

Pour citer cet article : Braibant, J.-M., Gerard, F.-M., & Billat, E. (2023, prépublication). QCM à points négatifs au début de l'enseignement universitaire : échecs abusifs et biais de genre. *La Revue L'EeE, Varia*. <https://revue.leee.online/index.php/info/article/view/188>

QCM À POINTS NÉGATIFS AU DÉBUT DE L'ENSEIGNEMENT UNIVERSITAIRE : ÉCHECS ABUSIFS ET BIAIS DE GENRE

Jean-Marc BRAIBANT, François-Marie GERARD, Étienne BILLAT

*Version de la prépublication : juin 2023
Évaluation ouverte et collaborative*

Résumé

Les QCM à points négatifs sont souvent utilisés pour réduire au mieux l'effet du hasard dans les réponses données. Pour des raisons fortuites, le même examen a été présenté à deux groupes d'étudiant·es en annonçant deux modalités de correction différentes : l'une sans pénalités, l'autre avec pénalités. Alors que l'équivalence entre les deux groupes a été attestée, les échecs sont plus nombreux dans le groupe « avec pénalités ». L'analyse de ces échecs abusifs montre qu'ils concernent davantage les étudiantes que les étudiants.

Mots-clés : QCM, questionnaire à choix multiple, points négatifs, pénalité, biais de genre

Abstract

Negative point MCQs are often used to minimize the effect of chance in the answers given. Due to fortuitous reasons, the same exam was presented to two groups of students with two different marking schemes announced: one without penalties, the other with penalties. While equivalence between the two groups was attested, there were more failures in the "with penalties" group. The analysis of these abusive failures shows that they concern more female students than male students.

Keywords: MCQ, multiple choice, negative points, penalty, gender bias

1. Introduction

Depuis de nombreuses années, l'effet du hasard sur les résultats des étudiant·es dans les examens par questions à choix multiple (QCM) a provoqué de nombreuses discussions. Selon certains auteurs, un candidat qui ne connaît pas la réponse à une question a une probabilité non négligeable de la sélectionner en la devinant. Selon d'autres, le hasard n'aurait aucune influence significative sur les résultats si l'examen est bien conçu (un nombre suffisant d'items et des leurres pertinents). Où se situe la vérité ? Et surtout quelle conduite faut-il adopter devant ce problème dont les enjeux sont importants vu leur impact significatif sur la réussite et l'avenir des étudiant·es ?

Grosso modo, les trois solutions apportées les plus fréquentes sont les suivantes :

1. corriger l'effet du hasard en pénalisant les réponses incorrectes (système à points négatifs) ;
2. corriger l'effet du hasard en utilisant un seuil de réussite supérieur à 50 % de bonnes réponses, seuil qui varie en fonction du nombre d'options de réponse proposées (c'est ce qu'on appelle le *standard setting*) ;
3. maintenir le seuil habituel de 50 % de bonnes réponses et ne pas corriger l'effet du hasard, mais augmenter la longueur et la qualité des questions pour assurer la fiabilité de l'examen.

Une revue de la littérature relative à ces différentes modalités présente des arguments en faveur et en défaveur de chacune de ces trois propositions. Le débat s'est concentré sur la méthode la plus couramment utilisée, la notation avec ou sans points négatifs, mais les conclusions de ces travaux ont présenté des résultats contradictoires. Le manque de consensus s'explique notamment par le fait que ces études ont négligé les différences de paramétrage des tests et/ou ont utilisé des critères d'analyse différents s'appliquant par ailleurs à des publics présentant des profils académiques diversifiés (Lesage et al., 2013).

À notre connaissance, il n'existe aucune étude permettant de comparer les performances d'un même échantillon d'étudiant·es qui seraient réparti·es aléatoirement en deux sous-groupes (examen avec ou sans notation négative) dans une situation réelle de certification débouchant sur une note officielle. Pour des raisons d'équité faciles à comprendre, ce type de design expérimental ne peut être mis en œuvre.

Les circonstances nous ont donné l'occasion de mener ce type d'analyse suite à une erreur de communication qui s'est produite à l'UCLouvain lors de la session d'examen de janvier 2020.

Ce jour-là, les étudiant·es réparti·es dans deux amphithéâtres différents pour l'évaluation certificative d'un même cours ont reçu au moment de la distribution des feuilles d'examen des consignes différentes à propos des modalités d'attribution des points.

Dans la situation A, on a informé les étudiant·es qu'ils·elles obtiendraient 1 point en cas de bonne réponse et 0 point en cas d'abstention ou de mauvaise réponse. On n'attribuait donc pas de points négatifs. C'est la situation sans pénalités.

Dans la situation B, on a annoncé aux étudiant·es que le barème suivant serait appliqué : 1 point pour une bonne réponse, 0 point en cas d'abstention et une pénalité de 0,25 point pour chaque mauvaise réponse. C'est la situation avec pénalités.

Les données issues de ces deux situations d'examen sont au centre de notre étude. Précisons qu'après l'examen, les étudiant·es placé·es dans la situation B ont déposé un recours



qu'ils-elles ont gagné. Ils-elles ont pu repasser l'examen selon les modalités initialement prévues, celles de la situation A (sans pénalités).

2. Population de départ et procédure d'échantillonnage

Ces circonstances particulières ayant débouché sur un recours, les informations que nous pouvons communiquer ici sont limitées pour des raisons évidentes de confidentialité et par respect du RGPD¹. Il s'agissait d'un cours obligatoire comptant pour quatre crédits sur 60 entrant en compte dans le tronc commun d'un programme de bachelier de 1^{re} année universitaire. L'examen s'est déroulé lors de la session de janvier 2020 pour un total de 664 étudiant-es. Ce nombre très élevé explique pourquoi les étudiant-es ont été réparti-es dans deux amphithéâtres pour passer l'examen. L'échantillon que nous étudierons ici se limitera à 60 % de cette population de départ : les 399 étudiant-es s'étant inscrit-es pour la première fois dans la faculté de référence en septembre 2019. Nous n'avons donc pas pris en compte les nombreux-ses étudiant-es redoublant-es qui repassaient ce cours pour la 2^e, voire la 3^e fois suite à leur(s) échec(s) antérieur(s). Nous voulions nous focaliser sur les étudiant-es qui passaient cet examen pour la première fois de manière à avoir un public novice face à cette situation. En effet, les stratégies de réponses à un examen par QCM à points négatifs sont impactées par l'expérience acquise d'une session d'examen à l'autre (Brogniaux, 2021 ; Cipriani, 2018). Nous n'avons pas non plus pris en compte une centaine d'étudiant-es en provenance d'autres orientations d'études qui avaient choisi de prendre ce cours comme une option à leur programme.

3. Étude de l'équivalence des deux échantillons

Le jour de l'examen, les étudiant-es ont été réparti-es vers l'un ou l'autre des deux amphithéâtres selon la première lettre de leur nom de famille. Cette procédure aléatoire simple constitue une technique habituelle dans le modèle expérimental pour s'assurer de l'équivalence de deux échantillons indépendants au départ d'une même population. Les 208 étudiant-es dont l'initiale commence par A jusqu'à la lettre J ont passé l'examen dans la condition A sans points négatifs. Les 191 étudiant-es dont l'initiale du nom commence par K jusqu'à la lettre Z ont passé l'examen en situation B avec points négatifs. Nous avons pu vérifier l'équivalence des deux échantillons en fonction de différentes informations.

3.1 Les variables d'entrée

3.1.1 Âge

Les deux groupes d'étudiant-es avaient un âge moyen de 18 ans et demi au moment de leur inscription en 1^{re} année en septembre 2019. Le test Khi carré (valeur 7,851 ; $p=,549$) confirme que les deux échantillons ne sont pas différents sur ce critère.

¹ Règlement général sur la protection des données



Tableau 1*Âge moyen des étudiant·es dans les deux amphithéâtres*

	Auditoire A (Sans pénalités)	Auditoire B (Avec pénalités)
Âge moyen lors de l'inscription	18,5 ans	18,5 ans

3.1.2 Expérience à l'université

En ce qui concerne le passé universitaire des étudiant·es, la proportion relative des étudiant·es entrant pour la première fois en 1^{re} année universitaire par rapport à ceux·celles qui s'étaient déjà inscrit·es précédemment dans une autre faculté ou dans une autre université est identique dans les deux groupes. Le test Khi-Carré (valeur=,000 ; $p=,987$) confirme qu'il n'y a pas de distinction entre les deux échantillons en ce qui concerne leur expérience à l'université.

Tableau 2*Expérience universitaire des étudiant·es des deux amphithéâtres*

	Auditoire A (Sans pénalités)		Auditoire B (Avec pénalités)	
	Fréquence	Pourcentage	Fréquence	Pourcentage
1 ^{re} génération universitaire	195	94,2	180	94,2
Inscription préalable dans l'enseignement universitaire	12	5,8	11	5,8
Total	207	100	191	100

3.1.3 Type d'enseignement suivi dans l'enseignement secondaire

En Belgique francophone, tou·tes les étudiant·es disposant d'un Certificat d'Enseignement Secondaire Supérieur (CESS) peuvent s'inscrire à l'université dans la faculté de leur choix quels que soient leurs résultats obtenus au terme de ces études ou la filière suivie précédemment dans l'enseignement secondaire. La probabilité de réussir des études dans l'enseignement supérieur est cependant plus élevée pour les élèves issu·es de l'enseignement général de transition que pour ceux·celles qui ont suivi les filières techniques de transition ou de qualification (ARES, 2022). Les proportions relatives d'étudiant·es issu·es de chacune de ces filières sont pratiquement les mêmes dans les deux amphithéâtres. Le test Khi carré (valeur=1,361 ; $p=,506$) confirme que les deux échantillons ne sont pas significativement différents sur ce critère.

Tableau 3*Type d'enseignement secondaire suivi par les étudiant·es des deux amphithéâtres*

	Auditoire A (Sans pénalités)		Auditoire B (Avec pénalités)	
	Fréquence	Pourcentage	Fréquence	Pourcentage
Baccalauréat étranger	11	5,3	13	6,9
Technique de qualification	32	15,5	29	15,3
Technique de transition	26	12,6	22	11,6
Général de transition	137	66,5	125	66,1
Total	206	100	189	100



3.1.4 Genre

En Fédération Wallonie-Bruxelles, comme dans tous les pays de l'OCDE, la réussite est influencée par le genre (OCDE, 2022 ; Todd, 2022) : les femmes sont significativement plus nombreuses à obtenir un diplôme de l'enseignement supérieur que les hommes. Dans notre échantillon, la proportion des femmes est légèrement supérieure dans la situation avec points négatifs, mais cette différence n'est pas significative au seuil de ,05 (Khi carré=,66 ; $p=,413$).

Tableau 4

Répartition des hommes et des femmes dans les deux amphithéâtres

	Auditoire A (Sans pénalités)		Auditoire B (Avec pénalités)	
	Fréquence	Pourcentage	Fréquence	Pourcentage
Femmes	155	74,5	149	78,0
Hommes	53	25,5	42	22,0
Total	208	100	191	100

L'hypothèse d'équivalence des deux échantillons est donc rencontrée en ce qui concerne les quatre critères pris en compte : l'âge, le passé universitaire, le parcours dans l'enseignement secondaire et le genre. Nous pouvons considérer qu'aucun des deux groupes n'a un avantage sur l'autre avant de commencer l'examen qui pourrait expliquer d'éventuelles différences de résultats entre eux.

3.2 Les variables de sortie

À ces variables d'entrée, mesurées en début de formation, nous avons ajouté deux indicateurs dits de sortie : le nombre de crédits validés par l'étudiant-e ainsi que la moyenne des points obtenus en fin d'année afin de vérifier l'hypothèse que les deux groupes d'étudiant-es n'étaient pas différents non plus entre eux en ce qui concerne leur aptitude à mener à terme une première année d'études universitaires. C'est l'occasion de souligner que, par rapport à notre échantillon de départ constitué des étudiant-es ayant passé l'examen en janvier, quelques-unes d'entre eux-elles ont arrêté leurs études par la suite : six dans le groupe sans pénalités ($208-6=202$) et deux dans le groupe avec pénalités ($191-2=189$).

Tableau 5

Résultats obtenus en fin de 1^{re} année par les étudiant-es des deux amphithéâtres

	Auditoire A (Sans pénalités)			Auditoire B (Avec pénalités)		
	N	Moyenne	Écart-type	N	Moyenne	Écart-type
Nombre de crédits validés en fin d'année (sur 60)	202	44,3	18,0	189	45,7	17,5
Moyenne des points obtenus (sur 100)	202	54,4	17,9	189	56,4	16,2

Les tests Khi Carré réalisés sur chacun de ces deux indicateurs montrent qu'il n'y pas de différence significative entre les deux groupes (Khi Carré Nombre de crédits=54,421 ; $p=,458$; Khi Carré Moyenne des points=282,547 ; $p=,643$).



3.3 Conclusion concernant l'équivalence des deux échantillons

Quelles que soient les variables prises en compte en entrée ou en sortie de cette première année universitaire, nous avons pu vérifier l'hypothèse d'absence de différences significatives entre les deux échantillons. Nous pouvons dès lors conclure à leur équivalence.

4. Présentation de l'examen

L'examen concerné est constitué d'un questionnaire de 24 questions QCM présentées sur format papier. Chaque question propose 4 options de réponses et une seule d'entre elles est correcte. Le score maximum est de 24 points et le seuil de réussite fixé par l'enseignant correspond à 50 % de bonnes réponses, soit à un score de 12 points sur 24. Les étudiant·es des deux groupes ont reçu les mêmes questions, mais les barèmes d'attribution de la note étaient par contre différents :

- Dans la situation A (examen sans pénalités), chaque bonne réponse valait 1 point. Chaque mauvaise réponse ou abstention valait 0 point.
- Dans la situation B (examen avec pénalités), les étudiant·es ont été informé·es au moment de l'examen que leur score serait calculé en utilisant le barème suivant :
 - 1 point pour chaque bonne réponse
 - 0 point en cas d'abstention
 - Le retrait de 0,25 point pour chaque mauvaise réponse.

L'examen a été administré au début de la session de janvier 2020. Pour la grande majorité des étudiant·es, il s'agissait non seulement de leur première session d'examen, mais aussi de leur première évaluation certificative faisant appel à un QCM.

5. Qualités psychométriques des deux évaluations

Avant de commencer la présentation et l'analyse des résultats obtenus par les étudiant·es dans les deux conditions, il est important de s'arrêter sur les qualités psychométriques du questionnaire. Les indicateurs les plus utilisés à cette fin sont l'indice de fidélité, l'indice de discrimination, l'indice de difficulté et l'erreur standard de mesure (Laveault & Grégoire, 2014).

Tableau 6

Indicateurs psychométriques pour les deux évaluations

Indicateurs	Questionnaire A (Sans pénalités)	Questionnaire B (Avec pénalités)
Indicateur de fidélité (alpha de Cronbach)	0,71	0,70
Indicateur de discrimination (Rir)	0,24	0,23
Indicateur de difficulté (p)	0,48	0,40
Erreur standard de mesure	2,22	2,21

Dans les deux situations, le questionnaire atteint la valeur minimale acceptable attendue pour l'alpha de Cronbach qui est de 0,70 et il dépasse la valeur minimale attendue pour un indice de discrimination qui est généralement fixée à 0,20. Cela signifie que les questions posées à



cet examen permettent d'estimer de manière globalement fiable le niveau de maîtrise de la matière des étudiant·es et qu'elles donnent également une information appropriée pour distinguer ceux·celles qui la maîtrisent et ceux·celles qui ne la maîtrisent pas. Les valeurs de ces deux indices ainsi que l'erreur standard de mesure sont les mêmes (à un centième près) : ce questionnaire appliqué à deux publics différents présente des qualités psychométriques similaires.

L'indice de difficulté de l'examen – obtenu en faisant la moyenne des indices de difficulté de toutes les questions – est le seul indice qui varie d'une situation à l'autre. Dans le cas d'items dichotomiques, cet indice est donné par la proportion des répondant·es qui ont sélectionné la bonne réponse (0,73 signifie que 73 % des participant·es ont sélectionné la bonne réponse à cette question). Plus l'indice est élevé, plus l'examen est considéré comme « facile ». Concrètement, toutes choses étant égales par ailleurs, l'examen pour lequel on a annoncé aux étudiant·es que les mauvaises réponses seront pénalisées s'avère plus exigeant que le même examen pour lequel on ne sanctionne pas les mauvaises réponses ($p=.48$ vs $p=.40$). Il est important de souligner que dans les deux cas, la valeur p observée dans les deux échantillons est proche d'une moyenne de 0,50 ou du moins suffisamment proche pour éliminer l'hypothèse d'avoir un effet plafond ou un effet plancher au niveau de la variance des résultats. S'agissant d'une épreuve utilisée à des fins de certification (voire de sélection), dans un contexte où la discrimination des aptitudes des étudiant·es prime, il est généralement recommandé d'avoir autant que possible un p -index moyen proche de .50, ce qui est le cas. En d'autres termes, aucun des deux examens n'est ni trop facile ni trop difficile. Ces résultats ainsi que les autres indicateurs témoignent que l'examen est bien calibré même si le nombre d'items pourrait être plus élevé (Braibant, 2015).

6. Résultats obtenus par les étudiant·es dans les deux examens

6.1 Statistiques descriptives

Les distributions des notes sont très différentes dans les deux conditions. Dans le groupe sans pénalités, la moyenne (11,46 pts) est proche de la médiane (11 pts) ainsi que du seuil de réussite (12 pts), ce qui n'est pas le cas dans le groupe avec pénalités. La moyenne des résultats ainsi que la médiane y sont beaucoup plus bas et on constate un nombre réduit de scores supérieurs au seuil de réussite (cf. figure 1). La majorité des étudiant·es sont en échec dans cette situation.

Figure 1

Distributions de fréquences des scores dans les deux conditions

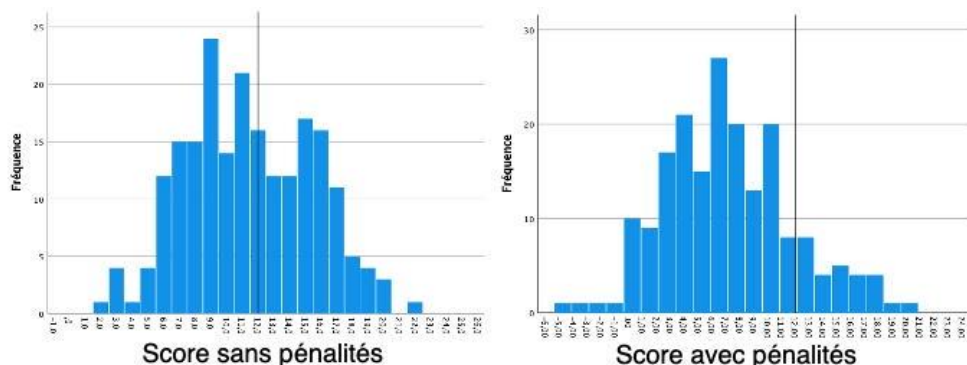
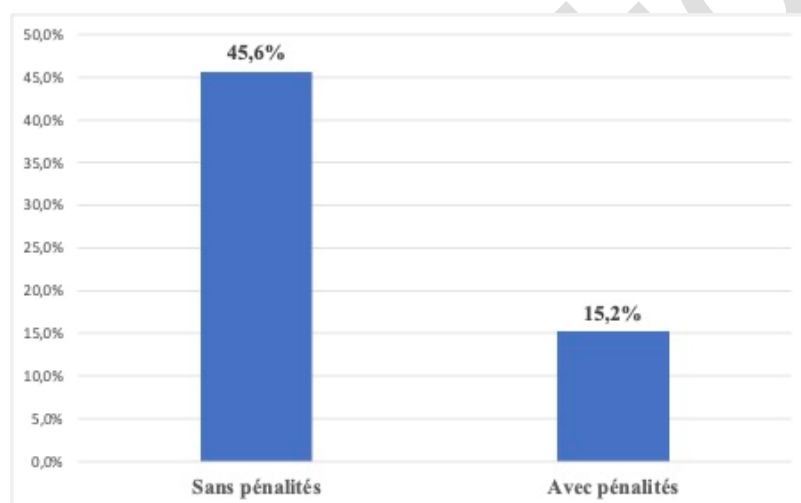


Tableau 7*Statistiques descriptives pour les deux examens*

	<i>N</i>	<i>Médiane</i>	<i>Moyenne</i>	<i>Écart-type</i>	<i>Min</i>	<i>Max</i>
Auditoire A (Sans pénalités)	208	11,00	11,46	4,04	2,00	22,00
Auditoire B (Avec pénalités)	191	6,75	7,22	4,58	-4,75	20,00

Une comparaison des moyennes à l'aide du test t de Student montre que les résultats obtenus par les deux groupes sont significativement différents ($F=,995$; $\text{Sigma}=,319$, p bilatéral $<,001$). Les notes obtenues au même examen varient fortement suivant que l'on fait appel ou non à un système de points négatifs. Si, toutes choses étant égales par ailleurs, on peut s'attendre à ce que les notes soient globalement diminuées lorsqu'on utilise des pénalités en cas de mauvaise réponse par rapport à la situation où on ne retire pas de points, ce qui apparaît ici d'emblée est l'impact considérable que peut avoir le choix de l'un ou l'autre barème de notation sur le niveau des performances obtenus sur deux groupes équivalents. Les étudiant·es avaient a priori les mêmes chances de réussir cet examen. Or, le pourcentage de réussite dans la condition sans pénalités est trois fois plus élevé que dans le groupe avec pénalités : la probabilité de réussir le même examen est de 45 % dans la condition « sans pénalités », mais seulement de 15 % lorsqu'il est associé à des pénalités.

Figure 2*Pourcentage de réussite selon les deux conditions*

Le fait de ne pas attribuer de points négatifs n'est cependant pas associé à un niveau élevé de performances. La moyenne obtenue pour le groupe A (11,46) est en effet inférieure au seuil de 12 attendu. Il en est de même pour le pourcentage de réussite, il est inférieur à la norme habituelle de 50 %. Si le hasard intervient réellement dans un QCM sans points négatifs, cet exemple illustre bien qu'il n'est pas nécessairement associé à la réussite du plus grand nombre.

La décision de faire appel ou non à un barème de pénalités est cruciale compte tenu de l'impact déterminant que ce choix aura sur la réussite ou non de chaque étudiant·e. Un examen sans pénalités permet de faire réussir près d'un·e étudiant·e sur deux, alors qu'avec pénalités, seul·e un·e étudiant·e sur six réussit (figure 2).

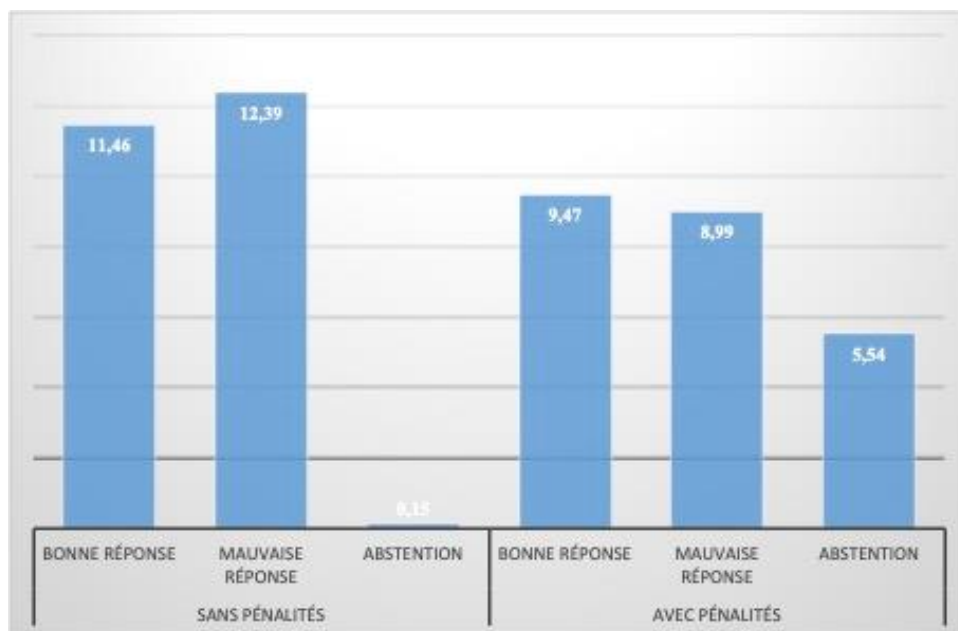


6.2 Des stratégies de réponses différentes

Le grand nombre d'abstentions dans l'examen avec pénalités permet d'expliquer cet écart. Les étudiant·es concerné·es par cet examen cumulent au total 1058 abstentions alors qu'on n'en relève que 31 pour l'ensemble de l'examen sans pénalités. Cela donne respectivement 5,54 abstentions en moyenne par étudiant·e pour l'examen avec pénalités pour seulement une moyenne de 0,15 abstention dans le format sans pénalités (cf. figure 3).

Figure 3

Moyenne des 3 types de réponses en fonction du type d'examen



Les stratégies de réponses sont fondamentalement différentes d'un examen à l'autre : l'utilisation des pénalités conduit un grand nombre d'étudiant·es à s'abstenir alors que ce n'est pas du tout le cas des étudiant·es qui passent un examen sans pénalités. Dans cette condition, ils « jouent » tous puisqu'ils n'ont mathématiquement rien à perdre. Ce n'est pas le cas dans la condition avec pénalités où un nombre conséquent d'étudiant·es « retiennent leur coup ». Pourtant, on voit que dans ce groupe le nombre moyen de bonnes réponses est légèrement supérieur à celui de mauvaises réponses. À l'inverse, les mauvaises réponses sont plus nombreuses dans le groupe sans pénalités. Ces observations mettent en évidence un paradoxe : le groupe dans lequel le nombre de réponses correctes est supérieur au nombre d'erreurs est en situation d'échec alors que le groupe où on observe la situation inverse s'en sort nettement mieux en termes de notes.

Au bout du compte, une note est attribuée à l'étudiant, sa réussite ou son échec sont décidés. L'une de ces deux formes d'examens permet-elle pour autant de certifier que l'étudiant·e maîtrise ou non ce qu'il·elle doit maîtriser ? Les docimologues savent depuis longtemps que les notes attribuées à une copie sont variables selon les correcteurs. Les chercheurs ont même renoncé à la notion de « note vraie » tant l'établissement de celle-ci est hasardeux (Merle, 2012). Position prudente et sage en ce qui concerne la recherche, elle interpelle quand on prend conscience que l'avenir d'un·e jeune est souvent lié à la réussite abusive ou à un échec abusif, ni l'une ni l'autre ne reflétant les réelles compétences de l'étudiant·e. De Ketele (2010) et Gerard (2017) rappellent que l'objectif de toute évaluation est d'éviter deux fléaux majeurs :



Les échecs abusifs Les "faux négatifs"	- Le risque est de refuser « à tort » un·e étudiant·e qui maîtrise les connaissances ou compétences visées par l'examen. - N'est-on pas en train d'exclure un·e candidat·e de façon inadéquate ou n'est-on pas en train de retarder inutilement son parcours alors qu'il·elle a le niveau requis ?	Sévérité excessive
Les réussites abusives Les "faux positifs"	- Le risque est de certifier « à tort » un·e étudiant·e qui ne maîtrise pas les connaissances ou compétences visées par l'examen. - N'est-on pas en train d'accepter de façon inadéquate un·e candidat·e à poursuivre son parcours alors qu'il·elle n'a pas le niveau requis ?	Indulgence excessive

Les qualités psychométriques de l'examen qui nous occupe permettent d'éliminer deux situations extrêmes : des questions tellement simples que toutes les étudiant·es, même ceux·celles qui n'ont pas ouvert leur cours, réussissent sans peine ou des questions tellement complexes qu'il faut être un·e expert·e pour y répondre. Cependant, les différences observées au niveau des pourcentages de réussites sont telles qu'elles amènent à nous questionner sur l'impact que peut avoir le système d'attribution des notes sur l'obtention de ces résultats.

- Dans le groupe sans pénalités, ne favorise-t-on pas la réussite abusive des étudiant·es alimentée notamment par le fait qu'ils·elles gagneraient des points en bénéficiant du hasard, avec une proportion élevée de faux positifs ?
- A contrario, dans le groupe avec pénalités, ne favorise-t-on pas les échecs abusifs chez les étudiant·es alimentés notamment par le fait qu'ils·elles sont nombreux·ses à s'abstenir plutôt que de démontrer leur potentielle maîtrise de la matière, avec au bout du compte une proportion élevée de faux négatifs ?

Pour répondre à ces questions, il est utile de faire *appel aux notions de « sensibilité » et de « spécificité »*.

7. Étude de la sensibilité et de la spécificité des deux examens

7.1 Méthodologie d'analyse

Les notions de sensibilité et de spécificité sont issues du domaine médical pour étudier la valeur prédictive d'un test ou d'un examen clinique, et par conséquent sa validité. La question est de savoir si le test atteint ou non son objectif. En médecine, l'objectif d'un test est de déterminer si la personne est malade ou non. Lors de la crise du Covid-19, qui n'a pas été testé en espérant chaque fois ne pas être « positif » ? Dans ce cas, le terme « positif » est utilisé pour dire que l'objectif du test est atteint : il a pu identifier une personne malade. Par contre, si le test est « négatif », cela signifie que la personne n'est pas malade. Dans ce cadre, nous avons été habitués aux termes « faux positifs » et « faux négatifs ». Ce n'est pas parce que le test est positif que nous sommes vraiment malades, tout comme ce n'est pas parce que le test est négatif que nous ne sommes pas porteurs du virus.



Ces notions sont utilisées dans le même sens en psychométrie pour analyser la validité d'un test qui vise à établir si une personne souffre ou non de troubles. Si le test est positif, cela signifie qu'il diagnostique la présence des dits troubles. Mais le test peut s'être trompé et diagnostiquer un « faux positif », ce qui signifierait qu'en réalité la personne ne souffre pas des troubles (Ecalte, 2010).

Le principe est le même en docimologie, mais l'objectif est inversé : le test, par exemple l'examen QCM qui nous occupe, vise à certifier la réussite ou non de l'étudiant·e. Le résultat est positif si l'étudiant·e réussit et négatif s'il·elle échoue (Laveault & Yerli, 2017). Mais il est possible, pour différentes raisons, qu'il y ait des erreurs dans le résultat de l'examen. Un·e étudiant·e peut réussir alors qu'il·elle ne maîtrise pas les compétences visées, c'est un « faux positif ». Un·e autre étudiant·e peut échouer alors qu'il·elle maîtrise les compétences, c'est un « faux négatif ». Le tableau suivant synthétise l'utilisation que nous faisons de ces termes :

Tableau 8

Classification en fonction du niveau de compétences et de son anticipation sur base de la réussite à un examen

	Maîtrise les compétences	Ne maîtrise pas les compétences
Réussit l'examen	Vrai positif (VP)	Faux positifs (FP)
Échoue à l'examen	Faux négatif (FN)	Vrai négatif (VN)
	Sensibilité $Se = VP / (VP + FN)$	Spécificité $Sp = VN / (VN + FP)$

Sur base d'un score seuil obtenu au test, il est possible de calculer plusieurs indicateurs de la validité du test :

- La **sensibilité** correspond au *taux de vrais positifs*, c'est-à-dire à la proportion d'individus réussissant l'examen sur la base de leur score au test qui se révèlent effectivement compétents. Elle se calcule à l'aide de la formule $VP / (VP + FN)$. Une sensibilité égale à 1 signifie que 100 % des personnes identifiées comme compétentes par le test le sont effectivement. Les valeurs inférieures à 1 indiquent que certains sujets compétents ne sont pas identifiés par le test. Ce sont les *faux négatifs*.
- La **spécificité** correspond au *taux de vrais négatifs*, c'est-à-dire à la proportion d'individus échouant au test alors qu'ils sont effectivement incompetents. Elle se calcule à l'aide de la formule $VN / (VN + FP)$. Une spécificité égale à 1 signifie qu'aucune personne incompetente ne réussit le test (absence de faux positifs).
- Le **taux de faux positifs** correspond à la proportion d'individus qui réussissent à tort l'examen. Il se calcule à l'aide de la formule $FP / (FP + VN)$ et est égal à $(1 - \text{spécificité})$.
- Le **taux de faux négatifs** correspond à la proportion d'individus qui échouent à tort à l'examen. Il se calcule à l'aide de la formule $FN / (FN + VP)$ et est égal à $(1 - \text{sensibilité})$.

La question est dès lors de savoir comment valider la valeur prédictive du test, et par là le test lui-même. En médecine, le test à évaluer doit être comparé à un test de référence déjà existant appelé « gold standard » (pour un cancer, le test de référence est la biopsie). Parfois le test de référence n'existe pas. Lorsque c'est le cas, on va classer les sujets avec un faisceau d'arguments dont la validité doit avoir été étudiée.



Dans notre cas, nous avons choisi de prendre comme « test de référence » la décision prise par le jury des examens en août 2020 qui après délibération prononce la réussite ou non de l'étudiant·e pour l'ensemble des cours qui composent le programme de 1^{re} année.

On aurait pu aussi mettre en relation notre examen de janvier 2020 avec la note obtenue dans le même cours en fin d'année académique attestant officiellement du niveau de maîtrise de la matière cible. Cette comparaison aurait eu peu de sens : ce serait comme si on voulait tester la validité prédictive (ou diagnostique) d'un test de dépistage en médecine avec ce même test utilisé six mois plus tard ! C'est pourquoi nous avons choisi un critère d'un niveau supérieur : la réussite de la première année attestée par la décision finale du jury d'examens.

Ce critère atteste officiellement de l'acquisition des acquis d'apprentissage de la première année de formation pour chaque étudiant·e. Il donne une idée de son profil global au terme de l'année d'études et les examens des différentes matières contribuent à s'assurer que les étudiant·es ont effectivement les compétences attendues. La validité prédictive de notre examen par rapport à ce critère de réussite ou d'échec en fin d'année prend la forme d'un tableau de contingence inspiré de ceux que l'on trouve dans le champ médical.

Tableau 9

Classification des étudiant·es sur base de leur score seuil à l'examen de janvier et de leur réussite en fin de 1^{re} année

		Réussite en fin de 1 ^{re} année	
		Oui	Non
Réussite à l'examen de janvier 2020	Oui (Score $\geq$ 12)	Vrais positifs (VP) <i>Ce sont les étudiant·es qui ont réussi en fin d'année et qui ont réussi l'examen</i>	Faux positifs (FP) <i>Ce sont les étudiant·es en échec en fin d'année mais qui ont réussi l'examen</i>
	Non (Score < 12)	Faux négatifs (FN) <i>Ce sont les étudiant·es qui ont réussi en fin d'année mais qui sont en échec à l'examen</i>	Vrais négatifs (VN) <i>Ce sont les étudiant·es en échec en fin d'année et qui sont en échec à l'examen</i>

Sachant qu'en Fédération Wallonie-Bruxelles, un·e étudiant·e est officiellement en situation de réussite lorsque le jury valide 60 crédits sur 60 au terme de son année, nous avons classé chaque étudiant·e réparti·e dans nos deux groupes en fonction de ce critère. Voici les tableaux de contingence qui en résultent.

Tableau 10

Tableau de contingence pour l'examen sans pénalités

		Réussite en fin de 1 ^{re} année	
		Oui (Les étudiant·es ont 60 crédits)	Non (Les étudiant·es ont moins de 60 crédits)
Réussite à l'examen de janvier 2020	Oui (Score $\geq$ 12)	Vrai positif (VP)=54	Faux positif (FP)=42
	Non (Score < 12)	Faux négatif (FN)=14	Vrai négatif (VN)=92
		68	134



Tableau 11*Tableau de contingence pour l'examen avec pénalités*

		Réussite en fin de 1 ^{re} année	
		Oui (Les étudiant-es ont 60 crédits)	Non (Les étudiant-es ont moins de 60 crédits)
Réussite à l'examen de janvier 2020	Oui (Score ≥ 12)	Vrai positif (VP)=19	Faux positif (FP)=10
	Non (Score < 12)	Faux négatif (FN)=44	Vrai négatif (VN)=116
		63	126

Nous avons calculé la valeur des 4 indicateurs cités plus haut auxquels nous avons ajouté 3 autres indicateurs statistiques permettant d'avoir une idée des performances respectives des deux examens en mesurant l'intensité de la relation entre leur réussite en janvier et celle en fin de 1^{re} année.

- L'**indice de Youden** est une mesure de la précision d'un test de diagnostic. Il dépend de la spécificité et de la sensibilité du test. Si la valeur de l'indice est négative, il est inefficace. Plus cette valeur se rapproche de 1 et plus le test est efficace.
- Le **coefficient Q de Yule** est une mesure permettant de quantifier la liaison entre deux variables dichotomiques. Plus le résultat est proche de 1, plus forte est l'intensité de la relation.
- Le **test χ^2** (Khi carré). La valeur critique pour un seuil de 0,05 étant de 3,84, plus le résultat du test est supérieur à cette valeur, plus la relation entre les deux variables est significative.

Les valeurs que prennent les différents indicateurs pour chacun des deux examens sont présentées dans le tableau 12.

Tableau 12*Analyse comparative des performances des deux formes d'examen*

	Examen sans pénalités	Examen avec pénalités
Sensibilité	79,4%	30,2%
Spécificité	68,7%	92,1%
Taux de faux positifs	31,3%	07,9%
Taux de faux négatifs	20,6%	69,8%
Indice de Youden	0,48	0,22
Coefficient Q de Yule	0,79	0,67
X (Khi Carré)	41,8	16,0

7.2 Interprétation de la sensibilité et de la spécificité des deux examens

Sur base de ces indicateurs, on peut observer que l'examen sans pénalités présente de meilleures qualités diagnostiques que l'examen avec points négatifs. Le seul avantage de ce dernier est de présenter une très haute spécificité puisqu'on peut voir qu'il y a peu d'étudiant-es incompetent-es en fin d'année qui réussissent cet examen. On enregistre par conséquent un taux réduit de faux positifs (8 %), mais cette qualité est associée à une sensibilité très faible



nettement insuffisante par rapport aux standards attendus. La sensibilité observée de 30 % signifie en effet qu'il y a 70 % de faux négatifs, c'est-à-dire qui ont échoué à l'examen alors qu'ils n'auraient pas dû. L'examen avec pénalités identifie donc à tort un problème qui n'existe pas pour un nombre conséquent d'étudiant·es. À l'inverse, l'examen sans points négatifs récompense les étudiant·es qui ont les meilleures capacités académiques puisque sa sensibilité est élevée (80 %).

Idéalement, il faudrait un examen à la fois très sensible et très spécifique. Mais cette situation n'existe pas. Le choix de privilégier la sensibilité ou la spécificité dépend souvent de l'utilisation que l'on veut faire du test, mais aussi des conséquences de ce choix.

À ce sujet, Laveault et Grégoire (2014) rappellent que la notion de validité ne se rapporte pas au test en lui-même, mais bien aux inférences que l'on peut faire à partir de ses résultats. Ces auteurs évoquent les travaux de Messick (1988) sur la validité de conséquence : la finalité des scores est de servir de base à l'action et leur valeur ultime ne dépend pas uniquement de leur validité de contenu, mais aussi de leur impact individuel et social.

Cette problématique est souvent passée sous silence. On peut l'approcher ici en analysant attentivement le tableau de contingence de l'examen avec pénalités (cf. Tableau 11 ci-dessus). On voit que parmi les étudiant·es ayant réussi leur année avec 60 crédits, la majorité (44 sur 63) était en échec à l'examen avec pénalités ! Suite au recours, ces étudiant·es ont eu l'occasion de représenter l'examen sans points négatifs et nous avons pu nous rendre compte que 42 d'entre eux·elles l'ont réussi lors de cette 2e tentative, cette fois sans recours aux pénalités. Les deux dernier·ères étudiant·es l'ont réussi en 2e session dans un examen également sans pénalités. Ils·elles ont donc finalement tous réussi et cela a contribué à leur succès en fin d'année. Rien ne dit qu'il en aurait été de même si on avait entériné officiellement la note initiale et que l'on aurait maintenu l'usage des pénalités pour leur deuxième tentative.

On rappellera aussi que l'examen avec pénalités est associé à un taux d'échec trois fois plus élevé que l'examen sans pénalités. De très nombreux·ses étudiant·es sont donc amenés à représenter l'examen avec tout ce que cela suppose en termes de conséquences psychologiques négatives (stress, perte de confiance, estime de soi...), mais aussi en termes de coûts organisationnels et économiques liés à la surcharge de travail demandée aux enseignant·es qui sont amené·es à réinterroger à deux ou trois reprises des dizaines d'étudiant·es qui auraient pu réussir dès leur premier essai, car ils·elles avaient en réalité une maîtrise suffisante de la matière.

Avec l'examen à points négatifs, nous sommes donc en présence d'un instrument dont la validité de conséquence est problématique. Sa sévérité excessive ainsi que le déséquilibre entre sa spécificité très élevée et sa sensibilité très basse remettent également en question sa validité puisqu'il est associé à de nombreuses erreurs de prédiction concernant l'aptitude à réussir une première année d'études universitaires.

Une façon simple de le vérifier consiste à calculer un coefficient de corrélation entre les scores de chaque examen et les résultats obtenus en fin d'année par les étudiant·es des deux groupes (nombre de crédits obtenus, pourcentage des points pour l'ensemble des matières et décision de réussite). Si nos analyses et interprétations sont correctes, on peut faire l'hypothèse que le degré de corrélation sera plus fort pour l'examen sans pénalités que pour l'examen avec pénalités, comme le montre le tableau 13.



Tableau 13

Corrélations entre les scores enregistrés dans les deux examens et les résultats obtenus en fin d'année

	Score à l'examen sans pénalités	Score à l'examen avec pénalités
Décision de fin d'année	0,535	0,421
Nombre de crédits	0,584	0,461
Points obtenus en moyenne	0,618	0,536
Moyenne des coefficients de corrélations	0,58	0,47

Tous les coefficients de corrélation de Pearson sont significatifs au niveau 0,01 (test bilatéral), mais les scores à l'examen sans pénalités et les résultats en fin d'année partagent une part de variance commune plus élevée que celle de l'examen avec pénalités. Dans le premier cas, la part de variance expliquée donnée par le carré du coefficient des corrélations (R^2 en moyenne 0,60) s'élève à 34 % alors qu'elle est de 22 % pour l'examen avec points négatifs (R^2 en moyenne 0,49). Cette analyse confirme la suprématie prédictive de l'examen sans pénalités.

En conclusion, nos analyses convergent pour indiquer que l'examen sans pénalités permet de mieux mesurer le niveau réel de connaissances des étudiant·es. Celui-ci n'est pas pour autant sans défaut puisqu'un nombre non négligeable d'étudiant·es peuvent être considéré·es comme des faux positifs. La faible spécificité des examens sans pénalités (vérifiée en partie ici) est une des principales préoccupations des enseignant·es. Ce type d'examens favoriserait la réussite d'étudiant·es qui ne maîtrisent pas la matière (les faux positifs, donc) et c'est la raison principale qui motive les professeur·es à utiliser des pénalités dans leurs barèmes de notation. Mais il y a une solution à ce problème : augmenter le seuil de réussite, en instaurant un « cut-off point » supérieur aux traditionnels 50 % (Dejonckheere, 2017 ; Lesage et al., 2013). Nous développerons cette solution dans un prochain article.

8. Existe-t-il un biais de genre dans l'examen avec pénalités ?

Le principe même de tout examen est de placer l'étudiant·e dans les meilleures conditions pour démontrer son niveau de maîtrise de la matière le jour J (Morissette, 2005). À partir du moment où on fait usage de pénalités, les étudiant·es sont obligé·es de faire des choix de type « je joue ou je ne joue pas » pour chacune des questions posées. Leur décision se base sur une estimation du coût et du gain compte tenu de la valeur des pénalités qui seront décomptées en cas d'erreur, mais aussi de leur niveau de confiance à choisir la bonne réponse parmi les options qui leur sont présentées. Sans oublier un autre paramètre : le nombre de points qu'ils pensent avoir engrangés sur base de leurs réponses aux autres questions. Le QCM à points négatifs donne par conséquent une grande place à des considérations stratégiques sans lien avec la maîtrise de la matière (Bar-Hillel et al., 2005 ; Chica & Tarrago, 2021). À compétences égales, cette attitude face au risque peut grandement varier d'un étudiant à l'autre, avec pour effet de modifier les résultats de l'évaluation (Espinosa & Gardeazabal, 2010). Ce biais sera d'autant plus grand que l'on pénalise fortement une mauvaise réponse (Lecroart, 2016).

Selon certaines études, le recours à la pénalisation des mauvaises réponses pousserait les étudiant·es très faibles à la prise de risque : ils n'ont rien à perdre et parfois ils peuvent gagner ! Ce système inciterait par contre de bon·es étudiant·es à une prudence excessive, voire à la paralysie – ils ou elles s'abstiennent de répondre – ce qui peut les conduire à l'échec



(Uyttebroeck, 2014). À ce « jeu », les femmes seraient désavantagées par rapport aux hommes. Plusieurs études mettent en effet en avant des différences significatives de résultats en faveur des hommes dans le cas des QCM par points négatifs (Arthur & Everaert, 2012 ; Baldiga, 2014 ; Coffman & Klinowski, 2019 ; Iriberry & Rey-Biel, 2019 ; Kacprzyk, et al., 2019 ; Ramos & Lambating, 1996). Dans cette situation, les hommes prendraient plus de risques et potentiellement, ils amélioreraient leurs scores. Les femmes, à l'inverse, auraient plutôt tendance à s'abstenir quand elles ne sont pas certaines de leur réponse et le fait de s'abstenir entraîne de moins bons résultats (Brogniaux, 2021 ; Roggemans & Spruyt, 2014). D'autres études nuancent ces conclusions quand elles ne les remettent pas en question (Akyol et al., 2022 ; Funk & Perrone 2016 ; Ricketts et al., 2010;). Ce manque de consensus s'explique notamment par le fait qu'il n'est pas possible de monter un design expérimental où on testerait les deux conditions (avec ou sans pénalités) dans une situation réelle d'examen. Une différence de performance entre deux groupes, par exemple femmes et hommes, n'implique pas nécessairement qu'il y ait un biais. En effet, pour parler d'un biais, il faut que deux individus, de compétences identiques mais appartenant à des groupes différents n'aient pas la même probabilité de réussite. Les circonstances dans lesquelles s'est déroulé notre examen auprès de deux groupes présentant les mêmes caractéristiques permettent de tester l'hypothèse du biais de genre. À notre connaissance, une telle étude est inédite.

8.1 Comparaison des résultats dans les deux conditions en fonction du genre

Le tableau 14 et la figure 4 permettent de visualiser les performances des étudiantes et des étudiants dans les deux types d'examen. On voit que l'utilisation d'un barème avec des pénalités a un impact important dans les deux groupes sur la moyenne des résultats. Les différences observées entre les deux conditions sont très significatives tant pour les étudiantes ($t=9,250$; p bilatéral $<,001$) que pour leurs condisciples masculins ($t=3,778$; p bilatéral $<,001$). On remarque que la perte des femmes est plus substantielle que celle des hommes. Leur moyenne chute de 4,56 points entre les deux situations (11,84-7,28) alors que la perte des hommes se limite à 3,34 points (10,36-7,02).

Tableau 14

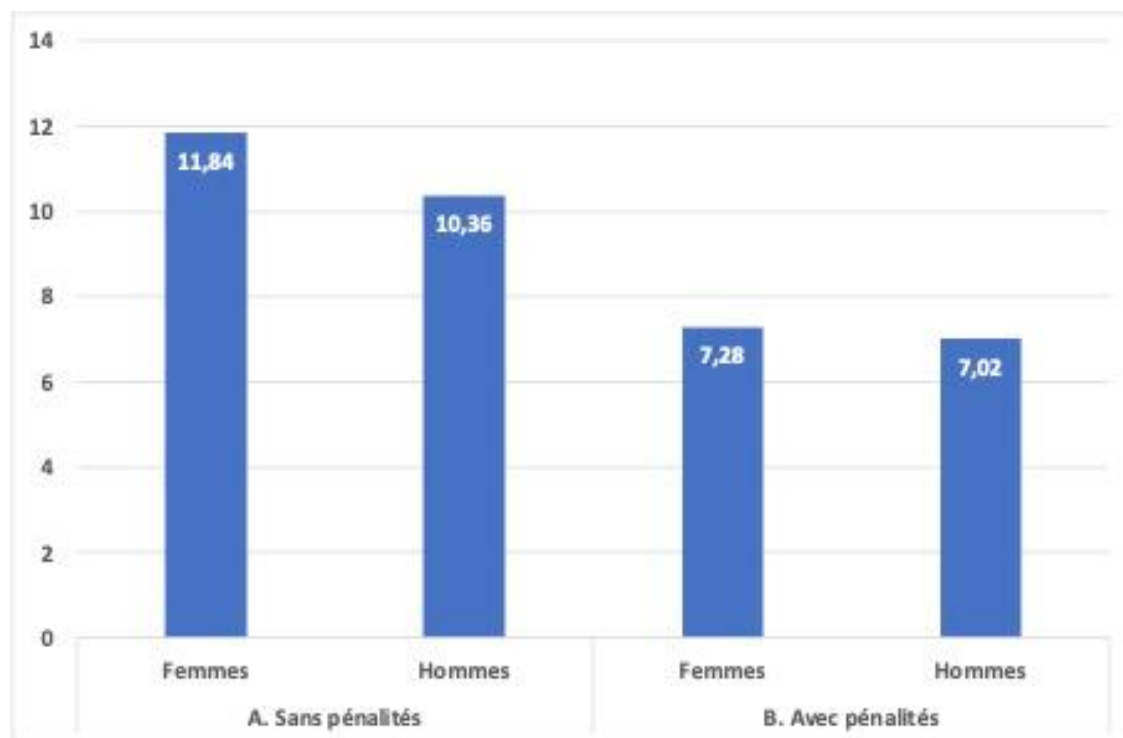
Statistiques descriptives dans les deux examens en fonction du genre

	Sans pénalités			Avec pénalités		
	N	Moyenne	Écart-type	N	Moyenne	Écart-type
Femmes	155	11,84	3,94	149	7,28	4,57
Hommes	53	10,36	4,01	42	7,02	4,66



Graphique 4

Moyennes des points obtenus dans les deux situations par les femmes et par les hommes

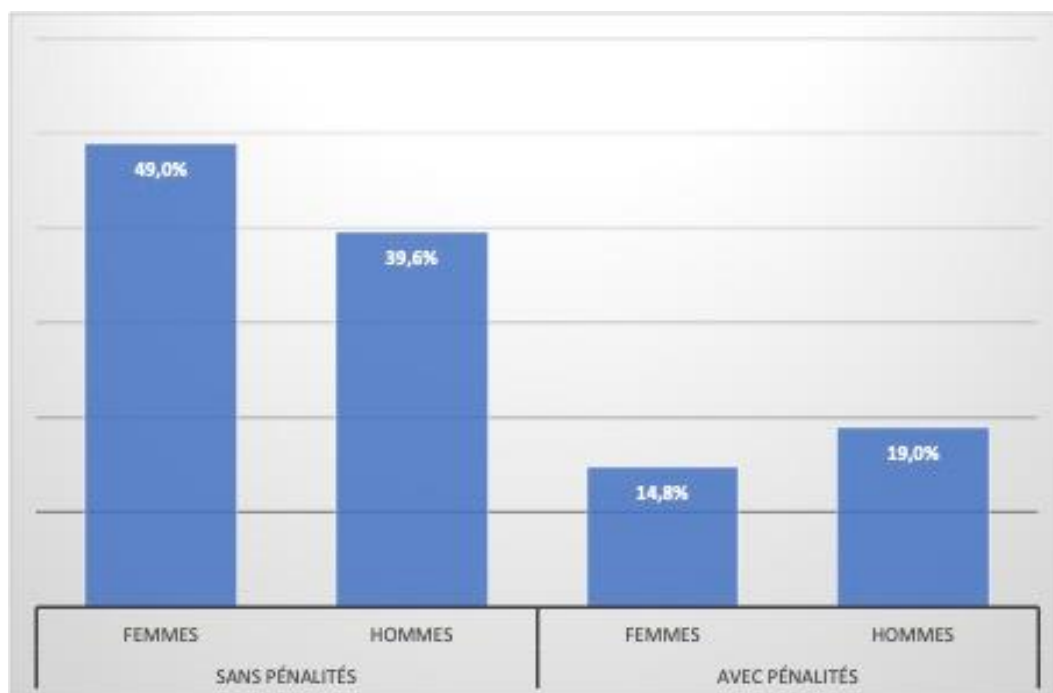


Plus fondamentalement, il apparaît que les femmes perdent dans l'examen avec pénalités l'avantage qu'elles ont sur les hommes lorsque sont examinées leurs aptitudes dans l'examen sans pénalités. La comparaison entre ces deux sous-groupes montre en effet que les différences de performances sont significatives en dessous du seuil de ,05 au test t de Student dans cette dernière condition ($t=-2,349$; p bilatéral=0,02) alors que ce n'est pas le cas dans l'examen avec pénalités ($t=0,316$; p bilatéral=0,75). Ce résultat est important, car il met clairement en évidence l'existence d'un biais à la défaveur des femmes.

8.2 Comparaison des deux groupes en pourcentages de réussite

La moyenne des points est une chose, le nombre de réussites en est une autre. Dans le cadre d'un examen certificatif, on sait que l'enjeu principal pour les étudiant-es est d'obtenir le nombre de points suffisant pour passer le seuil de réussite. Pour rappel, il est ici fixé à 50 % des points, soit un score de 12 points sur 24. La figure 5 montre le pourcentage des réussites à cet examen en fonction du genre et de la situation d'examen qui était proposée. Ces résultats confirment à quel point le recours à des points négatifs est défavorable aux femmes.



Graphique 5*Pourcentages de réussite en fonction du genre dans les deux conditions*

- Les femmes réussissent mieux que les hommes dans la condition sans pénalités, mais pas dans la condition avec pénalités. Dans ce type d'examen, les femmes ont même un niveau de réussite inférieur à celui des hommes alors que la condition sans pénalités indique clairement qu'on peut éliminer l'hypothèse qu'elles auraient une moins bonne connaissance de la matière sur laquelle on les interrogeait.
- Si on compare les taux de réussite dans les deux types d'examens, on voit que les femmes réussissent 3 fois moins dans la condition avec pénalités que dans la condition sans pénalités (14,8 % vs 49 %). Les hommes s'en sortent mieux puisque qu'ils ne sont « que » deux fois moins à réussir (19,0 % vs 39,6 %).
- Dans l'examen avec pénalités, alors que la performance moyenne des femmes est légèrement supérieure à celle des hommes (7,28 points vs 7,08 points), leur taux de réussite est paradoxalement moins bon (14,8 % vs 19,0 %). La différence entre les genres est importante en termes de points, mais son impact est encore plus marqué en ce qui concerne les taux de réussite.

Ces différents éléments renforcent l'hypothèse d'un biais à la défaveur des femmes. Leur sous-performance dans l'examen avec pénalités est associé à un risque d'échec très élevé. Ce risque est finalement plus important que celui des hommes alors que leur maîtrise de la matière est meilleure, comme l'indiquent leurs résultats dans l'examen sans pénalités. Le recours aux points négatifs aboutit à créer une injustice à l'encontre des femmes.

8.3 Comparaison des deux groupes en ce qui concerne la validité prédictive

Pour estimer l'ampleur et les conséquences de ce biais en défaveur des femmes, il est important de comparer pour les deux groupes la relation entre le score à l'examen avec



pénalités et le résultat obtenu en fin d'année académique. Cette démarche est résumée dans les tableaux de contingence présentés ci-dessous.

Tableau 15

Tableau de contingence pour les femmes

		Réussite en fin de 1 ^{re} année	
		Oui (Les étudiantes ont 60 crédits)	Non (Les étudiantes ont moins de 60 crédits)
Réussite à l'examen de janvier 2020	Oui (Score ≥ 12)	Vrai positif (VP)=15	Faux positif (FP)=7
	Non (Score < 12)	Faux négatif (FN)=40	Vrai négatif (VN)=85
		55	92

Tableau 16

Tableau de contingence pour les hommes

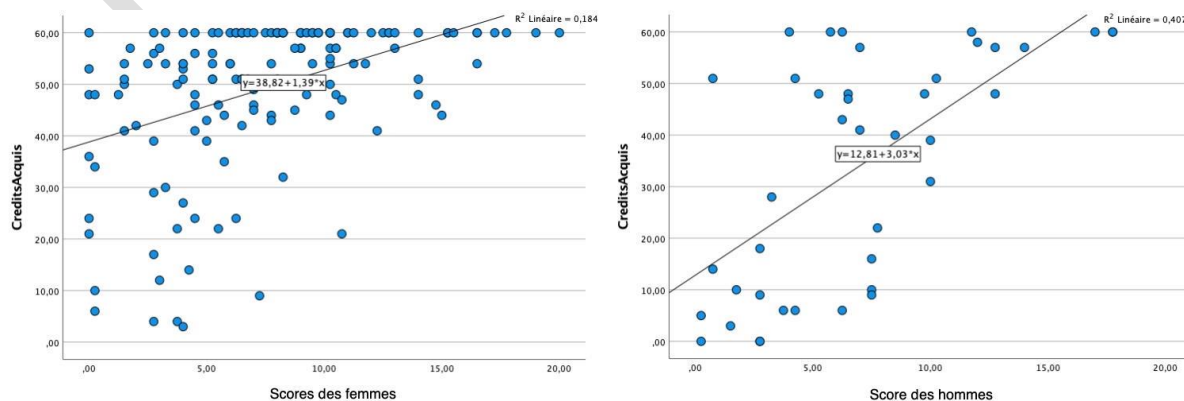
		Réussite en fin de 1 ^{re} année	
		Oui (Les étudiants ont 60 crédits)	Non (Les étudiants ont moins de 60 crédits)
Réussite à l'examen de janvier 2020	Oui (Score ≥ 12)	Vrai positif (VP)=4	Faux positif (FP)=3
	Non (Score < 12)	Faux négatif (FN)=4	Vrai négatif (VN)=31
		8	34

On voit que la proportion des hommes correctement classés (vrais positifs et vrais négatifs) est élevée puisqu'on obtient un indice de concordance entre les deux décisions qui s'élève à 0,83 ($CO=(VP+VN)/N$). Cet indice n'est que de 0,68 pour les femmes. Cela signifie qu'on fait une erreur de prédiction de 32 % pour les femmes alors qu'on ne rencontre ce problème que dans 17 % des cas pour les hommes. La validité prédictive de l'examen n'est pas du tout la même dans les deux groupes.

Pour illustrer ce constat d'une autre manière, nous avons représenté sous forme de nuages de points la relation qui existe entre les scores à l'examen et la réussite en fin d'année pour chacun des groupes.

Figure 6

Évaluation de la validité prédictive différentielle en fonction du genre



La pente de la droite de régression est différente pour les deux groupes. Contrairement à ce qu'on peut observer chez les hommes, il n'y a pratiquement pas de relation chez les femmes entre le score à l'examen et les résultats qui sont obtenus à la fin de l'année comme l'indique un R^2 linéaire beaucoup plus bas. Une valeur de 0,18 indique qu'il est pratiquement impossible de faire une prédiction des résultats qui seront obtenus par les femmes à la fin de l'année sur base de leur score à l'examen avec pénalités puisque 82 % de la variance observée sur la première variable est sans relation avec la seconde variable. Ces observations montrent bien que l'examen avec pénalités conduit à des prédictions différentes en fonction du groupe d'appartenance : il a une meilleure validité prédictive pour les hommes que pour les femmes. C'est un biais de validité prédictive (Laveault & Grégoire, 2014).

Ce type de biais a déjà été mis en avant notamment par Rothstein (2004), Ors, Palomino et Peyrache (2008) ou encore Pekkarinen (2015) : les femmes ont tendance à sous-performer dans un examen avec pénalités par rapport à leur performances académiques passées et futures. Cela est confirmé ici par la proportion importante de femmes classées comme « faux négatifs » (tableau 15) et par le nombre élevé de femmes qui se trouvent dans le cadran supérieur gauche de la figure 6.

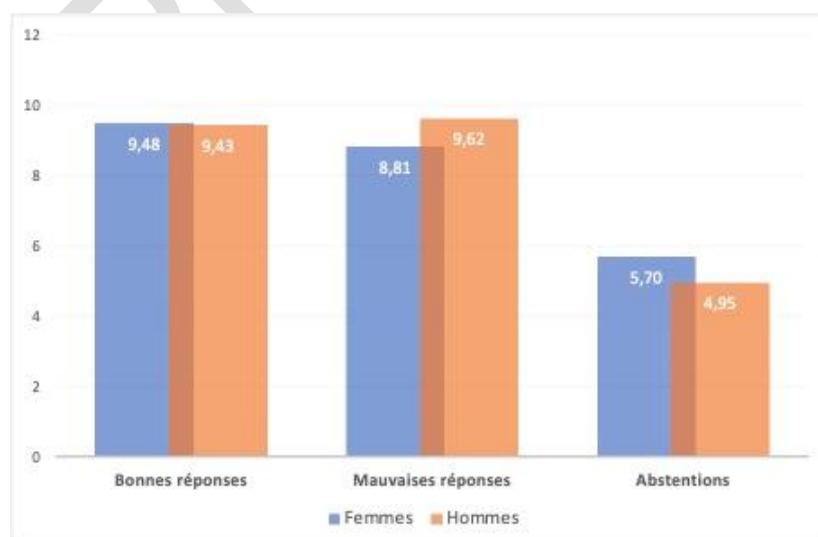
8.4 Comment expliquer ce biais de genre ?

Les stratégies de réponse sont fondamentalement différentes d'un examen à l'autre. L'utilisation des pénalités conduit un grand nombre d'étudiant-es à s'abstenir alors que ce n'est pas du tout le cas des étudiant-es qui passent un examen sans pénalités. De nombreuses données empiriques montrent que les femmes ont une plus grande aversion à la prise de risque que les hommes et que cette prudence se traduit dans les QCM à points négatifs par des taux d'abstention plus élevés, ce qui a pour effet de diminuer leur note (pour un résumé, voir Cipriani, 2018). Cette hypothèse est au cœur de la dernière partie de notre travail.

Dans un premier temps, nous avons calculé la moyenne des trois types de réponses possibles en fonction du genre dans les deux types d'examen. La figure 7 présente le résultat de cette comparaison.

Figure 7

Nombre moyen de bonnes réponses, de mauvaises réponses et d'abstentions en fonction du genre



On peut voir que les femmes répondent en moyenne à moins de questions que les hommes puisque qu'elles s'abstiennent un peu plus qu'eux. Quand elles ne s'abstiennent pas, les femmes répondent correctement à autant de questions que les hommes (9,48 vs 9,43). Ceux-ci, par contre, enregistrent un nombre moyen de mauvaises réponses plus élevé (9,62 vs 8,61).

Lorsqu'on convertit les moyennes en pourcentages, les femmes et les hommes répondent respectivement à 76,2 % et à 79,4 % des questions de l'examen, soit un différentiel d'un peu plus de 3 % qui exprime aussi l'écart qui existe en termes de pourcentages d'abstentions. Cette différence semble plutôt minime à première vue – elle n'est d'ailleurs pas statistiquement significative – mais une étude menée en Belgique par la VUB à propos de l'examen d'entrée en études de dentisterie montre que tout peut se jouer à un ou à un demi-point près (Roggemans & Spruyt, 2014). Cette étude a notamment montré que la propension des femmes à s'abstenir lorsqu'elles ont un niveau proche de la réussite entraînent une bonne partie d'entre elles à l'échec alors que si elles avaient pris le risque de répondre à l'une ou l'autre question supplémentaire, comme le font les hommes, elles auraient augmenté sensiblement leurs chances de réussite. D'autres études aboutissent à la même conclusion (Brogniaux, 2021 ; Marín & Rosa-García, 2011).

En d'autres termes, les femmes proches du score seuil répondent à moins de questions que les hommes et elle obtiennent un résultat inférieur qui les conduit plus souvent à l'échec. C'est la raison pour laquelle il semble important de nous pencher sur le nombre de réponses correctes, incorrectes et laissées ouvertes (abstentions) en fonction du niveau de performance chez les femmes et chez les hommes. Nous avons réparti les étudiant·es en quatre sous-groupes en fonction de leurs résultats à l'examen : le groupe des étudiant·es ayant obtenu moins de 5 points, le groupe de celles et ceux ayant obtenu un score compris entre 5 et 8,75, la catégorie des étudiant·es ayant récolté un score compris 9 et 11,75 points et enfin la catégorie des étudiant·es ayant réussi l'examen (score supérieur à 12 points sur 24).

Tableau 17

Répartition des hommes et des femmes en fonction de leur score à l'examen

	< 5 pts	Entre 5 et 8,75	Entre 9 et 11,75	> 12 points
Femmes	46	50	31	22
Hommes	15	15	5	7

Figure 8

Répartition du type de réponses pour les femmes en fonction de leur score à l'examen

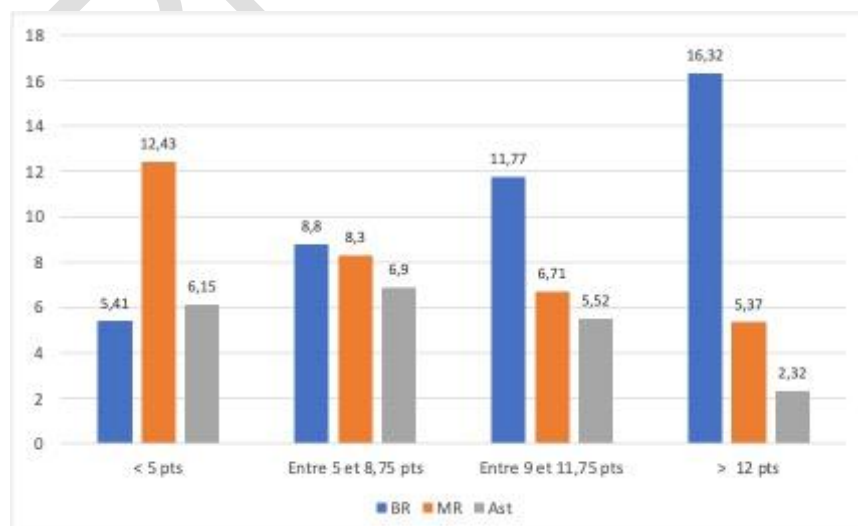
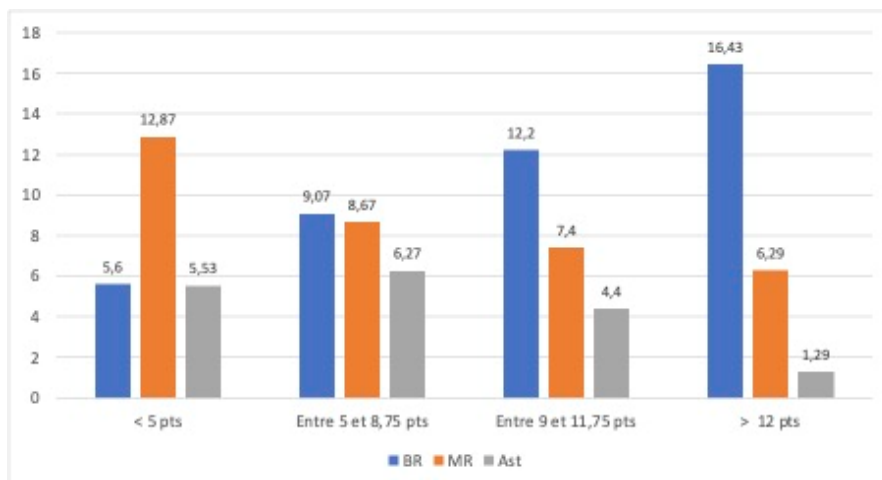


Figure 9

Répartition du type de réponses chez les hommes en fonction de leur score à l'examen



Certaines différences sont évidentes : quel que soit leur genre, les participant·es ayant un score élevé combinent un nombre élevé de bonnes réponses avec un faible nombre de réponses incorrectes et vice versa. On voit aussi que les femmes ont tendance à s'abstenir davantage que les hommes quel que soit le niveau de performance à l'examen.

Si on s'intéresse à présent à la catégorie qui nous intéresse le plus, celle des résultats proches du seuil de réussite (entre 9 et 11,75 pts), on observe que les femmes répondent proportionnellement à moins de réponses que les hommes. La différence est de 5 % : 77 % contre 82 %. On sera prudent en ce qui concerne cette comparaison, car le groupe de hommes repris dans cette catégorie est très faible (n=5) mais le fait que les femmes qui ne sont pas très loin du score seuil aient laissé près d'une question sur quatre sans réponse (23 %) mérite que l'on s'attarde un peu plus finement à leurs stratégies.

Nous avons effectué un zoom supplémentaire en répartissant les 31 étudiantes ayant obtenu entre 9 et 11,75 points en trois sous-groupes équivalents en termes d'effectifs en fonction du score qu'elles ont obtenu à l'examen. Les différents types de réponses qu'elles produisent est présenté dans le tableau 18.

Tableau 18

Répartition du type de réponses des étudiantes proches de la réussite

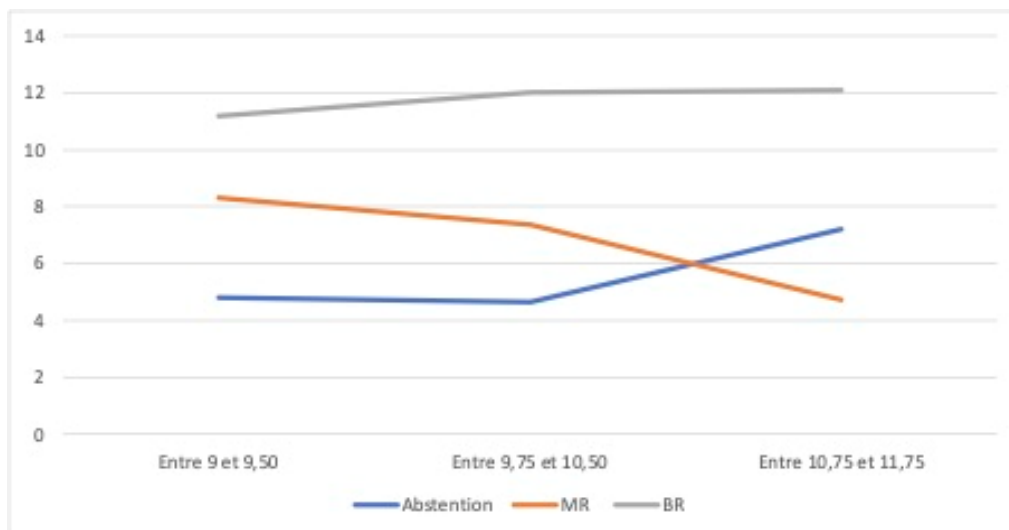
	Entre 9 et 9,50 (N=10)	Entre 9,75 et 10,25 (N=11)	Entre 10,50 et 11,75 (N=10)
Bonnes réponses	11,2	12	12,1
Mauvaises réponses	8,0	7,36	4,7
Abstentions	4,8	4,64	7,2

Lorsqu'on reproduit ces données dans un graphique, on voit que les étudiantes les plus proches de la note de 12 sur 24, synonyme de réussite, s'abstiennent davantage que leurs condisciples à répondre tout en réduisant sensiblement le nombre de mauvaises réponses. Il est intéressant de noter que les deux courbes se croisent. Cette configuration laisse à penser que ces étudiantes adoptent une stratégie d'évitement des risques : elles préfèrent s'abstenir plutôt que de cocher une réponse dont elles ne sont pas certaines. Cette stratégie s'avère finalement infructueuse puisque leur score se situe juste en dessous du *cut-off*. Elles échouent de peu, mais elles échouent.



Figure 10

Évolution des trois types de réponses pour les étudiantes ayant un score proche de la réussite



Ces étudiantes ont les prérequis nécessaires pour réussir l'examen puisqu'elles répondent correctement à la moitié des questions, mais elles font preuve d'une prudence sans doute excessive. Quand elles ne sont pas certaines de leurs réponses, elles préfèrent s'abstenir plutôt que de risquer une erreur, un peu comme ces joueurs de tennis qui retiennent leur coup et n'osent plus monter au filet alors qu'ils sont sur le point de gagner leur match. Ils finissent par laisser filer la victoire, paralysés par la peur de gagner.

Cette analogie nous conduit à nous poser une dernière question ? N'y aurait-il pas une interaction entre la prise de risques et le niveau académique des étudiantes ? En d'autres termes, celles qui ont les capacités pour réussir des études universitaires n'auraient-elles pas plutôt tendance à s'abstenir de répondre par rapport à leurs condisciples moins performantes. Ces dernières seraient davantage enclines à prendre plus de risques dans un examen à points négatifs afin de potentiellement améliorer leur score compte tenu d'une maîtrise moins affirmée de leurs connaissances.

Nous avons testé cette hypothèse en subdivisant les 125 étudiantes ayant échoué à l'examen en fonction de leurs résultats de fin d'année. Les résultats, présentés dans le tableau 19, montrent que les étudiantes ayant obtenu moins de 45 crédits s'abstiennent moins de répondre en moyenne que les deux autres catégories plus performantes. Une analyse des comparaisons des moyennes deux à deux à l'aide du test t de Student montre qu'il y a une différence significative au seuil de ,05 entre le 1^{er} groupe et le 3^e groupe ($t=-2,315$, p bilatéral=,02) ainsi qu'entre le 2^e et le 3^e groupes ($t=-1,999$; p bilatéral=,04). Il n'y a pas de différence par contre entre le 1^{er} et le 2^e groupes ($t=-0,442$; $p=,66$).

Tableau 19

Nombre d'abstentions des étudiantes en fonction de leur résultat en fin d'année

	N	Nombre moyen d'abstentions	Écart-type
60 crédits sur 60	40	7,20	5,06
Entre 45 et 60 crédits	51	6,73	5,11
< 45 crédits	34	4,59	4,63



Notre hypothèse est globalement confirmée : l'utilisation des pénalités conduit les étudiantes qui ont de bonnes capacités académiques à adopter une attitude plus prudente que les étudiantes plus faibles qui, elles, répondent à plus de questions.

Cette crainte de perdre des points quand on n'est pas totalement certain de la bonne réponse est sans doute à l'origine du sort peu enviable réservé aux étudiantes ayant obtenu 60 crédits sur 60 à la fin de l'année. Elles ne sont en effet que 15 sur 55 à avoir réussi l'examen avec pénalités (27,3 %) alors qu'on enregistre un taux de réussite trois fois supérieur de 77,6 % chez les étudiantes ayant réussi également leur année mais qui ont passé le même examen sans pénalités (45 réussites sur 58).

Ce résultat montre que le biais de genre touche aussi les étudiantes présentant les meilleures habilités académiques. Il serait à l'origine d'un paradoxe de genre : alors que les femmes performant beaucoup mieux que les hommes dans les études supérieures, notre étude permet de comprendre pourquoi les examens à points négatifs constituent la seule forme d'évaluation où elles obtiennent de moins bons résultats que les hommes (Roggemans & Spruyt, 2014).

9. Conclusion

Cette recherche met en évidence deux résultats importants. Premièrement, l'examen avec pénalités crée un nombre très élevé d'échecs en première session tout en présentant une validité prédictive moins bonne que l'examen sans points négatifs. Deuxièmement, ces échecs abusifs touchent principalement les étudiantes. Compte tenu de leur niveau de maîtrise de la matière tel que mesuré dans l'examen sans pénalités, elles devraient obtenir de bien meilleures performances que celles de leur pairs masculins dans l'examen avec pénalités. Ce n'est pas du tout le cas puisque leur taux de réussite est plus faible que celui des garçons sans pour autant que cette différence soit significative.

Ce type de résultats est généralement avancé dans la littérature pour remettre en question l'hypothèse du biais de genre. Puisqu'il n'y pas de différence significative, il n'y a pas non plus de biais. L'intérêt de notre recherche est de montrer que lorsque le niveau de maîtrise de la matière est vérifié par ailleurs, les filles sous-performent de manière très importante dans l'examen avec pénalités au point qu'on peut considérer que ce dernier porte atteinte à l'égalité des chances entre les femmes et les hommes. Le diable se cache dans le détail : des différences minimales en termes d'abstentions cachent en réalité un biais de genre, hypothèse avancée dans plusieurs études précédentes, mais sans qu'on puisse la vérifier expérimentalement.

Tel est l'apport de notre travail qui confirme ainsi la nécessité d'interdire l'usage des pénalités dans les examens par QCM comme ce fut le cas de plusieurs universités européennes ces dernières années². Plus rien ne semble justifier qu'on fasse appel aux points négatifs dans les évaluations certificatives ou dans les examens d'admission de nos étudiant·es.

Références

Akyol, P., Key, J., & Krishna, K. (2022). Hit or Miss? Test taking behavior in multiple choice exams. *Annals of Economics and Statistics*, 147, 3-50. <https://doi.org/10.2307/48684785>

² L'Université de Gand (dès 2014), l'École de biologie de l'University College Dublin (depuis 2017), l'UCLouvain (première université francophone) en 2020. Des changements de politique liés aux notes négatives ont également eu lieu dans le cadre des examens d'entrée comme par exemple pour le SAT, examen d'entrée des universités américaines (depuis 2016).



- ARES [Académie de recherche et d'enseignement supérieur] (2022). *Taux de réussite en 1re année de bachelier des étudiants de 1^{re} génération suivant le secondaire suivi*. <https://www.ares-ac.be/fr/statistiques/indicateurs#03-taux-de-reussite-en-1re-année-de-bachelier-des-étudiants-de-1re-génération-suivant-le-secondaire-suivi>
- Arthur, N., & Everaert, P. (2012). Gender and performance in accounting examinations: Exploring the impact of examination format. *Accounting Education*, 21(5). <https://doi.org/10.1080/09639284.2011.650447>
- Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science*, 60(2), 434-448. <https://doi.org/10.1287/mnsc.2013.1776>
- Bar-Hillel, M., Budescu, D., & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society*, 4, 3-12. <https://doi.org/10.1007/s11299-005-0001-z>
- Braibant, J.-M. (2015, 6 octobre). *QCM : problématiques et amélioration* [communication orale]. Université Paris-Saclay, France. https://www.universite-paris-saclay.fr/sites/default/files/media/2021-02/qcm.jm_braibant.pdf
- Brogniaux, A. (2021). *Gender bias in learning evaluation : Negatively marked multiple choice questions* [mémoire de maîtrise inédit]. Université Libre de Bruxelles.
- Chica J.C., & Tarrago, M.J.G. (2021). Analysis of the reliability of multiple-choice tests with the Monte Carlo method. *Revista de Education*, 392, 59-89. <https://doi.org/10.4438/1988-592X-RE-2021-392-479>
- Cipriani, G.P. (2018). Gender difference in willingness to guess after a failure. *The Journal of Economic Education*, 49(4), 299-306. <https://doi.org/10.1080/00220485.2018.1500958>
- Coffman K.B., & Klinowski, D. (2019). The impact of penalties for wrong answers on the gender gap in test scores. *Proceedings of the National Academy of Sciences*, 117(16). <https://doi.org/10.1073/pnas.1920945117>
- De Ketele, J.-M. (2010). Ne pas se tromper d'évaluation. *Revue française de linguistique appliquée*, 15(1), 25-37.
- Dejonckheere, D. (2017). *The effect of gender on multiple choice exams: retrospective correcting for guessing*. [mémoire de maîtrise, Université Gent]. https://lib.ugent.be/fulltxt/RUG01/002/350/967/RUG01-002350967_2017_0001_AC.pdf
- Ecalte, J. (2010). L'évaluation de la lecture et des compétences associées. *Revue française de linguistique appliquée*, 15(1), 105-120.
- Espinosa, M. P., & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology*, 54(5), 415-425. <https://doi.org/10.1016/j.jmp.2010.06.001>
- Funk, P., & Perrone, H. (2016). *Gender differences in academic performance: The role of negative marking in multiple-choice exams*. Center for Economic and Policy Research Working Paper, 12. https://www.researchgate.net/publication/311739762_Gender_Differences_in_Academic_Performance_The_Role_of_Negative_Marking_in_Multiple-Choice_Exams
- Gerard, F.-M. (2017). Objectiver la subjectivité. Dans D. Leduc & S. Béland (Eds.). *Regards sur l'évaluation des apprentissages en arts à l'enseignement supérieur* (Tome 1). Presses de l'Université du Québec, 29-47.
- Iriberry, N., & Rey-Biel, P. (2021). Brave boys and play-it-safe girls: Gender differences. In Willingness to guess in a large scale natural experiment. *European Economic Review*, 13, 103603. <https://doi.org/10.1016/j.euroecorev.2020.103603>
- Kacprzyk, J., Parsons, P., Maguire, P.B., & Stewart G.S. (2019). Examining gender effects in different types of undergraduate science assessment. *Irish Educational Studies*, 38(4), 467-480. <https://doi.org/10.1080/03323315.2019.1645721>
- Laveault, D., & Grégoire J. (2014). *Introduction aux théories de tests en psychologie et en sciences de l'éducation* (3^e éd). De Boeck Supérieur.



- Laveault, D., & Yerli, G. (2017). Modération statistique et modération sociale des résultats scolaires : approches opposées ou complémentaires ? *Mesure et Évaluation en Éducation*, 40(2). 1-26 <https://doi.org/10.7202/1043569ar>
- Lecroart, I. (2016). *L'évaluation certificative par QCM à l'université. La pénalisation des mauvaises réponses est-elle associée à une meilleure discrimination des étudiants ?* [mémoire de maîtrise, UCLouvain]. https://dial.uclouvain.be/downloader/downloader.php?pid=thesis%3A3621&datastream=PDF_02&cover=cover-mem
- Lesage, E., Valcke, M., & Sabbe, E. (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. <http://dx.doi.org/10.1016/j.stueduc.2013.07.001>
- Marín, C., & Rosa-García, A. (2011). *Gender bias in risk aversion: Evidence from multiple choice exams*. Working Paper 39987, MPRA, University Library of Munich. Allemagne. https://mpira.ub.uni-muenchen.de/39987/1/MPRA_paper_39987.pdf
- Merle, P. (2012). L'évaluation par les notes : quelle fiabilité et quelles réformes ? Dans *Regards croisés sur l'économie*, 2(12), 218-230. <https://www.cairn.info/revue-regards-croises-sur-l-economie-2012-2-page-218.htm>
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33-48). Lawrence Erlbaum Associates, Inc.
- Morissette, D. (2005). *Les examens de rendement scolaire*. Les Presses de l'Université de Laval.
- OCDE (2022). *Regards sur l'éducation 2022*. Indicateur A1. Tableau A1.2. Évolution du niveau de formation des 25-34 ans, selon le sexe (2011, 2021), p. 50. <https://www.oecd-ilibrary.org/sites/ea9f1e95-fr/index.html?itemId=/content/component/ea9f1e95-fr>
- Ors, E., Palomino, F., & Peyrache, E. (2008). Performance gender-gap: Does competition matter? *Journal of Labor Economics*, 31(3), 443-499. <https://doi.org/10.1086/669331>
- 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. <https://doi.org/10.1016/j.jebo.2014.08.007>
- Ramos, I., & Lambating, J. (1996). Risk taking: Gender differences and educational opportunity. *School Science and Mathematics*, 96(2), 94-98. <https://doi.org/10.1111/j.1949-8594.1996.tb15816.x>
- Ricketts, C., Brice, J., & Coombes, L. (2010). Are multiple choice tests fair to medical students with specific learning disabilities? *Advances in Health Sciences Education*, 15(2), 265–275. <https://doi.org/10.1007/s10459-009-9197-8>
- Roggemans, L., & Spruyt, B. (2014). *Toelatingsproef (tand)arts : een sociografische schets van de deelnemers en geslaagden*. Vrije Universiteit Brussel. <https://torvub.be/publication/toelatingsproef-tandarts-een-sociografische-schets-van-de-deelnemers-en-de-geslaagden/>
- Rothstein, J.M. (2004). College performance predictions and the SAT. *Journal of Econometrics*, 121(1-2), 297–317. <https://doi.org/10.1016/j.jeconom.2003.10.003>
- Todd, E. (2022). *Où en sont-elles ? Une esquisse de l'histoire des femmes*. Éditions du Seuil.
- Uytendaele, E. (2004). *QCM et points négatifs*. Les fiches-outils du Centre d'appui pédagogique, Université Libre de Bruxelles.

Jean-Marc Braibant

Psychopédagogue et psychologue, il a développé une expertise dans les démarches qualité pour l'enseignement supérieur. Intégré au service QOPA de l'UCLouvain, il est chargé d'évaluer la qualité psychométrique des examens par QCM. Il est membre de la Commission des évaluations de la



François-Marie Gerard

Fédération Wallonie-Bruxelles en tant qu'expert spécialisé dans la docimologie.

Étienne Billat

Psychopédagogue, tour à tour enseignant du primaire au supérieur (non-)universitaire, chercheur, formateur d'adultes, auteur de plusieurs ouvrages, il est intervenu dans 25 pays. Particulièrement intéressé par toutes les problématiques liées à l'évaluation (évaluation de projets, de systèmes, de compétences...), il est actuellement conseiller scientifique de BIEFOR.

Psychopédagogue, il est conseiller pédagogique au Laboratoire Learning Lab de l'UCLouvain où il est chargé de l'offre d'un service de traitement automatisé des QCM. L'application fournit différents rapports dont les notes d'examens. Il évalue aussi la qualité de ces questions sur base des analyses d'items.

