Notation Efficace des Questions à Choix Multiples

Alexis Direr

Abstract

Cet article étudie la notation efficace des questions à choix multiples dans lesquelles les points minimisent l'écart quadratique moyen entre le score et les connaissances des candidats. Les candidats présentent une aversion aux pertes et sont réticents à risquer des réponses sur la base de leurs connaissances. Je trouve qu'il est généralement efficace d'inciter les candidats les moins informés à omettre, sauf lorsque le test comporte un très grand nombre de questions. La note en cas d'omission est positive lorsque la taille du test est limitée et négative dans le cas inverse. L'aversion aux pertes améliore généralement l'efficacité des estimateurs en induisant spontanément davantage d'omissions. Le modèle éclaire les propriétés statistiques de deux méthodes de notation populaires : le comptage du nombre de bonnes réponses (Number right scoring) et la notation par formule (Formula scoring).

Efficient Scoring of Multiple-Choice Tests

This paper studies the optimal scoring of multiple choice tests in which the marks for wrong selections and omissions jointly minimize the mean square difference between score and examinees' abilities. Examinees are loss averse and, as a result, reluctant to risk answers on the basis of their knowledge. I find that it is efficient to incentivize the lowest able to omit, except when the test has a very large number of items. The mark for omission is positive when the test size is limited and negative when it is large. Loss aversion generally

improves estimators efficiency by spontaneously inducing more omission and thereby reducing the need to bias the mark upward to encourage omission. The model sheds light on the statistical properties of two widely used scoring methods, Number right scoring and Formula scoring.

Mots clés : théorie de l'estimation; questions à choix multiples; prise de décision; aversion aux pertes

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

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

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

Je remercie pour leurs commentaires utiles Marcel Voia, Christoph Heinzel, les deux rapporteurs de la revue ainsi que les participants de la Conférence AFSE 2019 et de la Conférence Internationale de Niort 2024 sur les Risques Économiques et Financiers.

1 Introduction

Les questions à choix multiples (QCM) représentent un mode d'évaluation populaire dans le domaine de l'éducation. Ils présentent plusieurs avantages comme une notation rapide et automatisable, un large échantillonnage du programme couvert et une évaluation exempte de biais des examinateurs. Un inconvénient majeur est toutefois la difficulté à traiter les réponses sélectionnées au hasard. Les candidats qui n'ont aucun indice sur la bonne réponse peuvent néanmoins obtenir un point s'ils ont de la chance. Les candidats ont par ailleurs souvent une connaissance partielle et sélectionnent les réponses qu'ils jugent les plus probables. Alors qu'une sélection incorrecte est toujours le résultat d'un manque de connaissances, une réponse correcte peut résulter soit d'une connaissance, soit d'une supposition, soit d'une réponse au hasard, sans qu'il soit possible de distinguer ces trois cas.

Le hasard ajoute une composante d'erreur aux scores. Supposons qu'un candidat ait une probabilité de 50 % de sélectionner la bonne option. Il peut avoir de la chance et obtenir un score moyen de 60 %, ou de la malchance et obtenir un score de 40 %. Dans les deux cas, son score est mal mesuré. Si le test comporte de nombreuses questions, la loi des grands nombres assure que l'erreur de mesure converge vers zéro. Mais pour des raisons pratiques, la plupart des tests ont un nombre limité de questions. Une notation efficace au sens statistique est susceptible atténuer ce facteur.

La méthode de notation doit également prendre en compte la possibilité laissée aux candidats de sauter une question s'ils sont incertains. La note attribuée en cas d'omission est un estimateur de la capacité moyenne des candidats qui omettent. L'omission supprime l'incertitude due au facteur chance mais introduit un autre type d'erreur de mesure qui découle de l'incapacité à classer les candidats ayant différents niveaux de connaissance partielle. Le problème est particulièrement important si une fraction significative des candidats omet de répondre.

La façon dont les notes affectent les incitations dépend également de la propension des candidats à risquer des réponses sur la base de leurs connaissances. Plusieurs études ont montré que les candidats ne répondent pas à toutes les questions même lorsque la note attendue par simple chance est supérieure à celle obtenue par omission (Sheriffs et Boomer [1954], Ebel [1968], Cross et Frary [1977], Bliss [1980], Pekkarinen [2015]). Ces observations ne sont pas compatibles avec des candidats neutres au risque. Une telle inclinaison est introduite en supposant que les candidats sont averses aux pertes : leur désutilité de perdre un point est supérieure à leur utilité de gagner un point. Cela crée un biais en faveur de l'omission, dont les conséquences pour la conception de la règle de notation sont aussi examinées.

À cette fin, un modèle de notation statistiquement efficace est étudié, dont les notes minimisent une fonction d'erreur de mesure. Le problème diffère d'une procédure d'estimation de moyenne standard car les notes servent deux objectifs à la fois. Elles fournissent une estimation de la compétence des testés à travers le calcul d'un score individuel, mais elles influencent également les candidats dans leur choix entre la sélection et l'omission, ce qui à son tour modifie les conditions dans lesquelles les compétences sont estimées. Pour étudier dans quelle mesure ces deux objectifs interagissent, la règle de notation minimise la différence quadratique moyenne entre les scores des candidats et leurs connaissance.

Je trouve que la règle de notation efficace est très sensible à la taille du test. Lorsqu'un nombre limité de questions est proposé aux candidats, les réponses des moins informés sont trop bruitées pour permettre une estimation précise de leurs connaissances. La note efficace pour l'omission est positive afin d'inciter les moins informés à omettre et à révéler leur type. Moins il y a de questions, plus il devrait y avoir d'omissions et plus la note pour l'omission devrait être élevée. L'aversion aux pertes améliore généralement l'efficacité des estimateurs en induisant spontanément davantage d'omissions, réduisant ainsi la nécessité de biaiser la note d'omission vers le haut. Lorsque le test comporte un large échantillon de questions, la capacité des candidats les moins informés est estimée avec plus de précision, éliminant la

nécessité de les inciter à omettre. La note en cas d'omission devient négative de sorte que tous les candidats répondent.

Les questions à choix multiples comme outil d'évaluation ont une longue histoire. Elles ont été administrées pour la première fois à grande échelle pendant la Première Guerre mondiale par l'armée américaine pour identifier rapidement les capacités de centaines de milliers de recrues (Ebel [1979]). Leur adoption s'est répandue rapidement dans divers domaines, comme les tests d'intelligence (Pintner [1923]) ou dans l'éducation. Kelly [1916] est le premier chercheur à rapporter et à étudier l'utilisation des questions à choix multiples pour mesurer les compétences en lecture des enfants. La standardisation du processus d'évaluation s'est avérée particulièrement adaptée aux examens à grande échelle et enjeux élevés, comme le Scholastic Aptitude Test (SAT) et le Graduate Record Examination (GRE), pour prendre deux exemples importants aux États-Unis.

Dans quelle mesure les tests fournissent des mesures précises et valides des compétences a été étudié pendant plus d'un siècle par la psychométrie, un domaine de recherche à l'intersection de la psychologie et des statistiques. Beaucoup de ses résultats ont été incorporés dans ce qui est considéré aujourd'hui comme la théorie classique des tests (voir par exemple, McDonald [1999]). Elle est basée sur l'hypothèse centrale que le score d'une personne à un test est la somme d'un vrai score (*true score*) et d'un score d'erreur (Harvill [1991]). Le programme de recherche s'est développé autour de deux concepts clés : la fiabilité et la validité. Une mesure est fiable si elle produit des résultats similaires dans des conditions cohérentes. Les scores fiables sont reproductibles d'un test à un autre (Traub et Rowley [1991]). Une mesure valide est celle qui mesure ce qu'elle est censée mesurer. Une volumineuse littérature théorique et empirique a appliqué ces concepts aux propriétés de différentes règles de notation (par exemple, Diamond et Evans [1973] ; Burton [2001] ; Lesage et al. [2013]).

Le modèle s'écarte des études psychométriques de deux façons. Premièrement,

une attention particulière est accordée à l'interaction entre la règle de notation, les préférences face au risque et l'estimation des compétences. Dans la plupart des études existantes, les préférences face au risque ne sont pas modélisées ou, lorsqu'elles le sont, les candidats sont considérés comme neutres au risque. En posant l'hypothèse réaliste conjointe d'aversion aux pertes et de cadrage étroit (*narrow framing*), les candidats affichent un biais vers l'omission conformément à la littérature empirique (par exemple, Akyol et al. [2016]). Deuxièmement, la littérature s'est concentrée sur des règles de notation ad hoc dans lesquelles les notes pour les mauvaises réponses et l'omission ne sont pas dérivées de principes premiers. Les deux notes sont rendues endogènes ici en modélisant explicitement l'écart du score réel par rapport au score vrai, qui est le score que les candidats obtiendraient si leur capacité était parfaitement observée.

Quelques articles ont également analysé la notation endogène. Espinosa et Gardeazabal [2010] simulent un modèle de notation optimale avec une aversion au risque hétérogène et une difficulté des questions variable. Ils trouvent une pénalité relativement élevée pour dissuader les réponses au hasard. Budescu et Bo [2015] simulent un modèle de notation optimale sous différentes hypothèses (aversion aux pertes hétérogène et mauvaise calibration des probabilités). Ils constatent qu'une pénalité négative aggrave le biais du score et l'écart-type, et diminue la corrélation entre les scores simulés et les scores vrais. Akyol, Key et Krishna [2016] modélisent le comportement des étudiants lors des tests, et utilisent le modèle pour estimer leurs préférences face au risque. Ils simulent ensuite des règles de notation contrefactuelles et constatent que l'augmentation de la pénalité pour les mauvaises réponses a un impact significatif sur l'omission, ce qui à son tour améliore l'estimation des capacités des candidats. L'hétérogénéité de l'aversion au risque a peu d'influence sur les scores simulés, ce qui plaide en faveur d'une pénalité négative. Dans ces articles, seule la pénalité pour mauvaises réponses est optimisée, alors que dans le présent modèle, les notes pour les mauvaises réponses et l'omission sont toutes deux endogènes. Une autre différence majeure est l'utilisation de l'erreur

quadratique moyenne, qui permet des résultats analytiques et des interprétations simples. En supposant que les candidats ne diffèrent que par leurs connaissances, et non par des traits de personnalité comme l'aversion au risque, le présent modèle n'aborde pas la question de l'impact des préférences hétérogènes sur la validité des mesures.

Finalement, l'article est relié à la littérature sur le design optimal dans les modèles de choix discrets (voir à ce sujet Train [2009] et Louviere et al. [2000]). Ces travaux partagent des préoccupations méthodologiques avec l'article, notamment sur l'élicitation optimale des préférences et la conception des mécanismes d'incitation. En estimant des caractéristiques inobservables à partir de décisions observables, le modèle est également relié au cadre économétrique de McFadden [1981] pour l'inférence de variables latentes à partir de choix discrets.

Le reste de l'article est organisé comme suit. La section 2 présente le modèle de notation et ses constituants : le score vrai, l'aversion aux pertes et l'erreur quadratique moyenne. La section 3 présente quelques propriétés analytiques du modèle de notation. La section 4 calibre un modèle stylisé et présente les résultats des simulations. La section 5 compare le modèle avec les propriétés des deux règles de notation les plus utilisées, le comptage du nombre de bonnes réponses (*Number right scoring*) et la notation par formule (*Formula scoring*). La section 6 conclut.

2 Le modèle de notation

2.1 La règle de notation

Le test est composé de n questions. Chaque question a m réponses possibles, une correcte et $m-1$ incorrectes. Les questions sont supposées être bien rédigées, sans réponses évidentes, pièges ou formulations ambiguës. Les options sont correctement randomisées au sein de chaque question. Il y a suffisamment de temps pour répondre

à toutes les questions. Toutes les questions sont d'une difficulté égale. Les candidats ont une probabilité constante p de répondre correctement à chacune. La probabilité varie selon les candidats et est une mesure de leur connaissance du domaine couvert par le test.

L'objectif du concepteur du test est de concevoir une règle de notation permettant aux candidats de recevoir un score aussi proche que possible de leur compétence. Le traitement d'une question par un candidat a trois résultats possibles auxquels sont attribués des points spécifiques. Les points donnés à une sélection correcte sont normalisée à 1. Les points attribués aux sélections erronées sont notée θ et les points des omissions γ. Des restrictions minimales sont imposées :

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

Le score final est la somme des points obtenus pour toutes les questions divisée par le nombre n de questions. Soit $z \in [0, n]$ le nombre de questions omises et $\tilde{x} \in [0, n-z]$ le nombre de sélections correctes parmi les $n-z$ questions auxquelles le candidat a répondu. $\tilde{x}$ suit une distribution binomiale $B(n-z, p)$. Le score des candidats est la somme des bonnes réponses, des mauvaises réponses et des questions omises pondérées respectivement par 1, θ et γ, et divisée par le nombre de questions :

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

2.2 Score vrai

Le score vrai $s(p)$ est défini comme le score espéré obtenu dans un test avec des points de 1 pour les sélections correctes et θ^* pour celles erronées, en supposant que les candidats n'omettent pas :

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

C'est la composante du score observé non influencée par les événements aléa-

toires (Harvill [1991]). Le score vrai sera en général inférieur à p ($\theta^* < 0$) afin de pénaliser les réponses au hasard. Un candidat sélectionnant des choix au hasard obtiendrait sinon un score strictement positif : $s(p) = p = 1/m$, avec m le nombre de choix par question. Une façon d'éliminer le facteur chance consiste à fixer $\theta^* = -1/(m-1)$, comme dans la notation par formule (*Formula scoring*) (voir Section 5). Le score attendu du candidat est nul dans le cas où il sélectionne au hasard pour toutes les questions :

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

D'autres corrections sont possibles. Si certains candidats disposent de fausses connaissances, ils pourraient obtenir de moins bons résultats qu'en sélectionnant au hasard. Dans de nombreuses situations, être conscient de son ignorance sur un sujet est susceptible d'encourager les individus à rechercher des informations avant de décider. Dans ce cas, le pur hasard (ou l'omission) reflète une capacité minimale qui pourrait être récompensée en fixant θ^* au-dessus de $-1/(m-1)$. Le cas de la désinformation est exclu dans cet article en supposant que la capacité la plus basse des candidats, notée p_0, est égale à $1/m$.

2.3 Préférences face au risque

L'omission procure des points sans risque, à l'inverse de la sélection, à moins que les candidats ne soient certains de l'option correcte. Le choix entre un résultat certain et un résultat risqué est modélisé à travers trois hypothèses. Premièrement, les candidats obtiennent l'utilité $u(x)$ des points x de chaque question, et non du score moyen ou agrégé. L'hypothèse de cadrage étroit (*broad framing*, Tversky et Kahnemman [1981]) selon laquelle les personnes ne regroupent pas toutes les sources de risque avant de décider, s'est avéré pertinente dans divers contextes de décision impliquant des risques multiples (Tversky et Kahnemman [1981], Read, Loewenstein et Rabin [1999]).

Deuxièmement, les candidats se concentrent sur les pertes et les gains et surpondèrent les pertes. Ils sont plus affectés par les résultats négatifs que par les résultats positifs de même ampleur. L'aversion aux pertes est une caractéristique centrale de la théorie des perspectives (*prospect theory*) de Kahneman et Tversky [1979]. Sa validité repose sur de nombreuses preuves expérimentales, particulièrement lorsqu'elle est associée au cadrage étroit. Bereby-Meyer, Meyer et Flascher [2002] fournissent des preuves de cadrage étroit et d'aversion aux pertes dans le contexte des examens.[1]

Troisièmement, l'utilité dérivée des points est linéaire : $u(1) = 1$, $u(\gamma) = \gamma$. Appliquée au contexte des examens, la perte d'utilité associée à une mauvaise sélection est proportionnelle aux points : $u(\theta) = \lambda\theta$, avec λ le coefficient d'aversion aux pertes. Une mauvaise sélection est perçue comme une perte par les candidats quel que soit le signe des points : $\lambda > 1$ si $\theta \leq 0$ et $\lambda < 1$ si $\theta > 0$. L'aversion aux pertes est synthétiquement définie par la condition de signe

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

La neutralité aux pertes est équivalente à la neutralité au risque si $\lambda = 1$. Les candidats averses aux pertes n'aiment pas le risque. Ils préfèrent toujours des points sûrs à des points aléatoires avec la même espérance.

Étant donné une règle de notation (γ, θ), l'omission est préférée à la sélection si les points pour l'omission sont supérieurs aux points attendus, pondérés par les pertes, d'une réponse :

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

[1] Voir aussi Budescu et Bo [2015]. L'hypothèse conjointe selon laquelle les personnes ont tendance à se concentrer sur les gains et les pertes individuels plutôt que sur les résultats moyens est parfois appelée aversion myope aux pertes (Barberis, Huang, et Thaler [2006]; Barberis et Huang [2008]). Le cadrage étroit est également en accord avec les observations montrant que les individus ne deviennent pas neutres au risque lorsqu'ils passent de grands tests comportant de nombreuses questions indépendantes, dont le risque disparaît une fois agrégé (Pekkarinen [2015]; Akyol, Key et Krishna [2022]; Iriberri et Rey-Biel [2021]).

$\bar{p}$ est défini comme la probabilité de succès des candidats indifférents entre choisir et omettre :

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

Les candidats omettent lorsqu'ils ne sont pas assez confiants dans leur sélection : $p \leq \bar{p}$, et répondent dans le cas contraire. $\bar{p}$ dépend positivement des points γ et négativement de la pénalité θ. Comparée au cas de la neutralité au risque ($\lambda = 1$), l'aversion aux pertes élève le seuil de probabilité $\bar{p}$:

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

2.4 Erreur quadratique moyenne

Le score vrai du candidat est estimé par le taux de réussite du répondant en cas de réponse, ou en attribuant des points constants en cas d'omission, ce qui signale une faible capacité en moyenne. Les deux méthodes produisent des erreurs de mesure.

Considérons d'abord un candidat dont la probabilité de succès est $p > \bar{p}$. Puisque p ne varie pas d'une question à l'autre et que toutes les questions ont la même difficulté, il répond à toutes et obtient le score :

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

qui est interprété comme un estimateur linéaire ponctuel du score vrai $s(p)$. Sa qualité peut être mesurée par des méthodes statistiques et optimisée par le choix adéquat de θ et γ. L'erreur quadratique moyenne (MSE) du score observé $\tilde{s}$ obtenu par un candidat avec une probabilité de succès p est la différence quadratique moyenne entre $\tilde{s}$ et le score vrai $s(p)$:

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

L'erreur quadratique moyenne est une mesure couramment utilisée de la performance d'un estimateur. Elle est analytiquement traitable et se prête à la décompo-

sition intuitive :

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

La première composante est la variance du score observé. La seconde est le biais au carré, qui mesure à quel point le score attendu s'écarte de sa moyenne théorique. Le critère de l'erreur quadratique moyenne contrôle à la fois les fluctuations d'échantillon et la précision de l'estimateur.

Considérons maintenant un candidat dont la probabilité de succès est $p \leq \bar{p}$. Puisque p est constant pour toutes les questions, il omet toutes les questions et obtient le score γ. Il obtiendrait le score vrai $s(p)$ si sa capacité était parfaitement mesurée. L'erreur quadratique du candidat est dans ce cas le biais au carré :

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

Bien que les probabilités individuelles de succès ne soient pas observées par le concepteur du test, leur distribution est supposée connue. Soit $f(p)$ la fonction de densité de probabilité de succès. Le concepteur choisit les points θ et γ de façon à minimiser l'erreur quadratique moyenne moyennée sur les candidats :

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

Comme la composante MSE de la sélection, la composante MSE de l'omission, normalisée par leur proportion $F(\bar{p})$ dans la population, se prête à une décomposition. Réécrivons d'abord l'erreur quadratique moyenne :

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

où $E_{|\mathrm{omit}}$ est l'espérance conditionnelle de la compétence des candidats qui omettent. Notons $\bar{s}(p)$ leur compétence moyenne :

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

La variance conditionnelle de leur compétence est :

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

La composante MSE des omissions s'écrit par conséquent :

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

L'erreur de mesure totale a deux composantes. Le terme de variance mesure avec quelle étendue la compétence s'écarte de sa moyenne. Plus il y a de candidats qui omettent (plus $\bar{p}$ est élevé), plus la dispersion est grande et plus l'erreur quadratique moyenne est élevée. Le second terme est le biais au carré qui mesure l'écart entre les points en cas d'omission et la compétence moyenne de ceux qui omettent. Il s'ensuit que, tant que certains candidats omettent, le terme de variance dans (6) est une borne inférieure quel que soit le nombre de questions dans le test. C'est une différence majeure avec la composante MSE des réponses où l'erreur moyenne peut être arbitrairement réduite avec n suffisamment grand.

3 Score efficace

Lorsque le nombre de questions augmente, la compétence est estimée à partir des réponses avec une précision illimitée. Au contraire, puisque l'omission ne signale une faible compétence qu'en moyenne, les candidats qui omettent créent des erreurs de mesure qui ne disparaissent pas avec la longueur du test. Il s'ensuit que lorsque le nombre de questions dans le test est arbitrairement grand, les scores sont des estimateurs parfaits de la compétence à condition qu'aucun candidat n'omette. Le score efficace est tel que l'erreur quadratique moyenne tend vers zéro, le score statistique converge vers le score vrai ($\hat{\theta} \to \theta^*$) et même les candidats les moins compétents répondent $\hat{\gamma} < p_0 + (1 - p_0)\lambda\theta^*$.

Lorsque le nombre de questions est fini, les compétences des candidats sont estimées avec des erreurs dues aux fluctuations d'échantillon fini. Les conditions de

premier ordre du programme de minimisation (4) sont :

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

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

Le terme commun aux deux équations

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

est l'effet net sur l'erreur quadratique moyenne des candidats marginaux avec une compétence $\bar{p}$ changeant leur choix de la sélection à l'omission. Les termes $d\bar{p}/d\gamma$ et $d\bar{p}/d\theta$ sont les effets des points sur le seuil de probabilité $\bar{p}$ (voir (2)). Augmenter γ ou baisser θ encourage tous deux l'omission et élargit le groupe des candidats qui omettent :

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

4 Notation simulée

Ce problème n'ayant pas de solutions analytiques dans le cas général, des règles de notation simulées sont présentées dans la section suivante.

4.1 Stratégie de simulation

Tversky et Kahneman [1992] estiment un coefficient d'aversion aux pertes $\lambda = 2.25$ dans la théorie cumulative des perspectives. Comme il n'est pas entièrement clair comment un paramètre estimé à partir de choix impliquant des résultats monétaires se traduit dans le contexte des notes, trois niveaux plausibles d'aversion aux pertes

sont considérés : neutralité aux pertes ($\lambda = 1$), aversion modérée aux pertes ($\lambda = 1.5$) et forte aversion aux pertes ($\lambda = 2.5$).[2]

La distribution de compétences parmi les candidats change avec la difficulté du test et le niveau de maîtrise des candidats. Estimer la distribution des compétences à partir de tests réels dépasse le cadre de cet article. Sans information spécifique sur la population et l'examen, je choisis une distribution uniforme sur l'espace des compétences $[p_0, 1]$.

L'erreur quadratique moyenne (4) est calculée sur une double grille de valeurs pour les paramètres $\theta \in [\underline{\theta}, \theta^*]$ et $\bar{p} \in [1/m, 1]$. Les points optimaux γ sont récupérés pour chaque couple $(\theta, \bar{p})$ par la condition $\gamma = \bar{p} + (1 - \bar{p})\lambda\theta$. Les deux grilles sont composées de 2500 points chacune, de sorte que $2500^2 = 6,250,000$ différentes valeurs d'erreur quadratique moyenne sont calculées. Les points efficaces correspondent à la valeur la plus basse calculée.

J'utilise comme métrique d'ajustement l'erreur quadratique moyenne (RMSE), la moyenne géométrique des erreurs de mesure pour tous les candidats :

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

Je calcule également le biais sur le score des candidats qui omettent $\hat{\gamma} - \bar{s}(p)$ (voir son expression (5)), qui informe sur la mesure dans laquelle l'estimateur de la compétence des candidats qui omettent est déformé pour encourager (si positif) ou dissuader (si négatif) l'omission. Les incitations à omettre sont mesurées par la différence de points $\gamma - \theta$.

[2] Les erreurs pourraient être positivement notées ($\theta > 0$) en théorie. Les mauvaises réponses seraient toujours perçues comme une perte par les candidats, c'est-à-dire $\lambda \in (0, 1)$. Cette situation ne se produit jamais dans les simulations.

4.2 Résultats

J'étudie d'abord un modèle de base dans lequel différents nombres de questions ($n = 1, 5, 10, 20, 40, 80, 200, \infty$) sont considérés. Chaque question a $m = 3$ options.[3] Les scores vrais $s(p)$ sont calculés étant donné des points notionnels corrigeant pour les réponses au hasard pur : $\theta^* = -1/(m-1)$. Le coefficient d'aversion aux pertes est fixé à 1.5.

Le tableau 4 en Annexe A présente les points efficaces et les principales statistiques en fonction de la taille du test n pour le calibrage de référence. Le graphique 1 en Annexe B montre comment les points efficaces $\hat{\gamma}$ varient avec n.

Deux stratégies de notation distinctes émergent selon la taille du test. Lorsque le nombre de questions est limité (moins de 170 dans le graphique 1), l'omission des individus les moins compétents est encouragée pour pallier l'imprécision d'estimation. Les points pour l'omission sont positifs et fixés au-dessus de la compétence moyenne des candidats qui omettent (voir le biais d'omission dans le tableau 4).

Lorsque le nombre de questions est suffisamment grand, l'omission est entièrement découragée. Le graphique 2 montre que la proportion de candidats qui omettent tombe à zéro. La différence de points $\hat{\gamma} - \hat{\theta}$, qui mesure les incitations à omettre, diminue de 0.6 à 0.33. Par conséquent, il est efficace de forcer la sélection lorsque le test comporte un nombre suffisant de questions.

L'intuition est la suivante. Comme les individus les moins compétents sélectionnent des options avec peu de connaissances, l'erreur d'estimation de leur compétence à partir de leurs réponses est élevée. Il est plus efficace de les inciter à omettre et à révéler ainsi leurs faibles connaissances. À mesure que des questions supplémentaires sont incluses dans le test, les compétences sont estimées avec une précision croissante, même pour les moins compétents.

[3] Des formules de score avec $m = 2$ à 5 options ont également été simulées. Les résultats, non inclus dans l'article, montrent que les mesures d'erreur ne varient pas beaucoup, à condition que le nombre total d'options $m \times n$ dans le test reste constant.

Notons que, bien que l'omission puisse être un moyen efficace de réduire la variance des estimateurs, elle est toujours soumise à un compromis entre deux types d'erreurs de mesure. Regrouper trop de candidats qui omettent ne signalerait pas beaucoup d'informations sur leur vraie compétence (le terme de variance (6) serait grand).

Sauf dans le cas extrême $n = 1$ dans lequel plus de 80% des candidats omettent, la pénalité efficace en cas de mauvaise sélection $\hat{\theta}$ est supérieure au point notionnel θ^* (Tableau 4), c'est-à-dire que la pénalité est plus douce que ce que prescrit une simple correction pour les réponses au hasard. Elle se situe néanmoins dans le voisinage proche du point notionnel, suggérant qu'une pénalité fixe égale au point notionnel pourrait s'avérer une bonne approximation de la règle efficace. Les points pour l'omission sont plus sensibles à la taille du test n que la pénalité pour les mauvaises réponses. Le comportement des moins compétents est en effet mieux ciblé par les points pour l'omission que par la pénalité pour mauvaise réponse qui impacte tous les candidats, y compris les plus compétents qui n'omettent jamais.

4.3 Notation efficace et préférences face au risque

Dans quelle mesure les préférences face au risque interagissent-elles avec la règle de notation et l'efficacité des estimateurs ? Les candidats averses aux pertes surpondèrent la perte d'utilité lorsqu'ils se trompent et ont tendance à s'abstenir plus souvent. Les conséquences pour l'omission sont cependant ambivalentes lorsque la règle de notation est efficace (voir Tableau 1). La proportion de candidats qui omettent augmente avec l'aversion aux pertes lorsque l'omission est autorisée, mais le passage à un régime sans omission se produit plus tôt à mesure que le nombre de questions augmente.

Pour comprendre pourquoi, rappelons que la proportion de candidats qui omettent dépend des incitations données par les points. Sachant n, le nombre de ques-

Table 1: Proportion efficace de candidats qui omettent (%) et aversion aux pertes

Nombre de questions (n)	1	5	10	20	40	80	200	∞
Neutralité au risque ($\lambda = 1$)	43.4	26.0	19.6	14.4	10.5	7.6	4.9	0.00
Aversion modérée aux pertes ($\lambda = 1.5$)	83.5	39.4	30.6	24.9	21.1	18.6	0.00	0.00
Forte aversion aux pertes ($\lambda = 2.5$)	85.7	46.8	38.0	32.8	29.9	0.00	0.00	0.00

Modèle : les points notionnels corrigent pour les réponses au hasard pur ($\theta^* = -0.50$). Voir les Tableaux 3, 4 et 5 pour des statistiques détaillées. Lecture : 26% des candidats neutres au risque omettent dans un test efficace avec 5 questions.

tions du test, réduire la proportion de candidats qui omettent nécessite de biaiser les points pour l'omission vers le bas, ce qui comporte ses propres coûts en termes d'erreurs de mesure. Plus les candidats sont averses aux pertes, plus les points pour l'omission nécessaires pour atteindre une part désirée de candidats qui omettent sont bas. Par conséquent, il est moins coûteux de passer à un régime sans omission lorsque n est suffisamment grand.

L'aversion aux pertes favorise l'efficacité dans une certaine mesure, comme le montrent les erreurs quadratiques moyennes rapportées dans le Tableau 2. Les erreurs diminuent avec l'aversion aux pertes pour les tests avec un nombre limité de questions $n \leq 40$. Il n'y a pas de différences visibles pour les tests avec un n plus grand.

Lorsque le nombre de questions est fini, il est généralement efficace d'inciter les moins compétents à omettre. Si les candidats sont averses aux pertes, ces derniers omettent spontanément sans qu'il soit nécessaire de déformer les points.

Table 2: Erreur quadratique moyenne (RMSE) et aversion aux pertes

Nombre de questions (n)	1	5	10	20	40	80	200	∞
Neutralité au risque ($\lambda = 1$)	0.406	0.241	0.181	0.133	0.097	0.070	0.045	0.00
Aversion modérée aux pertes ($\lambda = 1.5$)	0.386	0.221	0.166	0.122	0.089	0.066	0.046	0.00
Forte aversion aux pertes ($\lambda = 2.5$)	0.324	0.196	0.151	0.119	0.097	0.072	0.046	0.00

Modèle : les points notionnels corrigent pour les réponses au hasard pur ($\theta^* = -0.50$). Voir les Tableaux 3, 4 et 5 pour des statistiques détaillées.

5 Relation avec les méthodes de notation existantes

Le modèle éclaire l'efficacité des deux méthodes de notation les plus utilisées, le comptage du nombre de bonnes réponses (NRS) et la notation par formule (FS). NRS compte simplement le nombre de sélections correctes et divise la somme par le nombre total de questions. Les omissions et sélections erronées comptent pour zéro (θ, $\gamma = 0$). Une critique souvent faite à cette méthode est que les candidats sélectionnant des options au hasard obtiennent un score positif en espérance égal à $1/m$. FS fixe également les points pour omission à zéro ($\gamma = 0$) mais impose une pénalité pour une sélection incorrecte égale à $-1/(m-1)$. La formule égalise les scores attendus des réponses au hasard pur et de l'omission comme montré dans (1) (Thurstone [1919], Holzinger [1924]).

La supériorité d'une de ces règles par rapport à l'autre est encore débattue dans la littérature psychométrique. En mettant en œuvre une notation négative, FS encourage l'omission, ce qui augmente la fiabilité (Lord [1975], Mattson [1975], Burton [2001]). Certains auteurs ont soutenu que FS ne mesure pas seulement la maîtrise des connaissances du domaine mais reflète également les stratégies de réponse des candidats et leur comportement face au risque (par exemple, Votaw [1936]; Frary [1988]; Budescu et Bar-Hillel [1993]). NRS fournit plus d'incitations

à répondre à toutes les questions, ce qui minimise ce type de biais.

Les principales lacunes de NRS et FS par rapport au présent modèle résident dans la façon dont ils traitent l'omission. Les points ne sont pas ajustés pour l'échantillon fini afin d'induire suffisamment d'omissions lorsque le nombre de questions est limité. Des points pour l'omission à zéro ne sont pas efficaces quelle que soit la taille du test (voir le graphique 1). En fixant la différence de points $\gamma - \theta$ à zéro, NRS dissuade l'omission, ce qui n'est efficace que pour les grands tests. FS fournit plus d'incitations à omettre mais uniquement en augmentant la pénalité pour une réponse incorrecte. Au contraire, le modèle montre que pour un large éventail de tailles de test, les points pour l'omission sont strictement positifs car ils remplissent deux fonctions : la connaissance partielle est créditée et l'omission est encouragée.

Quelle que soit la méthode, les points dans FS et NRS ne sont pas dérivés d'un modèle d'estimation explicite. Par exemple, la correction pour les réponses au hasard faite dans FS part de l'hypothèse que les candidats ignorants choisissent de répondre à toutes les questions au hasard. L'hypothèse n'est pas cohérente avec le présent modèle selon lequel les candidats ayant des connaissances insuffisantes devraient être incités à omettre, pas à répondre. Cela implique que les points attribués pour omission aux candidats totalement ignorants ne devraient pas être zéro mais strictement positifs.

6 Conclusion

Trois principales leçons peuvent être tirées du modèle de notation. Premièrement, le concepteur du test devrait inclure un grand nombre de questions lorsque c'est possible afin d'exploiter la loi des grands nombres. Les simulations numériques suggèrent un nombre supérieur à 40 et pouvant aller jusqu'à 100. Cependant, l'élaboration de tests à choix multiples de grande taille nécessite des compétences

et du temps. Une source d'inefficacité non traitée dans cet article est l'inclusion de questions mal formulées, aux réponses ambiguës, redondantes ou évidentes.

Deuxièmement, l'efficacité dicte de cibler une proportion décroissante de candidats qui omettent avec la longueur du test. Si le nombre de questions est grand, la compétence est généralement mieux estimée par la sélection que par l'omission. La sélection peut être forcée en fixant des points négatifs pour l'omission. Si le nombre de questions est limité, l'omission peut être encouragée par des points positifs. Moins il y a de questions, plus l'omission est nécessaire et plus les points sont élevés. La proportion résultante de candidats qui omettent est assez significative dans les petits tests. Par ailleurs, les instructions données aux candidats devraient être cohérentes avec la stratégie de notation. Si le nombre de questions est limité, les candidats devraient être encouragés à omettre. Dans le cas contraire, ils devraient être invités à répondre à toutes les questions même s'ils ne sont pas sûrs des bonnes réponses.

Troisièmement, le comportement des candidats peu compétents est mieux ciblé par les points pour l'omission que par la pénalité pour les mauvaises réponses. Cette constatation reformule un débat de longue date sur les avantages relatifs de la notation par formule et de le comptage des bonnes réponses qui se concentre exclusivement sur la meilleure valeur que les points pour les mauvaises réponses devraient prendre.

Le modèle a fait un certain nombre d'hypothèses simplificatrices dont les implications pour la stratégie d'estimation pourraient être intéressantes à étudier à l'avenir. Premièrement, les études expérimentales en psychologie suggèrent que les testés sont généralement trop confiants dans leurs réponses (par exemple, Keren [1991]; Yates [1990]). La surconfiance réduit le taux d'omission et peut avoir un impact sur l'efficacité de l'estimation, surtout si la tendance est corrélée avec la compétence (Lichtenstein et Bishhoff [1977]; Heath et Tversky [1991]). Une question connexe est de savoir comment noter la désinformation, qui survient lorsque

les candidats ont des connaissances erronées (Burton [2004]).

Deuxièmement, les tests pourraient être modélisés de manière plus réaliste en considérant des questions de difficulté variable, par exemple si la difficulté des questions augmente à mesure que les candidats progressent dans le test. La probabilité d'une bonne réponse et les incitations à omettre pourraient fluctuer d'une question à l'autre. Il pourrait alors être intéressant d'adapter les points à la difficulté des questions.

Troisièmement, le modèle suppose que les candidats ne diffèrent que par leurs connaissances, et non par des traits de personnalité comme l'aversion au risque ou le degré de confiance en soi. Des recherches récentes d'Akyol et al. [2022] suggèrent que l'hétérogénéité des préférences face au risque peut affecter la façon dont les candidats répondent aux incitations de pénalité. Une partie de l'aléa provient de l'hétérogénéité non observée, ce qui nécessite des procédures d'estimation plus complexes.

Quatrièmement, l'erreur quadratique moyenne n'est pas un critère d'optimisation adapté lorsque le classement relatif est l'objectif principal de l'examen. L'optimisation basée sur les statistiques de rang nécessite généralement des méthodes non paramétriques qui pourraient être explorées dans des recherches futures.

References

AKYOL P., KEY J. et KRISHNA K. [2022], « Hit or miss? Test taking behavior in multiple choice exams », *Annals of Economics and Statistics*, 147, p. 3-50.

BARBERIS N., HUANG M. et THALER R. [2006], « Individual preferences, monetary gambles, and stock market participation: A case for narrow framing », *American Economic Review*, 96, p. 1069-1090.

BARBERIS N. et HUANG M. [2008], « The loss aversion/narrow framing ap-

proach to the equity premium puzzle », in MEHRA R. (ed.), *Handbook of the Equity Risk Premium*, Elsevier Science, NBER version.

BEREBY-MEYER Y., MEYER J. et FLASCHER O.M. [2002], « Prospect theory analysis of guessing in multiple choice tests », *Journal of Behavioral Decision Making*, 15, p. 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, p. 147-153.

BUDESCU D.V. et BAR-HILLEL M. [1993], « To guess or not to guess: A decision-theoretic view of formula scoring », *Journal of Educational Measurement*, 30 (4), p. 277-291.

BUDESCU D.V. et BO Y. [2015], « Analyzing test-taking behavior: Decision theory meets psychometric theory », *Psychometrika*, 80 (4), p. 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), p. 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, p. 585-595.

EBEL R.L. [1968], « Blind guessing on objective achievement tests », *Journal of Educational Measurement*, 5, p. 321-325.

EBEL R.L. [1979], *Essentials of educational measurement* (3rd ed.), Englewood Cliffs, NJ, Prentice Hall.

CROSS L.H. et FRARY R.B. [1977], « An empirical test of Lord's theoretical results regarding formula-scoring of multiple-choice tests », *Journal of Educational Measurement*, 14, p. 313-321.

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

ESPINOSA M.P. et GARDEAZABAL J. [2010], « Optimal correction for guessing in multiple-choice tests », *Journal of Mathematical Psychology*, 54 (5), p. 415-425.

FRARY R.B. [1988], « Formula scoring of multiple-choice tests (correction for guessing) », *Educational Measurement: Issues and practice*, 7, p. 33-38.

HARVILL L.M. [1991], « Standard error of measurement », *Educational Measurement: Issues and Practice*, 10, p. 33-41.

HEATH C. et TVERSKY A. [1991], « Preference and belief: Ambiguity and competence in choice under uncertainty », *Journal of Risk and Uncertainty*, 4 (1), p. 5-28.

HOLZINGER K.J. [1924], « On scoring multiple-response tests », *Journal of Educational Measurement*, 15, p. 445-447.

IRIBERRI N. et REY-BIEL P. [2021], « Brave boys and play-it-safe girls: gender differences in willingness to guess in a large scale natural field experiment », *European Economic Review*, 131, 103603.

KAHNEMAN D. et TVERSKY A. [1979], « Prospect theory: An analysis of decision under risk », *Econometrica*, 47(2), p. 263-92.

KELLY F.J. [1916], « The Kansas silent reading tests », *Journal of Educational Psychology*, 7(2), p. 63-80.

KEREN G. [1991], « Calibration and probability judgments: conceptual and methodological issues », *Acta Psychologica*, 77, p. 217-273.

LESAGE E., VALCKE M. et SABBE A. [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, p. 118-193.

LICHTENSTEIN S. et FISCHHOFF B. [1977], « Do those who know more also know more about how much they know? », *Organizational Behavior and Human Performance*, 20, p. 159-183.

LORD F.M. [1975], « Formula scoring and number-right scoring », *Journal of Educational Measurement*, 12, p. 7-12.

LOUVIERE J.J., HENSHER D.A. et SWAIT J.D. [2000], *Stated Choice Methods: Analysis and Applications*, Cambridge, Cambridge University Press.

MATTSON D. [1975], « The effects of guessing on the standard error of measurement and the reliability of test scores », *Educational and Psychological Measurement*, 25, p. 727-730.

MCDONALD R.P. [1999], *Test theory: A unified treatment*, Mahwah, NJ, Lawrence Erlbaum Associates.

MCFADDEN D. [1981], « Econometric models of probabilistic choice », in MANSKI C.F. et MCFADDEN D. (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, p. 198-272.

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, p. 94-110.

PINTNER R. [1923], *Intelligence testing*, New York, Holt, Rinehart and Winston.

READ D., LOEWENSTEIN G. et RABIN M. [1999], « Choice bracketing », *Journal of Risk and Uncertainty*, 19 (13), p. 171-197.

SHERIFFS A.C. et BOOMER D.S. [1954], « Who is penalized by the penalty for guessing? », *Journal of Educational Psychology*, 45, p. 81-9.

THURSTONE L.L. [1919], « A method for scoring tests », *Psychological Bulletin*, 16, p. 235-240.

TRAIN K. [2009], *Discrete Choice Methods with Simulation*, Cambridge, Cambridge University Press.

TRAUB R.E. et ROWLEY G.L. [1991], « Understanding reliability », *Educational measurement: Issues and practice*, 10(1), p. 37-45.

TVERSKY A. et KAHNEMAN D. [1981], « The framing of decisions and the psychology of choice », *Science*, 211, p. 453-458.

TVERSKY A. et KAHNEMAN D. [1992], « Advances in prospect theory: Cumulative representation of uncertainty », *Journal of Risk and Uncertainty*, 5, p. 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, p. 698-703.

YATES J.F. [1990], *Judgment and decision making*, Englewood Cliffs, NJ, Prentice Hall.

Annexe A : Tableaux

Notation efficace et préférences face au risque

Table 3: Propriétés de la notation avec neutralité face au risque ($\lambda = 1$)

Nombre de questions (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
Candidats qui omettent (%)	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

Modèle : pénalité notionnelle corrigeant pour les réponses au hasard pur ($\theta^* = -0.50$), $\hat{\theta}$: points efficaces pour les mauvaises sélections. $\hat{\gamma}$: points efficaces pour l'omission. $\hat{\gamma} - \hat{\theta}$ mesure les incitations à omettre. $100\left(\hat{\gamma} - \bar{s}(p)\right)$ est le biais d'omission avec $\bar{s}(p)$ la compétence moyenne des candidats qui omettent. RMSE : erreur quadratique moyenne. Pour $n = \infty$, $\hat{\gamma}$ est le point le plus élevé incitant tous les candidats à répondre (toute valeur inférieure serait également efficace).

Table 4: Propriétés de la notation avec aversion modérée aux pertes ($\lambda = 1.5$)

Nombre de questions (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
Candidats qui omettent (%)	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

Modèle : pénalité notionnelle corrigeant pour les réponses au hasard pur ($\theta^* = -0.50$), $\hat{\theta}$: points efficaces pour les mauvaises sélections. $\hat{\gamma}$: points efficaces pour l'omission. $\hat{\gamma} - \hat{\theta}$ mesure les incitations à omettre. $100\left(\hat{\gamma} - \bar{s}(p)\right)$ est le biais d'omission avec $\bar{s}(p)$ la compétence moyenne des candidats qui omettent. RMSE : erreur quadratique moyenne. Pour $n \geq 200$, $\hat{\gamma}$ est le point le plus élevé incitant tous les candidats à répondre (toute valeur inférieure serait également efficace).

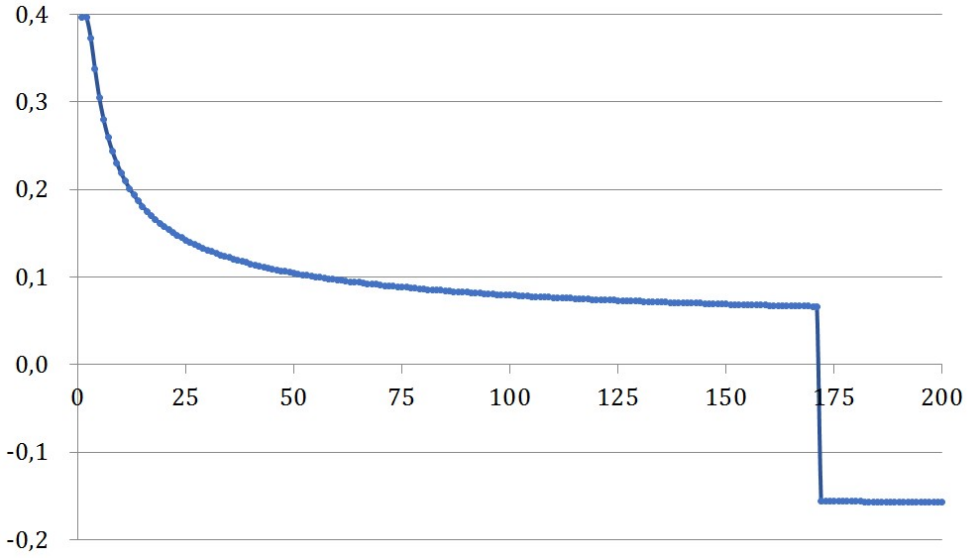
Table 5: Propriétés de la notation avec forte aversion aux pertes ($\lambda = 2.5$)

Nombre de questions (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
$100\,(\hat{\gamma} - \bar{s}(p))$	8.49	4.44	2.01	0.05	-1.27	0.00	0.00	0.00
Candidats qui omettent (%)	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

Modèle : pénalité notionnelle corrigeant pour les réponses au hasard pur ($\theta^* = -0.50$), $\hat{\theta}$: points efficaces pour les mauvaises sélections. $\hat{\gamma} - \hat{\theta}$ mesure les incitations à omettre. $100\,(\hat{\gamma} - \bar{s}(p))$ est le biais d'omission avec $\bar{s}(p)$ la compétence moyenne des candidats qui omettent. RMSE : erreur quadratique moyenne. Pour $n \geq 80$, $\hat{\gamma}$ est le point le plus élevé incitant tous les candidats à répondre (toute valeur inférieure serait également efficace).

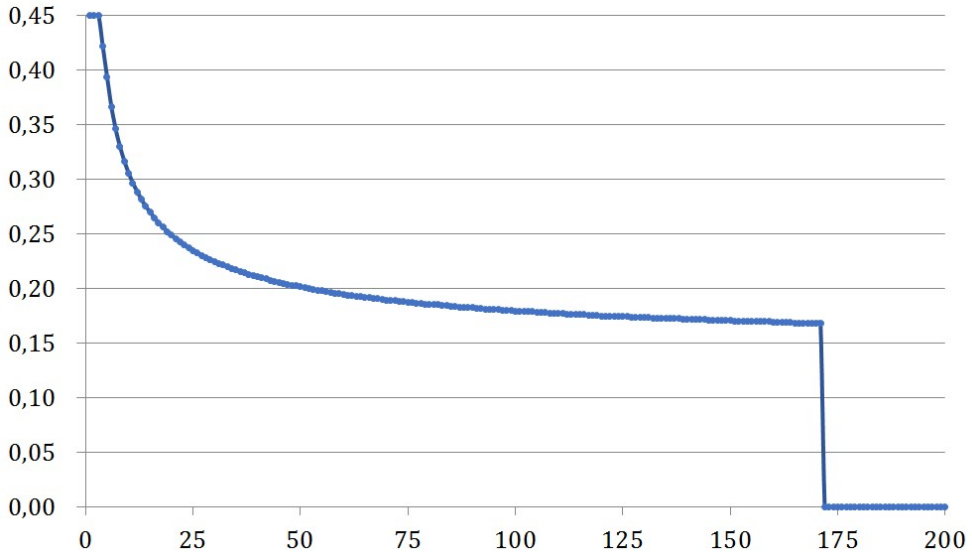
Annexe B : Graphiques

Figure 1: Points efficaces pour l'omission en fonction du nombre de questions



Modèle : points notionnels corrigeant pour les réponses au hasard pur ($\theta^* = -0.50$), aversion modérée aux pertes ($\lambda = 1.5$).

Figure 2: Proportion de candidats qui omettent en fonction du nombre de questions



Modèle : points notionnels corrigeant pour les réponses au hasard pur ($\theta^* = -0.50$), aversion modérée aux pertes ($\lambda = 1.5$).