notes. Baseline model: finite sample-adjusted formula scoring (the notional mark corrects for pure guessing: $\theta^* = -1/(m-1)$). Examinees are moderately loss averse: $\lambda = 1.5$. The number of items $\times$ the number of options is kept constant. RMSE: root mean square error. See Table 15 for detailed statistics.

The quasi-equivalence holds for efficient scoring. Table 7 shows similar results with FS, NRS, and extended NRS. RMSE varies weakly with test configuration for all three methods. At a fine level, two options is marginally best for NRS, and three options for FS and extended NRS.

distractors with the number of options per item are not investigated here. Likewise, including more items has the potential to cover more content, a benefit not investigated here.

Table 7: Number of items and number of options, tests with 100 options

number of options per item (m)	2	3	4	5
number of items (n)	50	33	25	20
FS RMSE	0.110	0.106	0.110	0.116
NRS RMSE	0.058	0.075	0.087	0.097
Extended NRS RMSE	0.115	0.112	0.115	0.121

Notes. RMSE: root mean square error. FS: the penalty corrects for pure guessing ($\theta = -0.50$), the mark for omission is set to zero ($\gamma = 0$). NRS: penalty θ and mark for omission γ set to zero. Extended NRS: penalty θ and mark for omission γ both set to $-1/(m-1) = -0.50$. No adjustment is made for finite sample. Examinees are moderately loss averse: $\lambda = 1.5$.

4.6 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 9 to 14). 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 10).

It suggests that a simplified scoring rule with a fixed penalty could provide satisfactory estimation of examinees ability. To check this possibility, scoring rule with $\theta = \theta^*$ and optimized mark for omission is compared to a fully efficient model with baseline calibration.

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

Table 8: RMSE in efficient scoring and 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

Notes. RMSE: root mean squared errors. 3 options per item. Examinees are moderately loss averse ($\lambda = 1.5$). Efficient scoring: notional mark corrects for pure guessing ($\theta^* = -0.50$). Actual mark is adjusted for finite sample. See Table 10 for detailed statistics. Quasi-efficient scoring: the penalty is fixed and corrects for pure guessing ($\theta = -0.50$). See Table 20 for detailed statistics.

5 Conclusion

Four main lessons can be drawn from the scoring model. First, a test-maker should include, if feasible, a large number of items to exploit the law of large numbers. Additional items proved 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, although the literature on this issue points to the difficulty of writing more than two plausible distractors (Rodriguez, 2005).

Second, the proportion of omitters and the mark for omission should vary with test length. If the number of items is large, ability is generally better estimated by answers than omissions. Omission is dissuaded by setting a negative mark. If it 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 may be quite significant in that case.

Third, the omissive behavior of low able examinees is better targeted by the mark for omission than by the penalty for wrong answers. The penalty is marginally lower than the notional mark, ie. dissuades omission rather than encourages it. It converges

gradually to the notional mark when the number of items increases. A fixed penalty is a satisfactory and easy to implement second best rule.

Last, the instructions, if any, given to examinees should be consistent with the scoring strategy. If the number of items is small, examinees should be encouraged to guess. In the contrary case, they should be instructed to answer all questions even if they are not sure that their answers are correct.

The scoring model allows comparison of the two most used scoring methods, formula scoring and number right scoring. Both scoring rules set the mark for omission to zero, which is not efficient. It induces too much or to few omission, depending on the number of items. By allowing omission, formula scoring is marginally better than number right scoring when the number of items is limited. The reverse is generally true for longer tests where ability is better estimated if all examinees answer.

The model has made some simplifying assumptions which implications for estimation efficiency could be interesting to investigate. 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 modeled more realistically by considering items with varying difficulty. Examinees' probability of being right and their incentives to omit would fluctuate from one item to another. 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.

Grandy J. (1987) "Characteristics of examinees who leave questions unanswered on the GRE general test under rights-only scoring", *ETS Research Report Series*, 2, i-71.

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.

Nagore I. 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 Psychol-*

ogy, 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 (1986) "Rational Choice and the Framing of Decisions" *The Journal of Business*, 59 (4): 251-278.

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 I Tables

Risk preferences

Risk neutrality

Table 9: Scoring statistical properties, finite sample-adjusted formula scoring, risk neutrality ($\lambda = 1$), 3 options per item

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
omission bias	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

Notes. Scoring: the notional mark corrects for pure guessing ($\theta^* = -0.50$). $\hat{\theta}$: efficient mark for wrong selections. $\hat{\gamma}$: efficient mark for omission. $\hat{\gamma} - \hat{\theta}$: a measure of the incentives to omit. Omission bias = $100\left(\hat{\gamma} - \bar{s}(p)\right)$: 100 × estimated omitters' ability minus 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.

Moderate loss aversion

Table 10: Scoring statistical properties, finite sample-adjusted formula scoring, moderate loss aversion ($\lambda = 1.5$), 3 options per item

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
omission bias	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

Notes. Scoring: the notional mark corrects for pure guessing ($\theta^* = -0.50$). $\hat{\theta}$: efficient mark for incorrect selection. $\hat{\gamma}$: efficient mark for omission. $\hat{\gamma} - \hat{\theta}$: a measure of the incentives to omit. Omission bias = $100\left(\hat{\gamma} - \bar{s}(p)\right)$: 100 × estimated omitters' ability minus 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.

High loss aversion

Table 11: Scoring statistical properties, finite sample-adjusted formula scoring, high loss aversion ($\lambda = 2.5$), 3 options per item

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
omission bias	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

Notes. Scoring: the notional mark corrects for pure guessing ($\theta^* = -0.50$). $\hat{\theta}$: efficient mark for incorrect selection. $\hat{\gamma} - \hat{\theta}$: a measure of the incentives to omit. Omission bias $= 100\left(\hat{\gamma} - E\bar{s}(p)\right)$: 100 × estimated omitters' ability minus 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

Two options

Table 12: Scoring statistical properties, finite sample-adjusted formula scoring, moderate loss aversion, 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
omission bias	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

Notes. Scoring: the 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. $\hat{\gamma} - \hat{\theta}$ is a measure of the incentives to omit. Omission bias = $100\left(\hat{\gamma} - \bar{s}(p)\right)$: 100 $\times$ estimated omitters' ability minus 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.

Three options

See Table 10.

Four options

Table 13: Scoring statistical properties, finite sample-adjusted formula scoring, moderate loss aversion, 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
omission bias	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

Notes. Scoring: the 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. $\hat{\gamma} - \hat{\theta}$ is a measure of the incentives to omit. Omission bias = $100\left(\hat{\gamma} - \bar{s}(p)\right)$: 100 × estimated omitters' ability minus 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.

Five options

Table 14: Scoring statistical properties, finite sample-adjusted formula scoring, moderate loss aversion, 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
omission bias	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

Notes. Scoring: the 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. $\hat{\gamma} - \hat{\theta}$ is a measure of the incentives to omit. Omitters (%): share of examinees who omit. Omission bias = $100\big(\hat{\gamma} - \bar{s}(p)\big)$: 100 × ability estimator of omitters minus 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.

Tradeoff between number of items and options

Table 15: Efficiency and number of options for a test with a total of 100 options, baseline model

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
omission bias	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

Notes. Baseline model: finite sample-adjusted formula scoring (the notional mark corrects for pure guessing: $\theta^* = -1/(m-1)$). All tests have exactly or approximately a total of 100 options. Examinees are moderately loss averse: $\lambda = 1.5$. $\hat{\gamma} - \hat{\theta}$ is a measure of the incentives to omit. Omission bias = $100\left(\hat{\gamma} - \bar{s}(p)\right)$: 100 × estimated omitters' ability minus average omitters' ability. Omitters (%): share of examinees who omit. RMSE: root mean square error.

Number right scoring

Table 16: Number right scoring, 3 options per item

number of items (n)	1	5	10	20	40	80	200	∞
RMSE	0.430	0.192	0.136	0.96	0.068	0.048	0.030	0.000

Notes. Scoring: penalty θ and mark for omission γ set to zero, no adjustment made for finite sample. It is never optimal to omit under NR scoring, whatever loss aversion level. The proportion of omitters and omission bias are both zero as a result. RMSE: root mean square deviation.

Formula scoring

Risk neutrality

Table 17: Formula scoring, risk neutrality, 3 options per item

number of items (n)	1	5	10	20	40	80	200	∞
RMSE	0.645	0.289	0.204	0.144	0.102	0.072	0.045	0.000

Notes. Scoring: the penalty corrects for pure guessing ($\theta = -0.50$), the mark for omission is set to zero ($\gamma = 0$), no adjustment is made for finite sample. It is not optimal to omit under formula scoring, when examinees are risk neutral ($\lambda = 1$). The proportion of omitters and omission bias are both zero as a result. RMSE: root mean square deviation.

Moderate loss aversion

Table 18: Formula scoring, moderate loss aversion, 3 options per item

number of items (n)	1	5	10	20	40	80	200	∞
omitters (%)	14.3	14.3	14.3	14.3	14.3	14.3	14.3	14.3
omission bias	−7.14	−7.14	−7.14	−7.14	−7.14	−7.14	−7.14	−7.14
RMSE	0.584	0.263	0.187	0.134	0.097	0.072	0.051	0.031

Notes. Scoring: the penalty corrects for pure guessing ($\theta = -0.50$), the mark for omission is set to zero ($\gamma = 0$), no adjustment is made for finite sample. Examinees are moderately loss averse ($\lambda = 1.5$). Omitters (%): share of examinees who omit. Omission bias $= 100\big(\hat{\gamma} - \bar{s}(p)\big)$: 100 × ability estimator of omitters minus average omitters' ability. RMSE: root mean square deviation.

Strong loss aversion

Table 19: Formula scoring, high loss aversion, 3 options per item

number of items (n)	1	5	10	20	40	80	200	∞
omitters (%)	33.3	33.3	33.3	33.3	33.3	33.3	33.3	33.3
omission bias	−16.7	−16.7	−16.7	−16.7	−16.7	−16.7	−16.7	−16.7
RMSE	0.497	0.243	0.189	0.155	0.135	0.124	0.117	0.111

Notes. Scoring: the penalty corrects for pure guessing ($\theta = -0.50$), the mark for omission is set to zero ($\gamma = 0$), no adjustment is made for finite sample. Examinees are highly loss averse ($\lambda = 2.5$). Omitters (%): share of examinees who omit. Omission bias $= 100\big(\hat{\gamma} - \bar{s}(p)\big)$: 100 × ability estimator of omitters minus average omitters' ability. RMSE: root mean square deviation.

Quasi-efficient scoring

Table 20: Scoring statistical properties, fixed penalty ($\theta = \theta^*$), finite sample-adjusted marking of omissions, moderate loss aversion, 3 options per item

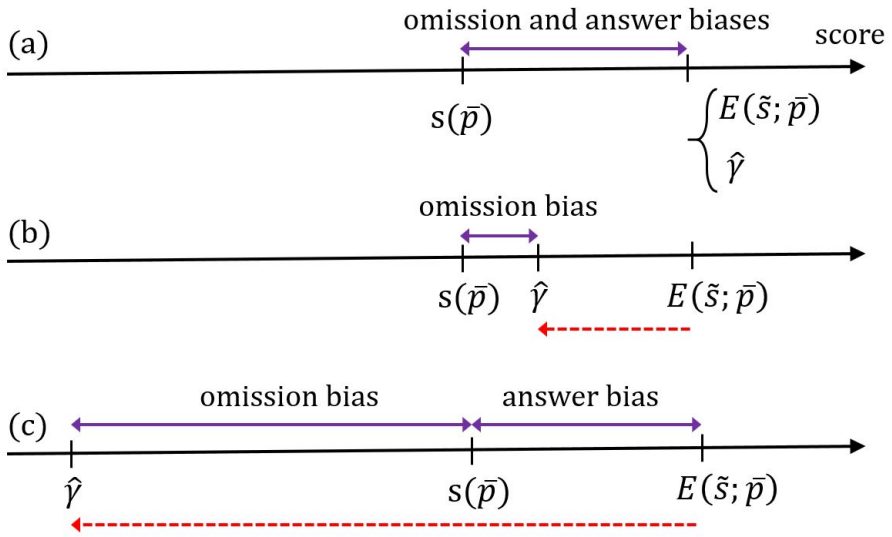
number of items (n)	1	5	10	20	40	80	200	∞
$\hat{\gamma}$	0.59	0.31	0.22	0.16	0.12	0.09	−0.17	−0.17
$\hat{\gamma} - \theta$	1.09	0.81	0.72	0.66	0.62	0.59	0.33	0.33
omission bias	26.5	10.3	5.56	2.09	−0.31	−1.90	0.00	0.00
omitters (%)	64.8	40.6	33.3	28.1	24.5	22.2	0.00	0.00
RMSE	0.383	0.221	0.166	0.123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

Notes. Quasi-efficient scoring: the penalty corrects for pure guessing ($\theta = -0.50$), but is not adjusted for finite sample. $\hat{\gamma}$: efficient mark for omission. $\hat{\gamma} - \hat{\theta}$: a measure of the incentives to omit. Omission bias = $100\left(\hat{\gamma} - \bar{s}(p)\right)$: 100 × estimated omitters' ability minus 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 II Figures

Omission and answer biases

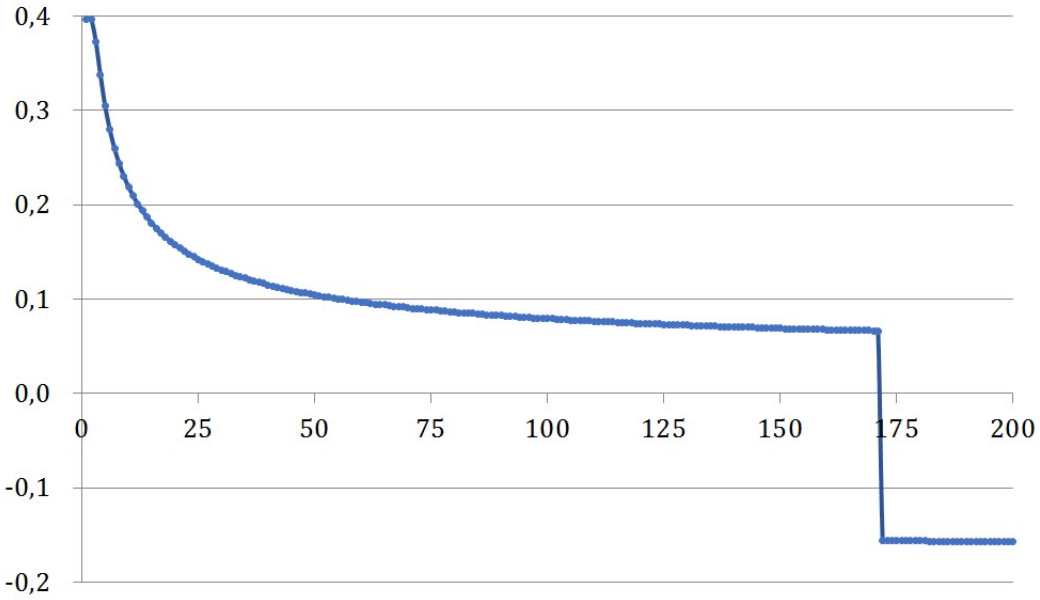
Figure 1: Omission and answer biases for different loss aversion levels



Notes. $\bar{p}$: ability of marginal examinees indifferent between answering and omitting, $s(\bar{p})$: true score, $E(\tilde{s};\bar{p})$: expected actual score, $\hat{\gamma}$: efficient mark for omission, $E(\tilde{s};\bar{p}) - s(\bar{p})$: answer bias, $\hat{\gamma} - s(\bar{p})$: omission bias. (a) Omission and answer biases are equal when examinees are risk neutral. (b) If examinees are loss averse, the mark $\hat{\gamma}$, which induces marginal examinees to omit, is drifting to the left (dotted arrow). The resulting omission bias is lower than the answer bias. (c) If examinees are "excessively" loss averse, the mark $\hat{\gamma}$ is moving away from true score. The omission bias may become larger than the answer bias.

Efficient mark for omission and test length

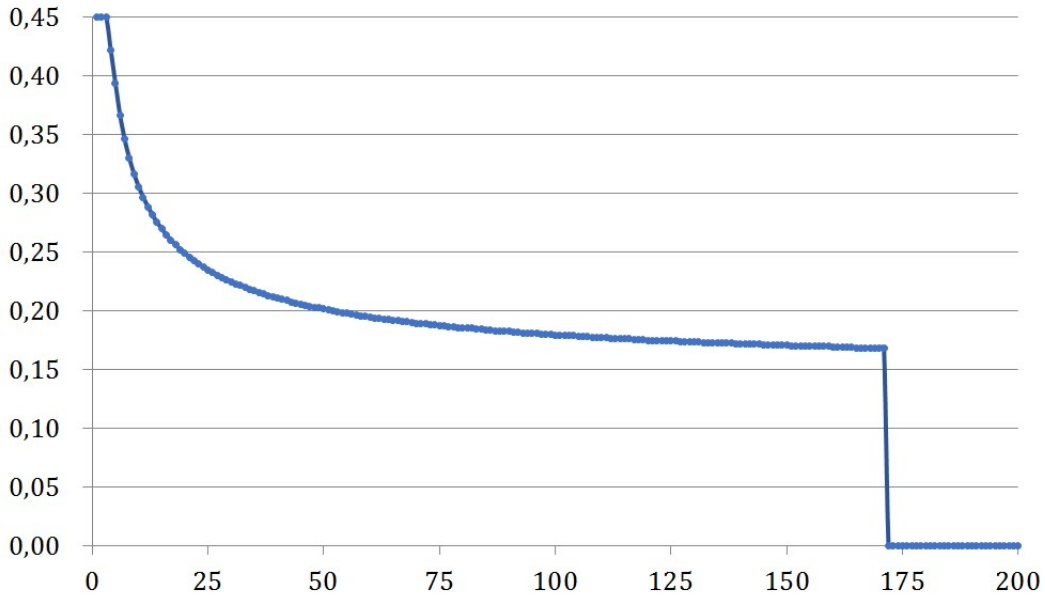
Figure 2: Efficient mark for omission in function of number of items



Notes. Efficient mark for omission ($\hat{\gamma}$) in function of number of items n (horizontal line). Baseline calibration: 3 options per item, the notional mark corrects for pure guessing ($\theta^* = -0.50$); examinees are moderately loss averse ($\lambda = 1.5$), uniform ability distribution.

Efficient omission and test length

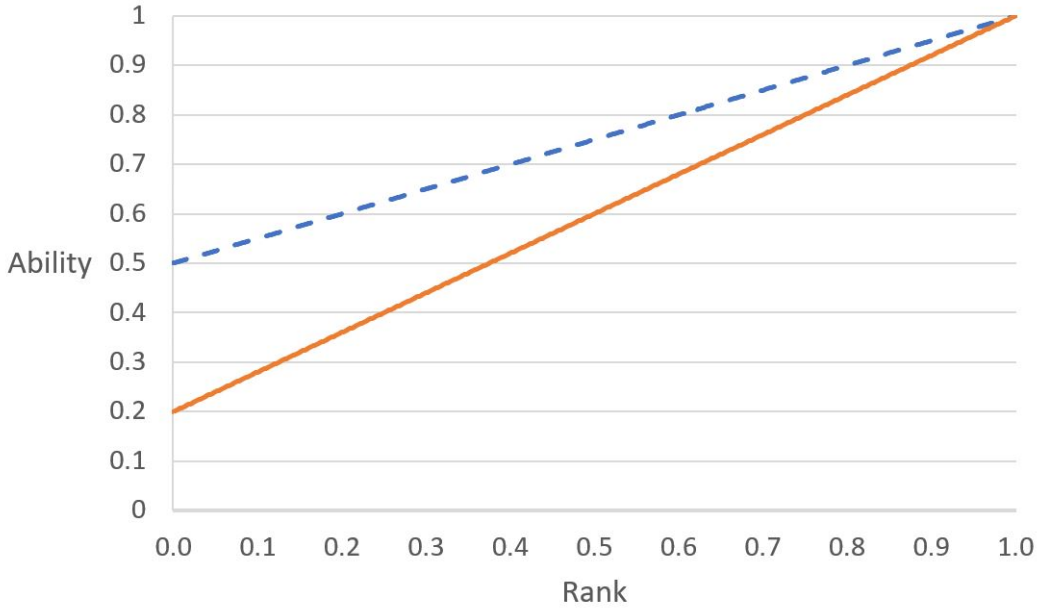
Figure 3: Proportion of omitters in function of number of items



Notes. Horizontal line: number of items. Baseline calibration: 3 options per item, the notional mark corrects for pure guessing ($\theta^* = -0.50$); examinees are moderately loss averse ($\lambda = 1.5$), uniform ability distribution.

Examinees' rank and ability

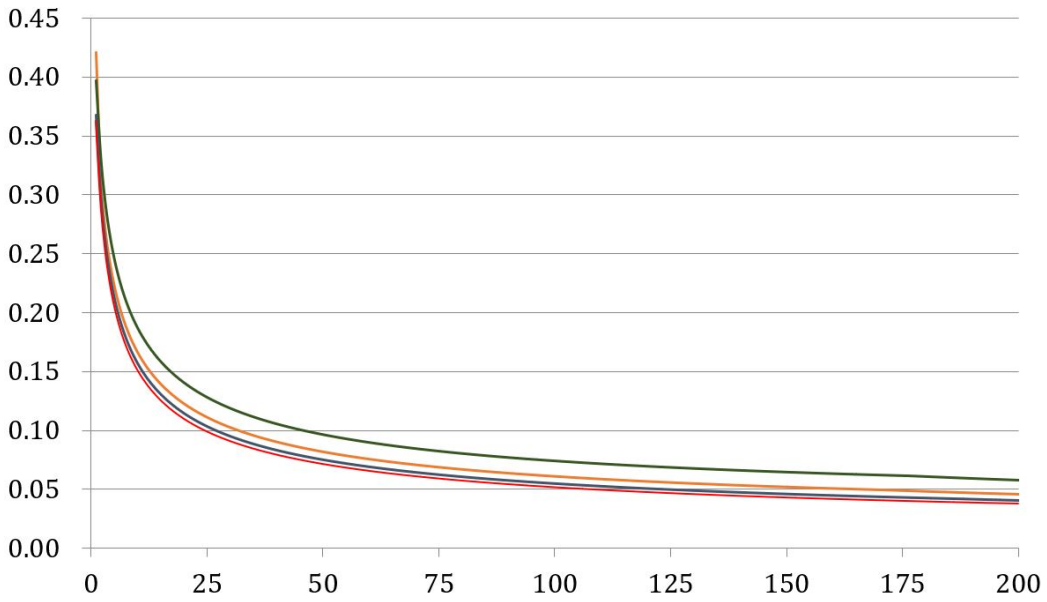
Figure 4: Examinees' rank and ability in function of the number of options per item, uniform ability distributions



Notes. Blue solid line: examinees ability for $m = 5$ options. Orange dotted line: examinees ability for $m = 2$ options. Rank: examinees' relative standings. Reading: half of examinees are less able than examinee whose rank is 0.5. Her chance of correctly selecting the right option is 75% with two options and 60% with five options.

Measurement errors and test length

Figure 5: Root mean squared errors and number of items for 2, 3, 4 and 5 options per item



Notes. Horizontal line: number of items. Green upper line: RMSE for $m = 2$ options per item; orange line: $m = 3$ options; blue line: $m = 4$ options; lower red line: $m = 5$ options. Baseline calibration: the notional mark corrects for pure guessing ($\theta^* = -0.50$); examinees are moderately loss averse ($\lambda = 1.5$), uniform ability distribution.