

Appréciations et extraits des débats autour du texte soumis

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 publication : septembre 2023

Évaluation ouverte et collaborative

Rétroacteurs : Sébastien Béland, Jean-Marie De Ketele

Ce fichier présente l'appréciation globale des deux rétroacteurs avant leurs échanges avec les auteurs, puis des extraits des débats qui ont eu lieu autour du texte entre les auteurs et les deux rétroacteurs. Ces extraits ont été choisis d'entente entre tous.

Appréciations globales à propos du texte de la prépublication

Sébastien BÉLAND

Bonjour,

J'ai eu beaucoup de plaisir à lire le manuscrit. Celui-ci devra, par contre, être retravaillé. Voici quelques pistes d'amélioration:

- Il sera important de clarifier la posture de la validité adoptée et proposer une argumentation en conséquence.
- Veuillez clarifier les objectifs de la recherche.
- Il y a des ajouts à faire dans les résultats présentés.
- Je suggère d'ajouter une analyse factorielle pour étudier la dimensionnalité des items.
- Je suggère de retravailler considérablement la section sur le biais de genre à la lumière du fonctionnement différentiel d'item. La version actuelle n'est pas convaincante pour décrire un biais.
- Les résultats obtenus sont peu mis en relation avec les écrits sur le sujet: je suggère fortement de produire une section "Discussion".
- Veuillez proposer des pistes de recherche à explorer dans le futur (la plus importante me semble celle de produire une étude de simulation pour généraliser vos résultats).

Vous allez trouver plus de commentaires sur le Google Doc.
Bonne réflexion!

Jean-Marie De Ketele

Je trouve cet article très important.

Comme membre de la Commission des sages de l'examen d'entrée en médecine, je suis très sensibilisé aux effets pervers des épreuves d'évaluation certificative (ici de sélection ... donc aux enjeux personnels et institutionnels importants). Beaucoup d'écrits ont été faits sur le sujet, mais aucun d'entre eux n'a eu l'occasion de pouvoir faire une démonstration tout à fait probante. Les auteurs de l'article ont eu la chance (rarissime) de pouvoir comparer deux procédures différentes de notation sur des groupes d'étudiants strictement équivalents. En effet, une session d'évaluation certificative pendant la pandémie a dû être annulée suite à des fraudes. Une nouvelle session a été organisée en présentiel dans deux amphithéâtres avec une répartition des étudiants telle que les deux groupes soient équivalents (comme le démontre l'article) et avec une passation de la même épreuve dans le même cadre horaire. La seule différence était la procédure de notation annoncée aux étudiants. On ne peut donc rêver d'un dispositif mieux contrôlé et d'une démonstration aussi convaincante sur les effets différentiels des deux procédures de notation sur (1) globalement les taux de réussite, (2) l'effet genre (une des procédures pénalisant gravement les filles). L'article est pour moi d'une grande clarté didactique et correspond en tous points aux critères de la revue.

A LIRE ABSOLUMENT PAR LES CONCEPTEURS D'ÉPREUVES PAR QCM.
ET DONC À PUBLIER (corrections mineures fournies aux auteurs)

Extraits d'échanges lors de la révision collaborative

1. Introduction

Les questionnaires à choix multiple (QCM) sont devenus – à tort ou à raison – un outil d'évaluation très répandu dans l'enseignement supérieur, surtout en première année. Un avantage des QCM est la possibilité d'évaluer un échantillonnage plus important de la matière que les autres formes d'examen auprès de grands groupes d'étudiants. Un autre avantage est qu'il s'agit en principe d'un outil d'évaluation objectif et fiable puisque toutes les réponses sont évaluées de la même façon. La correction est rapide, simple et précise et elle peut être automatisée par lecture optique ou traitement informatique. Par contre, un écueil fréquent des QCM consiste à poser des questions qui ne portent que sur des points de détails. Cela conduit à un risque de parcellisation des connaissances sollicitant majoritairement la capacité de mémorisation de l'apprenant et la simple restitution de connaissances. Une autre critique importante est de laisser la possibilité à l'étudiant de deviner la bonne réponse sans la connaître, ce qui lui permettrait de gagner des points « immérités ».

Cette dernière question a fait couler beaucoup d'encre au cours des trente dernières années. Depuis de nombreuses années, l'effet du hasard sur les résultats des étudiants dans les examens par questions à choix multiple (QCM) a provoqué de nombreuses discussions. Selon certains auteurs (par exemple, Choppin, 1988 ; Frary, 1988 ; Reisz et al., Hennequin & Langlois, 2016), un candidat qui ne connaît pas la réponse à une question a une probabilité non négligeable de la sélectionner en la devinant. (par exemple, Choppin, 1998). Selon d'autres (par exemple, Haladyna, 1994 ; Morissette, 2005), si l'examen est bien conçu – un nombre suffisant d'items et des leurres pertinents –, 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 (par exemple, Haladyna, 1994, Morissette, 2005)). 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 étudiants ?

Commenté [1]: Sébastien BÉLAND

Il sera important, dans cet article, d'adopter une position sur la validité et argumenter en conséquence de celle-ci. Voir:
<https://sk.sagepub.com/books/validity-and-educational-assessment>

Commenté [2]: Sébastien BÉLAND

Le QCM est un mal-aimé de l'évaluation de connaissances dans mon système éducatif. Je suggère aux auteurs de présenter quelques avantages de ce type d'item dans l'introduction pour donner de la pertinence sociale à votre très intéressant manuscrit



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 fidélité des scores ~~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, Valeke & Sabbe et al. 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'éthique et d'équité, ce type de design expérimental ne peut être mis en œuvre : imposer à des étudiant·es deux formes différentes d'un examen alors qu'il est possible qu'une de ces formes défavorise ceux et celles qui y seront soumis·es est inconcevable. ~~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 d'un examen 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 – c'est-à-dire une absence de réponse – 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. Objectifs de la recherche

On pourrait qualifier notre recherche d'empirique, voire même d'occasionnelle, en ce sens que nous avons saisi une « occasion » qui s'est présentée à nous. Cette opportunité est due à un malentendu au moment de la passation d'un examen : les consignes sur ses modalités de correction ont été différentes dans les deux amphithéâtres, contrairement à ce qui était prévu.

Il est fondamental de comprendre que nous utilisons donc des données d'une situation réelle, non construite à des fins de recherche, mais constitutive d'une situation institutionnelle d'évaluation. L'avantage de pouvoir travailler sur des données réelles nous semble évident, mais cela a aussi des limites. Nous n'avons eu aucun pouvoir d'action sur le choix d'une évaluation par QCM ou non, sur les items qui ont été proposés aux étudiants, sur leur nombre, sur la discipline évaluée, sur la population étudiante concernée, etc. Nous travaillons avec les données telles qu'elles existent et telles qu'elles sont encodées dans le logiciel ConTest de la société Consilium (<https://consiliumdcs.com/en/products/contest/>) utilisé à l'UCLouvain pour gérer les résultats des QCM de toutes les facultés.

N'ayant aucun contrôle sur la construction des items, nous n'avons pas de prise sur sa validité dite de contenu. Nous nous basons sur le postulat (vraisemblable) que l'enseignant qui a réalisé le questionnaire maîtrise parfaitement la matière qu'il enseigne et qu'il a proposé dans son examen des questions représentatives du domaine enseigné en vue de certifier les acquis d'apprentissage de ses étudiant-es. Remarquons au passage que si la validité de construit n'était pas présente, cela ne changerait rien à la réalité empirique étudiée, celle d'un examen universitaire administré en situation réelle et de son impact sur le cursus des étudiants.

C'est bien à ce dernier niveau que se situe notre objectif de recherche. Il se décline lui-même en deux questions importantes : quelle est la validité prédictive des deux formes d'examen et qu'en est-il de leur validité de conséquences ?

Dans un premier temps, après avoir réalisé une étude d'équivalence des deux échantillons et des qualités psychométriques des deux formes d'évaluation, nous réaliserons une analyse comparative des scores de l'examen avec ou sans points négatifs avec les résultats obtenus par les étudiants à l'issue de leur programme de première année universitaire.

Dans un second temps, nous vérifierons si les deux examens ont des conséquences semblables ou différentes auprès des femmes et des garçons. La question du biais de genre fait débat depuis une trentaine d'années et nous avons ici l'opportunité d'apporter notre contribution pour trancher cette question.

La validité des conséquences a souvent été négligée. Reprenant les travaux de Messick consacrés à ce sujet, Laveault et Grégoire (2014) rappellent que « la finalité des scores est de servir de base à des prises de décision et leur valeur ultime provient de leur impact individuel et social ».

La question des conséquences de l'usage des tests est en réalité analogue à celle soulevée à propos des médicaments. Les chercheurs ne doivent pas seulement montrer que l'outil qu'ils utilisent traite bien ce qu'il doit traiter (ou mesure bien ce qu'il doit mesurer), mais ils doivent aussi évaluer les effets secondaires ou indésirables, en informer les utilisateurs et fournir des indications sur la manière de les minimiser. C'est pour pouvoir répondre à ces questions que nous avons élaboré notre méthodologie expérimentale.



2.3. Population de départ et procédure d'échantillonnage

...

3.4. Étude de l'équivalence des deux échantillons

Commenté [3]: Sébastien BÉLAND

Il serait utile, pour la forme, de produire des objectifs de recherche à la fin de l'introduction

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.

3.14.1 Les variables d'entrée

3.1.14.1.1 Âge

...

3.1.24.1.2 Expérience à l'université

...

3.1.34.1.3 Type d'enseignement suivi dans l'enseignement secondaire

...

3.1.44.1.4 Genre

...

3.24.2 Les variables de sortie

...

4.5. Présentation de l'examen

...



5.6. Comparaison des 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'assurer que le questionnaire administré dans deux groupes différents présente les mêmes niveaux de difficulté, de discrimination et de fidélité. Ce sont les trois indicateurs identifiés par Beller, Gafni et al. (1999) pour déterminer l'équivalence de deux tests, indicateurs auxquels nous avons ajouté l'erreur standard de mesure. Nous avons ajouté l'erreur (1999) pour présenter les mêmes niveaux de difficulté, de discrimination et de fidélité, qualités psychométriques équivalentes, des deux questionnaires s'arrêter sur les qualités psychométriques du questionnaire afin de vérifier si. 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).

Commenté [4]: Sébastien BÉLAND

Il serait important d'ajouter une analyse de la dimensionnalité des réponses à l'examen. En gros: peut-on dire que les items mesurent bien la dimension/construit/trait latent qu'ils devraient mesurer?

Tableau 6

Indicateurs psychométriques pour les deux évaluations

Indicateurs	Questionnaire A (Sans pénalités)	Questionnaire B (Avec pénalités)
Indicateur de fidélité (alpha de Cronbach) ¹	0,71	0,70
Indicateur de discrimination (Rir) ²	0,24 (0,17)	0,23 (0,11)
Indicateur de difficulté (p-index) (p-index) ³	0,48 (0,16)	0,40 (0,13)
Erreur standard de mesure	2,22	2,21

Commenté [5]: Sébastien BÉLAND

Nous ne savons pas si il y a plusieurs construits/dimensions/traits latents dans cet examen: est-ce le cas? Si oui, il faudra rapporter ces informations par construits/dimensions/traits latents, car elles ont été pensées en contexte d'unidimensionnalité

Commenté [6]: Sébastien BÉLAND

Le coefficient alpha est l'objet de très nombreuses et légitimes critiques. Je vous suggère d'ajouter l'oméga de McDonald, disponible dans le logiciel R et JASP

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.

¹ L'alpha de Cronbach est la méthode la plus connue d'estimation de la précision (la stabilité par consistance interne) du score total d'un test. PCes dernières années, plusieurs auteurs ont fait valoir des critiques contre l'usage de ce coefficient et suggèrent de le compléter par l'oméga de McDonald (Béland et al., -Cousineau & Love, 2017). Le logiciel ConTest ne fournit pas cet indicateur et c'est la raison pour laquelle nous ne l'avons pas repris dans nos analyses. Comme Rappelons aussi que notre objectif n'est pas de valider un instrument de mesure, mais bien de s'assurer que les deux examens présentent les mêmes qualités psychométriques, ce qui est le cas pour l'alpha de Cronbach en tant qu'estimation de leur fidélité respective est suffisant.

² L'indice de discrimination est un indice important, car il permet de vérifier que le test que l'on utilise remplit bien sa fonction : départager à bon escient les étudiant-es qui ont eu un score total élevé par rapport à ceux qui ont eu un score total faible. Dans cette perspective, une "bonne" question est une question qui serait réussie par une plus grande proportion des répondant-es ayant obtenu un score élevé à l'examen que par celles/ceux ayant obtenu un score faible. L'indice de discrimination repose donc sur l'idée suivante : un sujet qui a répondu correctement à un item présente une plus grande probabilité d'obtenir un résultat élevé au test que celui qui a échoué. L'indice Rir de ConTest repose sur le calcul d'une corrélation entre le score à chaque question et le score total à l'examen en retirant la question-cible de ce calcul. Le Rir mentionné dont il est question ici est obtenu en faisant la moyenne des indices de discrimination de toutes les questions de l'examen (avec l'écart-type entre parenthèses).

³ L'indice de difficulté ou p-index est simplement le rapport entre le nombre de personnes qui réussissent l'item et le nombre de personnes qui l'ont passé. (p x 100 donne donc directement le pourcentage de réussite à l'item). Cet indice de difficulté varie entre 0 et 1 (0 signifiant qu'un item est systématiquement échoué [0% de réussite] et à l'inverse 1 [100% de réussite] signifie qu'il est systématiquement réussi. Le p-index moyen de l'examen calculé ici par ConTest est obtenu en faisant la moyenne des indices de difficulté de toutes les questions de l'examen.



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 ($p\text{-index}=,48$ vs $p\text{-index}=,40$). Il est important de souligner que dans les deux cas, la valeur $p\text{-index}$ 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 $p\text{-index}$ 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 administré dans les deux amphithéâtres est bien calibré même si le nombre d'items pourrait être plus élevé (Braibant, 2015).

6.7. Résultats obtenus par les étudiant-es dans les deux examens

6.17.1 Statistiques descriptives

...

6.27.2 Des stratégies de réponses différentes

...

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

7.18.1 Méthodologie d'analyse

...

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

...



8.9. 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, ~~Budescu & Attali~~, 2005 ; ~~Budescu & Bar-Hillel~~, 1993 ; 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 & Gardezabal, 2010). Ce biais sera d'autant plus grand que l'on pénalise fortement une mauvaise réponse (Lecroart, 2016).

Selon ~~plusieurs~~ certaines études, ~~le~~ le recours à la pénalisation des mauvaises réponses pousse ~~fa~~ les étudiant·es très faibles à la prise de risque : ils n'ont rien à perdre et parfois ils peuvent gagner ~~↓~~ (pour une revue de la littérature, voir Betts et al., 2009 ; Chica & Tarrago, 2021 ; Lesage et al., Valcke & Sabbe, 2013). Ce système inciterait par contre de bon·nes 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 (~~Bliss~~, 1980 ; Uytbroeck, 2014). À ce « jeu », les femmes dont il est admis qu'elles sont des "risk adverse" dans de nombreux contextes (~~Badiga~~, 2014 ; Cipriani, 2018 ; Spell & Bezrukova, 2010) 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 (Brognaux, 2021 ; Roggemans & Spruyt, 2014). D'autres études nuancent ces conclusions quand elles ne les remettent pas en question (Akyol ~~et et~~ al., 2022 ; ~~Bond et al.~~, 2013 ; Funk ~~& &~~ Perrone 2016, 2023 ; Ricketts ~~et et~~ al., 2010). Ce manque de consensus s'explique notamment par le fait qu'il n'est éthiquement 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 (~~Laveault et Grégoire~~, 2014) – (Reynolds, 1982). 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 (~~Laveault & Grégoire~~, 2014). – Il faut également que les individus soient extraits de la population selon une procédure aléatoire et simple pour s'assurer de l'équivalence des échantillons. Les circonstances dans lesquelles s'est déroulé notre examen remplit cette dernière condition et nous donne l'occasion de tester l'hypothèse d'un biais de genre. À notre connaissance, une telle étude est inédite. 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.

Commenté [7]: Sébastien BÉLAND

La question du biais de genre, en psychométrie, a une définition précise. Pourriez-vous consulter la page 12 du lien qui suit pour être certain que c'est ce que vous faites, ici ?

<https://faculty.educ.ubc.ca/zumbo/DIF/handbook.pdf>

Je ne suis pas satisfait de cette section, collègues. Nous ne pouvons pas étudier de façon précise la question du biais à partir du score: il faudra plutôt passer par l'étude détaillée des items, ce que nous nommons du fonctionnement différentiel d'item. Voici la définition proposée par Zumbo:

DIF occurs when examinees from different groups show differing probabilities of success on (or endorsing) the item after matching on the underlying ability that the item is intended to measure.

Commenté [8]: François-Marie GERARD

Nous ne pouvons répondre à votre demande de réaliser une analyse différenciée des items.

Outre des difficultés techniques liées au logiciel ConTest, non prévu pour cela, réaliser une analyse de la dimensionnalité des réponses pose par ailleurs deux difficultés importantes :

- selon les formulaires papier utilisés (ce qui exclut d'ailleurs les questionnaires adaptatifs), les items n'ont pas été présentés aux étudiants dans le même ordre (pour éviter que ceux qui sont proches l'un de l'autre dans l'amphithéâtre puissent copier les réponses de leurs voisins). Il pourrait donc y avoir un biais lié à l'ordre des items : on sait que les derniers items sont souvent traités différemment, par manque de temps, par fatigue, par désinvestissement ou bien, dans le cas particulier des points négatifs, lorsque l'étudiant estime qu'il n'est pas nécessaire de répondre aux dernières questions de son examen, car il pense (à tort ou à raison) avoir accumulé suffisamment de points en répondant aux items qui précèdent.

- la plupart des méthodes de TRI nécessite de dichotomiser les résultats, ce qui est en contradiction avec la triple modalité "bonne réponse - abstention (non-réponse) - réponse incorrecte" au cœur de notre analyse. Certes, certaines modélisations comme le rating scale model ou le partial credit model le permettent, mais dans l'état actuel des données en notre possession ces analyses ne sont pas possibles.

Nous estimons que l'analyse des scores globaux est suffisante dans le cadre de ce travail pour analyser l'impact des deux modalités de notation des QCM concernés et pour mettre en évidence l'éventuel biais de genre. Il est néanmoins possible qu'une analyse de la dimensionnalité des réponses permettrait d'affiner les résultats. Il s'agit pour nous d'une piste qui serait à investiguer pour des recherches futures, à proposer aux chercheurs intéressés (les données telles qu'elles existent sont disponibles sur simple demande).



8.19.1 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 $<,001$; d de Cohen= $0,822$) que pour leurs condisciples masculins ($t=3,778$; p bilatéral $<,001$; d de Cohen = $0,781$). 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).

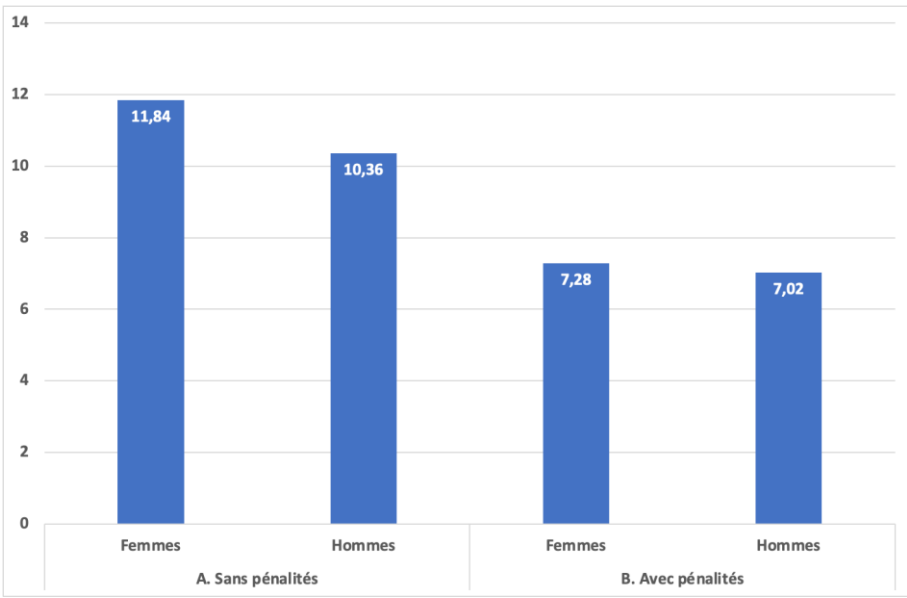
Tableau 14

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

Figure 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= $0,02$; d de Cohen = $0,47$) alors que ce

Commenté [9]: Sébastien BÉLAND

Les analyses produites dans cette section sont instructives, mais il existe une stratégie bien plus intéressante: le fonctionnement différentiel d'item (differential item functioning). La raison pourquoi cette approche est supérieure est qu'elle permet d'étudier les réponses à chacune des questions. Je crois qu'à ce moment-ci, il serait important d'aller plus loin en utilisant par exemple l'approche de Mantel-Hanszel ou par la régression logistique. Le problème avec votre proposition et qu'elle est "poolée" et demanderait de nombreuses nuances.



n'est pas le cas dans l'examen avec pénalités ($t=0,316$; p bilatéral= $0,75$; d de Cohen = $0,06$). Ce résultat est important, car il met clairement en évidence l'existence d'un biais à la défaveur des femmes.

8.29.2 Comparaison des deux groupes en pourcentages de réussite

...

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

...

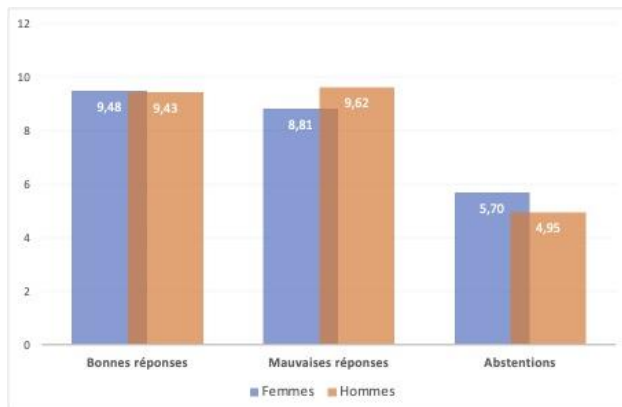
8.49.4 Comment expliquer ce biais de genre ?

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



Commenté [10]: Sébastien BÉLAND

Cette section est fort intéressante, mais il faudra étudier plus en détail les réponses aux items (plutôt que les scores totaux) pour vous prononcer sur la question du biais. J'attends des informations sur le fonctionnement différentiel d'item



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 elles 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

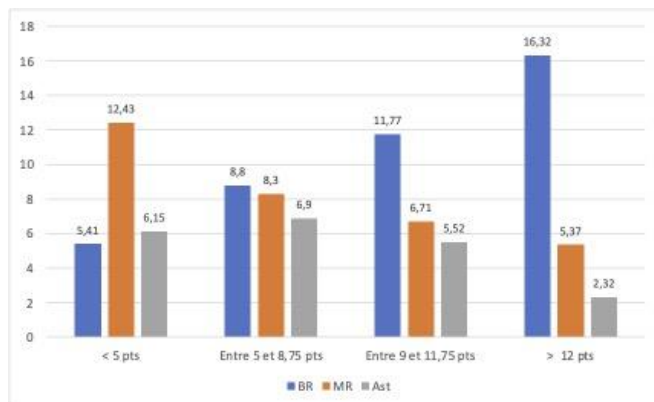
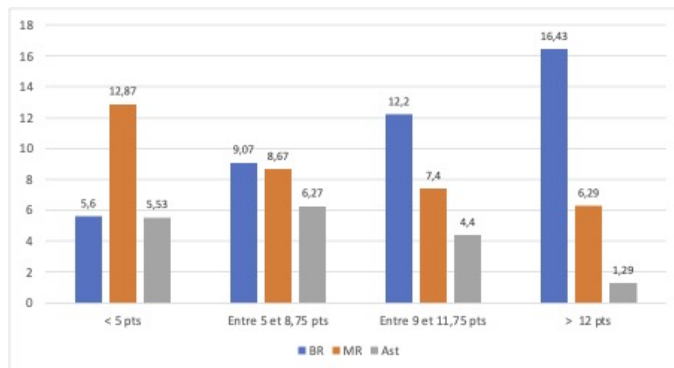


Figure 9

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 sont présentés dans le tableau 18.

Tableau 18

Répartition du type de réponses des étudiantes proches de la réussite

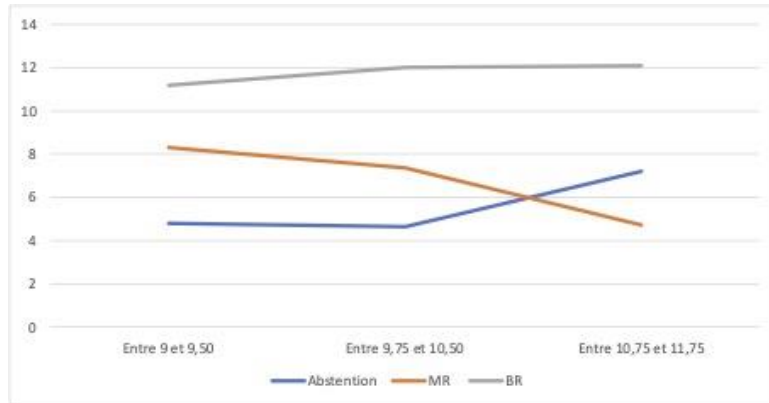
	Entre 9 et 9,50 (N=10)	Entre 9,75 et 10,25 (N=11)	Entre 10,50 et 11,75 (N=10)
Bonnes réponses	11,2	12	12,1
Mauvaises réponses	8,0	7,36	4,7
Abstentions	4,8	4,64	7,2

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-elles 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 ; d de Cohen = 0,536) ainsi qu'entre le 2^e et le 3^e groupes ($t = -1,999$; p bilatéral = ,04 ; d de Cohen = 0,434). Il n'y a pas de différence par contre entre le 1^{er} et le 2^e groupes ($t = -0,442$; $p = ,66$; d de Cohen = 0,093).

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).*

9.10. Conclusion

Cette recherche présente dès le départ un caractère particulier : elle ne découle pas d'un design expérimental élaboré pour vérifier des hypothèses de départ, mais elle résulte d'un "incident" qui s'est déroulé lors d'un examen officiel basé sur des questions à choix multiple. Deux consignes différentes ont été données dans les deux groupes d'étudiant-es soumis au même questionnaire. Dans un cas, il était question d'un "examen sans pénalités" (ce qui était la modalité prévue) alors que dans l'autre, les encadreurs ont parlé d'un "examen avec pénalités", ce qui signifie qu'une non-réponse est peu pénalisée au contraire d'une mauvaise réponse. Impliqués à différents titres dans les problématiques d'évaluation des acquis, nous avons été interpellés par cette situation et avons cherché à savoir si les résultats de cet examen étaient différents selon ces deux modalités, question abordée dans de nombreuses autres recherches (##quelques références##) sans réponse définitive (-Lesage, Valcke & Sabbe, 2013).

Le traitement rigoureux de ces résultats met en évidence deux éléments 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 constat est généralement avancé dans la littérature pour remettre en question l'hypothèse du biais de genre. Puisqu'il n'y pas de différence significative, certains chercheurs concluent qu'il n'y a pas non plus de biais (Bond & al, 2013 ; Funk & Perrone, 2016, 2023) ##références de ces chercheurs##. 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.

~~Ce sont avant tout le diable se cache dans le détail :~~ des différences minimes en termes d'abstentions ~~qui permettent d'expliquer ce cachent en réalité un~~ biais de genre : les



étudiantes ont plus tendance à s'abstenir en ne prenant pas de risques, alors même qu'il est vraisemblable, comme le montre l'examen sans pénalités, que si elles répondaient, elles donneraient la bonne réponse. Cette hypothèse a déjà été avancée dans plusieurs études précédentes (Brogniaux, 2021 ; Roggemans & Spruyt, 2014)###références###, mais sans qu'on puisse la vérifier expérimentalement, de manière aussi que, ce que les circonstances "accidentelles" à la source de notre étude permettent de faire dans une situation réelle.

Tel est l'apport de notre travail qui incite plus que jamais à confirmer ainsi la nécessité à d'interdire l'usage des pénalités dans les examens par QCM, comme ce fut le cas de plusieurs universités européennes l'ont d'ailleurs fait ces dernières années⁴. Sur la base de notre étude, le recours à des QCM à 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 devrait être absolument évité.

Il faudrait bien sûr pouvoir répliquer notre étude sur d'autres étudiant-es, sur d'autres questionnaires, sur d'autres matières, etc. Pour nous, l'éthique scientifique rend cette réplication quasiment impossible, ou du moins inacceptable. Soumettre volontairement dans une situation réelle d'évaluation certificative des groupes d'étudiant-es à deux modalités différentes d'examen alors qu'on sait, grâce à notre étude, que la modalité "examen avec pénalités" défavorise nettement ceux qui y sont soumis par rapport à la modalité "examen sans pénalités" et en particulier les étudiantes ne serait pas recevable. On ne peut en effet pas « tromper/truquer » les étudiant-es dans une situation d'évaluation réelle en leur faisant croire une fausse information, puis ensuite leur demander de passer un autre test.

Des études de simulation seraient possibles, mais elles nécessiteraient de ne pas s'intégrer dans une réelle situation d'évaluation certificative et que les étudiant-es sachent qu'ils participent à une situation expérimentale. L'impact de ces deux exigences serait tel que les résultats seraient difficilement comparables à ceux que nous avons obtenus dans cette situation réelle involontaire.

Il est possible également d'analyser de manière plus approfondie et/ou avec une autre approche les données dont nous disposons. Ces données sont accessibles à toute personne qui nous solliciterait pour approfondir nos analyses. Une piste de travail pourrait être de traiter les données par une analyse différenciée des items. Nous n'avons pas estimé nécessaire de faire une telle analyse dans le cadre de notre étude, car les résultats sur les scores globaux nous permettaient de répondre à nos questions de recherche. Néanmoins, il est possible qu'une analyse différenciée des items permettrait de mieux comprendre comment le biais de genre mis en évidence est à l'œuvre non seulement de manière globale, mais aussi au sein même des items. Une difficulté à gérer serait alors liée au fait que pour des raisons de sécurité lors de la passation de l'examen, les items proposés aux étudiant-es ne sont pas toujours dans le même ordre. Or, on sait que les derniers items sont souvent traités différemment, par manque de temps, par fatigue, par désinvestissement ou bien, dans le cas particulier des points négatifs, lorsque l'étudiant-e estime qu'il n'est pas nécessaire de répondre aux dernières questions de son examen, car il-elle pense (à tort ou à raison) avoir accumulé suffisamment de points en répondant aux items qui précèdent. À nouveau, nos données sont à disposition, même si leur état actuel, géré par le logiciel ConTest, ne permet pas d'avoir nécessairement toutes les informations indispensables pour une analyse différenciée des items.

⁴ 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).



Nous encourageons de tels prolongements, pour autant qu'ils contribuent à répondre aux questions fondamentales qui nous ont guidés dans notre recherche : quelle est la validité prédictive des deux formes d'examen et qu'en est-il de leur validité de conséquences ?

*

Commenté [11]: Sébastien BÉLAND

Vous savez probablement aussi bien que moi que la recherche est une entreprise collective. Or, je suis étonné de ne pas vous voir relancer la communauté de recherche sur d'autres études à conduire pour aller plus loin. Ainsi, après votre étude empirique, je crois que la prochaine étape en sera une où il faut chercher à généraliser avec une étude de simulation, non? J'aimerais vous voir détailler davantage de pistes de recherche à suivre dans le futur. Mille mercis !

