Prise de risque et QCMs : conséquences sur la validité des méthodes de notation

Jean-Charles Quinton, Lisa Molto, Alan Chauvin

Prise de risque et QCMs : conséquences sur la validité des méthodes de notation

Jean-Charles Quinton [1] & Lisa Molto [2] & Alan Chauvin [2]

[1] *Laboratoire Jean Kuntzmann, UMR5224*
Université Grenoble Alpes / CNRS, France (quintonj@univ-grenoble-alpes.fr)

[2] *Laboratoire de Psychologie et NeuroCognition, UMR5105*
Université Grenoble Alpes / CNRS, France

Résumé. Cet article combine étude empirique (échantillon d'étudiants) et computationnelle (simulations de Monte Carlo) pour estimer l'impact des différences inter-individuelles dans la complétion de QCM (compréhension des probabilités et des méthodes de notation, prise de risque en répondant au hasard) sur la mesure de connaissance en examen (ici en statistique). Des solutions pour réduire les biais associés sont envisagées.

Mots-clés. QCM, notation, hasard, prise de risque

Abstract. This paper combines empirical (student sample) and computational (Monte Carlo simulations) studies to estimate the impact of inter-individual differences on QCM completion (understanding of probabilities and marking schemes, risk-taking when choosing random responses) on the measure of knowledge in exams (here in statistics). Solutions are proposed to reduce the associated biases.

Keywords. MCQ, marking, random responses, risk-taking

1 Introduction

L'enseignement de la statistique en sciences humaines et sociales (et plus spécifiquement en psychologie dans le contexte de cette étude) souffre d'une désaffection auprès des étudiants. Celle-ci est la conséquence de multiples facteurs, dont : 1) la non-anticipation de la place conséquente prise par le traitement de données et la statistique dans le cursus (quasiment une unité d'enseignement par semestre de la L1 au M2 en psychologie à Grenoble), 2) la crainte antérieurement construite vis-à-vis des mathématiques (population issue de filières non scientifiques pour la majorité), 3) l'impression que la matière joue un rôle de sélection plus que de formation (mauvaise perception de l'utilité de la matière pour les futurs métiers) [4]. Le traitement de données étant une matière obligatoire listée dans le référentiel des compétences en psychologie mais aussi dans les compétences transversales, la faible assiduité et participation d'une part non négligeable des étudiants accroissent les lacunes potentiellement accumulées depuis le collège et le lycée, et constituent ainsi une source d'échec en licence et en master.

Pour pallier ces difficultés, de nombreuses innovations pédagogiques sont mises en place : boîtiers de vote pour augmenter la participation, applications interactives, exercices en ligne et pédagogie inversée. Néanmoins, l'évaluation finale généralement présente reste un élément critique, qui conditionne en partie la motivation des étudiants. En effet, la compréhension et l'acceptation des méthodes et du contenu des évaluations facilite l'investissement dans la matière et par la même son acquisition. En début de cursus, et du fait d'effectifs importants (e.g. environ 600 étudiants en L1 à Grenoble), l'évaluation reste majoritairement basée sur des QCMs pour des raisons pratiques plus que pédagogiques. La stucturation et méthode de notation de tels QCMs est donc importante, afin d'éviter un abandon de la matière, de favoriser une acquisition sur le long terme de connaissances, et non des stratégies visant uniquement la performance à l'examen.

Or, les méthodes de notation de QCMs incluent parfois un terme de correction (calculé à partir de l'espérance de note obtenue en répondant au hasard), et sont sélectionnées sur la base de calculs probabilistes considérant les étudiants comme des agents économiques rationnels (rapport coût-bénéfice de la sélection d'une réponse) [2, 3]. Les barèmes de notation valorisent ainsi des dynamiques de réponse aux QCMs qui reposent sur des notions en cours d'apprentissage, et que l'on ne souhaite pas nécessairement évaluer. Lorsque l'évaluation porte sur des notions plus larges (e.g. statistiques descriptives et inférentielles), la note qui mène à la sélection des étudiants pourrait plus généralement dépendre de facteurs individuels non directement liés aux compétences statistiques. Dans cet article, on s'intéresse aux conséquences d'une mauvaise appréhension des règles de calcul de notes associée au niveau de prise de risque des étudiants sur la validité de la note, et en conséquence sur la validité de l'évaluation des compétences.

Dans la section suivante, on introduit l'ensemble des notations et méthodes de calcul considérées. On procède ensuite à deux études pour estimer l'impact de la variabilité inter-individuelle (en termes de compétence et de prise de risque) sur la validité psychométrique de la notation : étude empirique sur un échantillon d'étudiants en section 3, et étude computationnelle en section 4. Enfin, la section 5 explore les solutions envisageables au problème soulevé.

2 Notations

On introduit ici les notations, au sens mathématique dans un premier temps, puis au sens des méthodes de calcul de notes sur les QCMs dans un second temps. Dans l'étude empirique comme computationnelle, et afin d'avoir une mesure indépendante de la structure du QCM ou de la méthode de notation employée, on inclut des questions ouvertes dans l'examen final. Celles-ci sont en théorie sensées mieux refléter le score vrai (compétence réelle de l'étudiant en statistique), avec des contenus reposant directement sur des méthodes acquises en cours d'année. Surtout, et contrairement à ce qui est possible sur un QCM, elles laissent une place négligeable au hasard et à l'apprentissage par coeur.

Au niveau de la notation de ces questions, des critères indépendants permettent de valider les compétences des étudiants, même si des erreurs ont été commises sur des calculs intermédiaires. On notera O_i la note obtenue sur l'ensemble des questions ouvertes par l'étudiant i, ramenée à un taux de bonnes réponses dans $[0, 1]$.

On considère un QCM constitué de questions à réponses potentiellement multiples. Soit $R_{ijk} \in \{0, 1\}$ l'indication que l'étudiant i a coché la réponse k à la question j. Le fait que cette réponse soit correcte ou non est noté $C_{jk} \in \{0, 1\}$. Soit $M_{ij} = f_m(R_{ij}, C_j)$ la note obtenue à la question j par l'étudiant i, dépendante de la méthode de notation retenue (indicée m) :

1. *Tout ou rien* : $f_1(...) = 1$ si $\forall k, R_{ijk} = C_{jk}$, sinon 0. L'étudiant obtient ainsi le maximum de point (1 / question) si il a choisi l'intégralité des bonnes réponses, et n'obtient aucun point sinon.
2. *Pénalité avec report* : $f_2(...) = ({}^tR_{ij}(2C_j - 1))/||C_j||_1$. La transformation affine de C_j génère une pénalité pour chaque mauvaise réponse cochée, ici opposée au nombre de points gagnés pour une bonne réponse. Même si le score peut donc être négatif, la normalisation par le nombre de réponses correctes garantit que le score maximal sur chaque question soit de 1 (ainsi comparable entre méthodes).
3. *Pénalité sans report* : $f_3 = f_2^+$. En ne gardant que la partie positive de la fonction précédente, et donc une note minimale de 0 sur chaque question, on garantit que les points perdus à une question n'impacteront pas les suivantes.

Le total des points obtenu sur l'ensemble des questions est défini par $S_i = g_c(M_i)$:

- *Sans correction du hasard* : $g_0(...) = sum_j(M_{ij})$ (i.e. somme des points obtenus à chaque question).
- *Avec correction du hasard* : $g_1(...) = g_0(...) - \mathbb{E}(g_0)$. On déduit de la somme précédente l'espérance des points obtenus en répondant au hasard pour la méthode m choisie (calculée analytiquement ou empiriquement sur la base de f_m).

Enfin, la somme S_i est rectifiée pour garantir une note finale N_i dans l'intervalle $[0, n_q]$, où n_q est le nombre de questions du QCM (i.e. toute note hors de cet intervalle est ramenée à ses bornes).

L'intitulé des questions indiquant la présence certaine d'une réponse correcte dans la liste proposée (e.g., "choisir la réponse correcte... parmi les suivantes") ou cette liste incluant une réponse par défaut (du type "Aucune de ces réponses n'est correcte", telle qu'intégrée au package *automultiplechoice* pour LaTeX et Moodle [1]), on peut détecter les questions non traitées. On note le fait qu'une question soit traitée par $T_{ij} = 1$ si $||R_{ij}||_1 > 0$, 0 sinon. Le taux de réponse est alors défini par $P_i = \sum_j T_{ij}/n_q$.

La méthode de notation la plus simple et permettant de maximiser la participation des étudiants est évidemment d'allouer des points à toute réponse correcte, sans pénaliser les mauvaises réponses. Néanmoins, cette solution n'est pas retenue du fait que pour

	0	**1**	**2**	**3**
0	0	0	1/15	0
1	0	1/5	0	0
2	1	0	0	0

(1) Tout ou rien

	0	**1**	**2**	**3**
0	0	-1/6	-1/3	-1/2
1	1/2	1/5	-0.1	-2/5
2	1	1/2	0	-1/2

(2) Pénalité avec report

	0	**1**	**2**	**3**
0	0	1/6	1/15	1/10
1	1/2	1/5	1/5	0
2	1	1/2	0	0

(3) Pénalité sans report

FIGURE 1: Note espérée pour chaque méthode de notation de QCM, pour une question comportant 6 réponses possibles, dont 2 correctes. Chaque ligne correspond à un nombre de réponses connues, chaque colonne à un nombre de réponses fournies au hasard.

obtenir la note maximale, il suffit de cocher toutes les réponses possibles. Cette stratégie reste pourtant minoritaire pour cette méthode de notation, preuve que le calcul rationnel de coût-bénéfice n'est pas nécessairement réalisé par les étudiants. Pour les méthodes présentées, les corrections du hasard introduites se basent elles-aussi sur la supposition que les étudiants cocheront des réponses au hasard si cela peut statistiquement leur permettre d'accumuler des points.

Bien sûr, la probabilité d'obtenir des points à une question en répondant au hasard dépend du nombre de réponses disponibles et du nombre de réponses correctes. Néanmoins, pour des paramètres de QCM donnés, la notation influe également sur la distribution. Les tables de la figure 1 indiquent le nombre de points espérés pour une question (1 point au maximum) selon le nombre de réponses connues (garanties de rapporter des points) et celles cochées au hasard, lorsque 2 réponses sont correctes parmi 6. En pratique, le nombre de réponses à cocher est contraint par l'intitulé de la question (e.g. "choisir la meilleure représentation...", ou "indiquer comment l'écart type et la moyenne évoluent..." avec des réponses partiellement exclusives). Dans la suite de l'article et pour simplifier, on considèrera donc que les étudiants peuvent déterminer le nombre de réponses à cocher (ce paramètre ne conditionnant pas les résultats présentés). On remarque en effet que quelle que soit la configuration, la méthode de *pénalité sans report* est celle qui rapporte le plus de points. De plus, cette méthode est celle qui devrait encourager les étudiants à répondre à toutes les questions ; en effet, la méthode *pénalité avec report* pénalise un étudiant sur la compétence échouée mais également sur les compétences évaluées par d'autres questions, et la méthode *tout ou rien* ne valorise aucune compétence partiellement maîtrisée.

3 Etude empirique

Procédure : Afin de maximiser le taux de réponses, la méthode de notation *pénalité sans report* et *sans correction du hasard* a été choisie. Les étudiants de L1 de psychologie ont eu accès au barème de QCM durant tout le semestre (en ligne, et présentation dès le début de l'UE). Le dernier CM et dernier TD ont servi d'entraînement à l'examen (avec

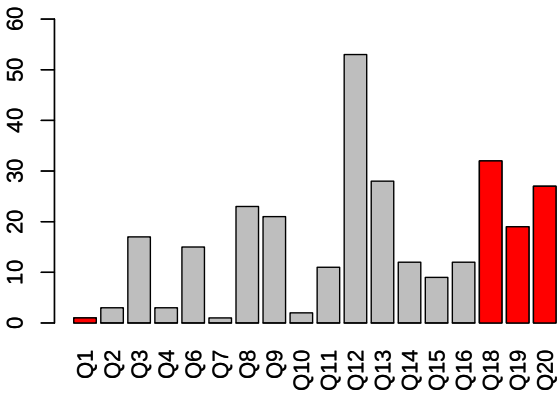

FIGURE 2: Nombre de réponses manquantes par question (sur les étudiants ayant intégralement répondu aux questions ouvertes). En rouge, celles portant sur les mêmes compétences que les questions ouvertes (Q5 & Q17).

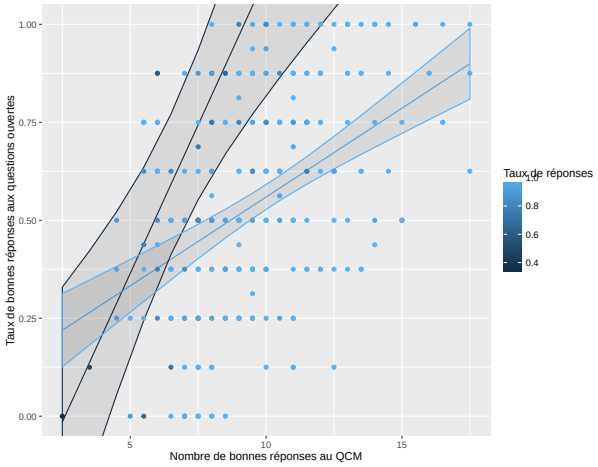

FIGURE 3: Graphe d'interaction du lien entre score au QCM et questions ouvertes, modéré par le taux de réponses (méthode 3, sans correction du hasard). Les droites et bandes de confiance correspondent aux taux de réponses extrêmes des données.

QCM et questions ouvertes, correction croisée par les étudiants, et explicitation de la notation), et ces séances ont permis de confirmer que les calculs de coût-bénéfice étaient non intuitifs pour la majorité. Les étudiants ont ensuite réalisé l'examen terminal (18 questions de QCM, et 2 questions ouvertes). Les échantillons des 4 sessions réparties sur les années universitaires 2015/16 et 2016/17 n'étant pas indépendants, on considère ici exclusivement les données de l'examen de 1ère session de 2016/17 ($n = 439$). En effet, les données étant anonymes, il est impossible de dissocier les étudiants des sessions 1 et 2 (compensation avec d'autres matières), ni de ceux redoublants. Néanmoins, les résultats présentés sont cohérents avec les 3 autres sessions ($N = 1100$).

Analyse : Afin de pouvoir correctement estimer la compétence des étudiants, seuls ceux ayant répondu aux questions ouvertes sont conservés ($n = 305$). Malgré ce filtrage, ainsi que les consignes et entraînements réalisés en cours de semestre, de nombreux étudiants n'ont pas intégralement répondu au QCM (40%). Ceux-ci ayant complété les questions ouvertes, il ne s'agit probablement pas d'un manque de motivation ou de temps, comme confirmé par le diagramme en barre (voir figure 2, respectant l'ordre des questions). De même, il pourrait s'agir d'étudiants qui n'ont révisé que la partie du cours associée aux questions ouvertes, mais celles-ci pouvaient porter sur l'intégralité des TDs, et ce n'est pas traduit par la distribution (en rouge sur la figure, avec un taux d'abstention moyen de 6.5% sur celles liées aux questions ouvertes vs. 4.9% sur les autres).

Quelle que soit la méthode de notation employée (1 à 3, avec/sans correction du hasard), la corrélation entre QCM et questions ouvertes est comprise dans l'intervalle $[0.43, 0.47]$ ($p < 10^{-14}$). Afin de voir si le lien est modéré par le taux de réponse (et donc

potentiellement la prise de risque), le modèle linéaire suivant a été testé (sur les variables P et N centrées) : $O_i = \beta_0 + \beta_1 N_i + \beta_2 P_i + \beta_3 N_i \times P_i + \epsilon_i$. L'effet principal de la note au QCM (N) reste systématiquement significatif ($b \in [0.13, 0.14]$, $t(301) > 8.3$, $p < 10^{-14}$), alors que l'effet principal du taux de réponse (P) est négatif, et uniquement significatif pour les méthodes 1 et 3 ($b \in [-0.08, -0.06]$, $t(301) < -3.0$, $p < 0.003$). Ce résultat est cohérent avec le non report de pénalité pour ces méthodes, et le fait que le remplissage du QCM puisse être réalisé au hasard pour une part des étudiants. L'interaction $N \times P$ est de plus significative pour toutes les méthodes ($b \in [-0.05, -0.03]$, $t(301) < -2.5$, $p \leq 0.01$), et illustrée sur la figure 3. Globalement, les étudiants ayant un faible taux de réponse obtiennent en moyenne des notes aux questions ouvertes plus élevées pour une note de QCM donnée. Dit autrement, plus le taux de réponse augmente, moins la réussite aux questions ouvertes s'explique par les réponses au QCM, et moins le QCM évalue les compétences réelles. Sur l'ensemble de l'examen, cela signifie également que les étudiants prenant moins de risque s'auto-pénalisent, à minima pour les méthodes 1 et 3. Si on considère toujours la note obtenue aux questions ouvertes comme référence, la validité de la notation de QCM peut être décomposée en faisant varier le taux de réponse dans l'intervalle observé. On obtient alors une corrélation r_{NO} décroissante de 0.99 pour $P \in [0.33, 0.5]$ ($p = 0.02$) à 0.42 pour $P \in [0.95, 1]$ ($p < 10^{-9}$).

4 Simulation

Procédure : Pour intégrer la non-réalisation de calculs de coût-bénéfice complexes, on ajoute le niveau de prise de risque individuel comme un facteur psychologique pouvant expliquer les observations. A partir du score vrai (la compétence en statistique pour l'unité d'enseignement considérée), notée V, on considère que la mesure de compétence (les réponses correctes fournies) et la compétence réelle ont une corrélation de r_{VO} (via questions ouvertes) $> r_{VR}$ (via QCM). On tire la compétence V et la prise de risque D selon des distributions uniformes (ou bien normale pour la compétence, et beta pour la prise de risque afin de fitter les données réelles), éventuellement négativement corrélées. En effet, on a d'autant moins à perdre en répondant au hasard que l'on sait peu répondre aux questions (selon la méthode de notation appliquée). Les individus répondent correctement aux questions 1 à $\tilde{V}$ (avec $\tilde{V} < V$ satisfaisant r_{VR}), aléatoirement aux A questions suivantes (avec $A = (n_q - \tilde{V}) * D$, i.e. une proportion de questions parmi celles non maîtrisées égale au niveau de prise de risque), et rien aux questions restantes. On reproduit ainsi la dynamique probable des étudiants (moins on maîtrise, plus on laisse de place au hasard) expliquant la baisse de la corrélation r_{NO}. On tire aléatoirement O en respectant r_{VO}. Afin de confirmer que les observations peuvent être expliquées par la combinaison des différences inter-individuelles de compétence et de prise de risque, on réalise des simulations de Monte Carlo avec des échantillons de 1000 étudiants simulés, en faisant varier les distributions des facteurs inter-individuels, les corrélations, et en appliquant les 3×2

méthodes de notation précemment introduites.

Analyse : Trivialement, on vérifie que la corrélation entre taux de réponse et la note obtenue au QCM augmente avec la compétence V. On effectue les mêmes analyses que dans l'étude empirique, mais en testant aussi directement la corrélation r_{NV} en plus de r_{NO} pour vérifier la validité de la notation. On reproduit dans la majorité des cas les résultats obtenus sur l'échantillon humain, dont les effets principaux de la note N et du taux de réponse P, ainsi que l'interaction $N \times P$. Néanmoins, l'effet du taux de réponse diminue lorsque le nombre de questions, le nombre de réponses possibles, et le nombre de réponses correctes augmentent, car associés à une diminution de la probabilité d'obtenir beaucoup de points au hasard.

5 Conclusion

Dans note étude empirique, et étant donné les consignes, la préparation aux examens et le barème de QCM, le comportement optimal est de répondre à toutes les questions. Or 40% des étudiants produisent à la place des comportements non rationnels, déjà décrits dans [3], qui peuvent être attribués à un déficit sur des compétences en probabilité et statistique elles-mêmes en cours d'apprentissage. Ces comportements non optimaux induisent un biais dans l'évaluation des "compétence vraies" par les méthodes classiques de notation de QCM, pénalisant les étudiants avec une aversion au risque, et baissant ainsi la validité de la note comme mesure de compétence. En plus du barème annoncé, les simulations réalisées confirment le possible impact 1) des profils vis à vis de la prise de risque et 2) des compétences statistiques des apprenants.

La méthode de correction employée, avec ou sans correction du hasard, ne semble pas changer drastiquement les résultats, mais peut influencer la motivation des étudiants et leur attitude lors de l'examen, et ainsi permettre la minimisation du taux d'abstention. Les autres solutions au problème soulevé incluent bien sûr l'usage plus massif de questions ouvertes, moins sujettes au hasard, mais coûteuses en correction pour de large effectifs, et plus sensibles aux différences inter-individuelles d'interprétation si mal posées (augmentant par la même occasion l'erreur de mesure dans la correction). Alternativement, augmenter le nombre de réponses correctes et de questions de QCM permettrait de baisser le nombre moyen de points gagnés par case cochée (sur un total de points fixe à l'examen). En complément, rendre non-prédictible le nombre de réponses correctes, et augmenter le nombre de réponses possibles baisserait potentiellement la variance des points obtenus au hasard. On pourrait ainsi baisser la part de variance expliquée par le hasard, et réduire ainsi le biais d'estimation de la compétence dû aux différences inter-individuelles de compréhension des méthodes de notation et de prise de risque.

Remerciements

Nous remercions les intervenants de l'UE Traitement de Données de L1 du département de psychologie, au sein de l'UFR SHS de l'Université Grenoble Alpes pour leur participation active à la mise en place des TDs de révision, à la constitution des QCMs, et à la correction des examens ayant abouti aux jeux de données utilisés dans cet article.

Références

[1] Alexis Bienvenüe. Auto Multiple Choice (AMC). `http://auto-multiple-choice.net`.

[2] María Paz Espinosa and Javier Gardeazabal. Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, 54(5) :415–425, 2010.

[3] Daniel Kahneman and Amos Tversky. Prospect theory : An analysis of decision under risk. *Econometrica : Journal of the econometric society*, pages 263–291, 1979.

[4] Annique Smeding, Céline Darnon, Carine Souchal, Marie-Christine Toczek-Capelle, and Fabrizio Butera. Reducing the socio-economic status achievement gap at university by promoting mastery-oriented assessment. *PLoS One*, 8(8) :e71678, 2013.