Mise à l'épreuve de méthodes de détection de patrons de réponses suspects

Conference Paper · March 2019

CITATIONS
0

READS
149

3 authors:

Alhassane Aw
Chambre de commerce et d'industrie de Paris Île de France
9 PUBLICATIONS **26** CITATIONS
SEE PROFILE

Marc Demeuse
Université de Mons
417 PUBLICATIONS **1,800** CITATIONS
SEE PROFILE

Dominique Casanova
Chambre de commerce et d'industrie de Paris Île-de-France
39 PUBLICATIONS **31** CITATIONS
SEE PROFILE

Mise à l'épreuve de méthodes de détection de patrons de réponses suspects

Aw Alhassane, aaw@cci-paris-idf.fr, CCI Paris Ile-de-France, France
Zhang Adrien[1], adrien.zhang@eleve.ensai.fr, Ecole nationale de la statistique et de l'analyse d'information, France
Casanova Dominique, dcasanova@cci-paris-idf.fr, CCI Paris Ile-de-France, France
Demeuse Marc, marc.demeuse@umons.ac.be, Université de Mons, Belgique

---

Lorsque la passation d'un test présente un enjeu fort, il peut être tentant pour certains candidats de recourir à des stratagèmes destinés à fausser l'évaluation de leur compétence. Lors des tests de placement à l'université, cela peut être fait dans le souci d'être orienté dans un cours de niveau inférieur afin de s'assurer l'obtention de crédits universitaires (Raîche et Blais, 2003). Pour des tests de langue utilisés dans le cadre de l'accès à un territoire, le but sera d'obtenir une surévaluation de la compétence testée.
Dans ce dernier cas, il faut distinguer les stratégies de préparation licites (cours de renforcement avant le test, entraînement au moyen d'annales ou de matériel pédagogique public) des tentatives de triche. Ces dernières peuvent notamment consister à copier sur un voisin lors de l'examen ou à se procurer de matériel de test confidentiel diffusé illégalement (délit d'initié). De tels comportements remettent en cause le principe d'équité entre les candidats et portent atteinte à l'intégrité des résultats. Il est donc important de prévenir et détecter ces fraudes, afin de garantir aux institutions utilisatrices du test la validité des résultats délivrés.
Les centres d'examen sont un maillon essentiel du dispositif de sécurité, mais il leur est difficile de déjouer certaines stratégies, comme le fait qu'un candidat ait mémorisé des réponses à des questions avant le test. Les développements technologiques offrent également de nouvelles possibilités aux fraudeurs, qui sont parfois difficiles à contrecarrer. Les éditeurs de tests doivent donc mettre en place un dispositif de monitorage des sessions et analyser les réponses des candidats pour détecter d'éventuelles situations problématiques, susceptibles d'avoir échappé à la vigilance des centres de passation.

Pour les épreuves constituées d'items à réponse automatique, où les candidats ont à choisir entre plusieurs options de réponse prédéfinies, il s'agit en général de calculer des indices statistiques pour détecter les patrons de réponses aberrants (Magis & al., 2015 ; Teindeiro et al., 2016). Ces indices quantifient la normalité d'une série de réponses d'un candidat. Ainsi, pour un candidat donné, un indice trop éloigné des indices habituels sera source de suspicion. De nombreux indices ont été proposés pour la détection de patterns aberrants, et diverses études comparatives de leur puissances ont été entreprises (Artner, 2016 ; Béland, 2016 ; Huang, 2012).
Il convient toutefois de distinguer, parmi l'ensemble des patrons de réponses aberrants détectés, ceux qui sont le plus susceptibles de caractériser une tentative de fraude. En effet, un patron de réponse pourra être détecté comme aberrant du fait qu'un candidat faible

[1] Cette communication fait suite à un stage de deuxième année réalisé par Adrien Zhang au sein de la Chambre de commerce et d'industrie de région Paris Ile-de-France.

réponde fréquemment au hasard, ou qu'un candidat de niveau de compétence élevé manque d'attention dans ses réponses à des questions élémentaires. Si la passation s'effectue au format papier-crayon, cela peut également être dû à un simple décalage d'une ligne dans les réponses reportées par le candidat sur la fiche de réponses.

Dans cette étude, nous nous sommes intéressés plus spécifiquement à l'efficacité de tels indices pour la détection de tentatives de fraude visant à améliorer la performance des candidats, à partir de données simulant plusieurs cas de figure. L'objectif était d'identifier un sous-ensemble d'indices pertinents pour une application systématique sur les données de passation d'un test d'évaluation de français diffusé à l'international (le TEF). Nous avons tenu compte des conditions d'exploitation de ce test, pour lequel l'acquisition de données résultats pour une même version s'effectue dans la durée et qui s'appuie sur des paramètres d'items prédéterminés pour délivrer les résultats aux candidats.

Détection de patterns de réponses aberrants

Modélisation par simulations et identification des méthodes les plus efficaces

Nous nous sommes appuyés pour nos travaux sur la librairie Perfit de R (Teindeiro et al., 2016), qui permet le calcul d'une variété d'indices pour la détection de patterns aberrants. Ces indices quantifient la normalité d'une série de réponses d'un candidat. Le patron des réponses d'un candidat dont l'indice est trop éloigné de l'indice de référence sera considéré comme improbable. C'est notamment le cas si le candidat répond correctement à diverses questions difficiles alors qu'il échoue par ailleurs à des questions plus faciles. Pour déterminer les tests statistiques les plus efficaces, nous étudions leur risque d'erreur et leur puissance statistique dans des situations simulées représentatives des types de fraudes que nous souhaitons pouvoir mettre en évidence.

Analyse et résultats

Dans une première analyse, nous avons simulé une situation avec 100 fraudeurs parmi 2500 candidats qui ont pris part à un examen fictif comportant 60 questions, en les faisant tous tricher sur les mêmes questions (simulant ainsi une situation de divulgation illicite d'un ensemble d'items auprès de candidats). Nous avons fait varier le nombre de questions faisant l'objet de triche entre 1 et 45, en limitant les risques de triche aux 45 questions les plus difficiles parmi les 60. Les résultats montrent que pour une proportion de fraude de 40% (24 questions sur 60), certaines méthodes comme Cstar sont assez puissantes (Cf. figure 1) : elles détectent bien la fraude dans 90% des cas environ. Pour des situations très frauduleuses (plus de 60% des items), certaines statistiques perdent en puissance. C'est le cas des modèles lz ou D.KB. La raison la plus probable est que ces méthodes détectent mal les candidats ayant un très bon score. En effet, si des candidats ont triché sur 75% des questions, alors ils ont au moins 75% de bonnes réponses : l'estimation de leur compétence sera de fait biaisée et donc la fraude, plus difficile à détecter[2]. Rappelons que les comportements aberrants se caractérisent principalement par de mauvaises réponses aux

[2] Snijders (2001) a proposé une correction (lz*, lzstar dans cet article) pour corriger la moyenne et la variance de la distribution de l'indice lz afin de neutraliser ce biais entre compétence réelle et compétence estimée (voir notamment Magis et al., 2012 ; Béland et al., 2016).

questions faciles et de bonnes réponses aux questions difficiles. La puissance des méthodes étudiées est donc d'autant plus élevée que le comportement des candidats vis-à-vis du test est aberrant.
La pente des courbes de puissance est la plus élevée entre les proportions 20 et 40% de questions fraudées (9 à 18 questions) pour quasiment toutes les méthodes. Cela indique que les statistiques sont assez sensibles au nombre de questions dévoilées et à leur difficulté. Les puissances de ces statistiques sont toutefois variables selon la méthode. Dans cette configuration, la méthode C* ou Cstar (courbe noire) est celle qui, en moyenne, est la plus performante en termes de détection des cas aberrants. La statistique « lz » étant puissante jusqu'aux environs de 30% des questions dévoilées avant de chuter au-delà de cette proportion.

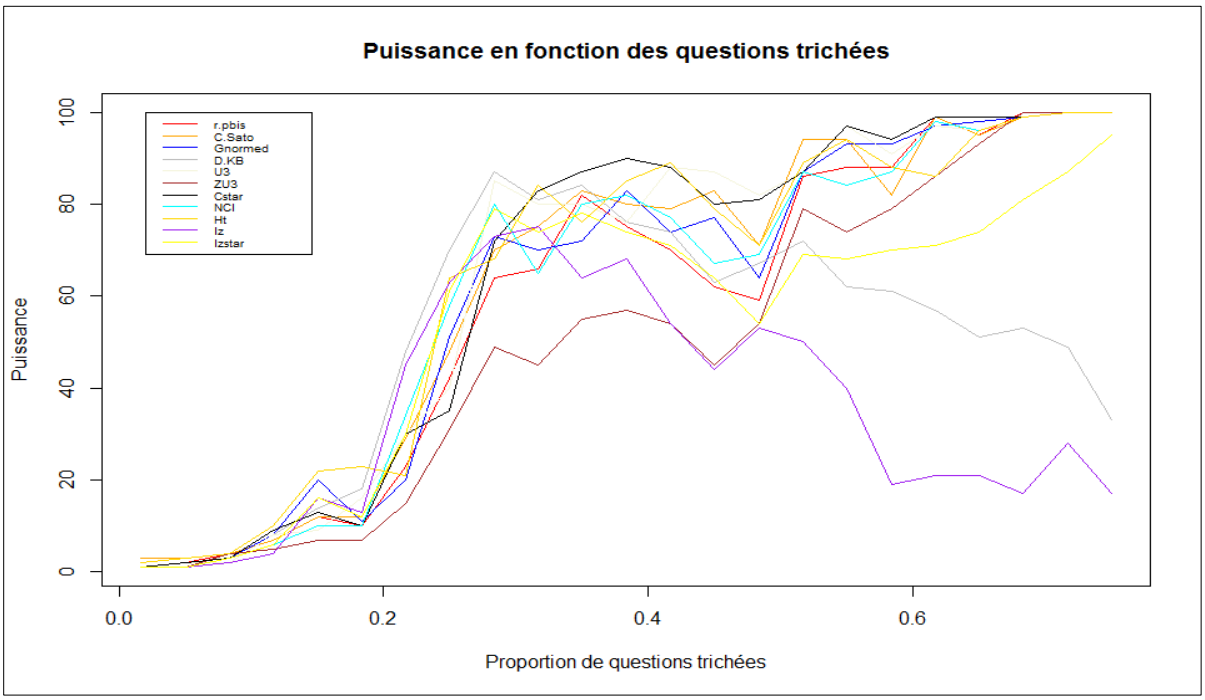

Figure 1 : Puissances (en %) de différents modèles en fonction de la proportion de questions trichées (seules les 75% des questions les plus difficiles sont susceptibles d'être fraudées)

Dans d'autres configurations, comme par exemple le cas où seules les 15 questions les plus difficiles sont sujettes à fraude, les indices de tests sont logiquement plus performants. Lorsque les candidats fraudent uniquement sur les questions les plus difficiles, ils sont plus facilement détectés. Ainsi, en fraudant sur 15% des questions (9 questions sur 60) qui font parties des 25% les plus difficiles, le taux de détection est d'environ 50% pour certains indices comme Ht ou lzstar, et atteint 90% lorsque la triche concerne les 25% (15 sur 60) des questions les plus difficiles.
Les performances des différentes méthodes dépendent de la nature de la fraude et de son étendue. Une connaissance a priori des items les plus susceptibles d'être sujets à la fraude permettrait d'optimiser le choix des méthodes à mettre en œuvre pour mieux s'adapter à la situation. Des questions sont-elles effectivement apprises par cœur et partagées ? Les réponses des questions les plus difficiles sont-elles vraiment les plus susceptibles d'être apprises avant le test ?

Modélisation de la détection de triches par copie sur un candidat

Méthode

La triche par copie sur un autre candidat est un cas particulier de fraude où des informations complémentaires peuvent être prises en considération pour évaluer les risques. Elle implique en effet la présence de deux acteurs : un candidat, dit copieur, recopie certaines réponses d'un autre candidat, dit fournisseur, qui peut délibérément être consentant ou non. L'enjeu est de déterminer si un couple de candidats ayant pris part au test a triché ou pas. La similitude de leurs réponses est quantifiée par un indice de test. Si cet indice est trop éloigné des indices standards, alors on considère que les réponses des deux candidats constituent un cas de triche. L'idée générale est que deux candidats qui répondent correctement à une même série de questions, et incorrectement à une autre, seront fortement suspectés de triche. L'indice utilisé est le suivant (Romero et al., 2015) :

$$\boldsymbol{T} = \frac{\boldsymbol{M_{C_1C_2}} - \sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\pi_i}}{\sqrt{\sum_{\boldsymbol{i=1}}^{\boldsymbol{N}} \boldsymbol{\pi_i}(\mathbf{1} - \boldsymbol{\pi_i})}}$$

$M_{C_1C_2}$ est le nombre de réponses identiques entre les candidats C_1 et C_2 ;
π_i est la probabilité que les deux candidats aient la même réponse (correcte ou erronée) à la question i ;
N est le nombre de questions du test.

Résultats

Une comparaison de cinq modèles testés sur des données simulées a été réalisée. Des réponses de candidats ont été générées et une triche a été simulée en forçant certaines réponses d'un candidat (copieur) à être identiques à celles du fournisseur. À chaque modèle correspond une probabilité de répondre correctement différente, à compétence et paramètres de questions fixés. La figure ci-dessous nous renseigne sur la puissance des modèles utilisés en fonction de la proportion du nombre de questions copiées. Les analyses présentées ici sont effectuées avec un risque théorique de 1%[3], sur un examen fictif de 50 items. Lorsque la proportion de questions recopiées augmente, le taux de détection augmente logiquement. Pour une proportion de questions recopiées située entre 30 et 40%, les modèles ont une puissance de détection d'environ 50%. Si la moitié des questions est copiée, les candidats fraudeurs sont détectés environ 8 fois sur 10 quel que soit le modèle utilisé. Ce taux de détection approche les 100% lorsque 70% des questions sont recopiées. À partir de 25% de questions recopiées, la méthode « *Birnbaum* » avec indépendance apparait comme la plus puissante.

[3] *Il y a ainsi 1% de chance d'accuser à tort deux candidats d'avoir triché ensemble.*

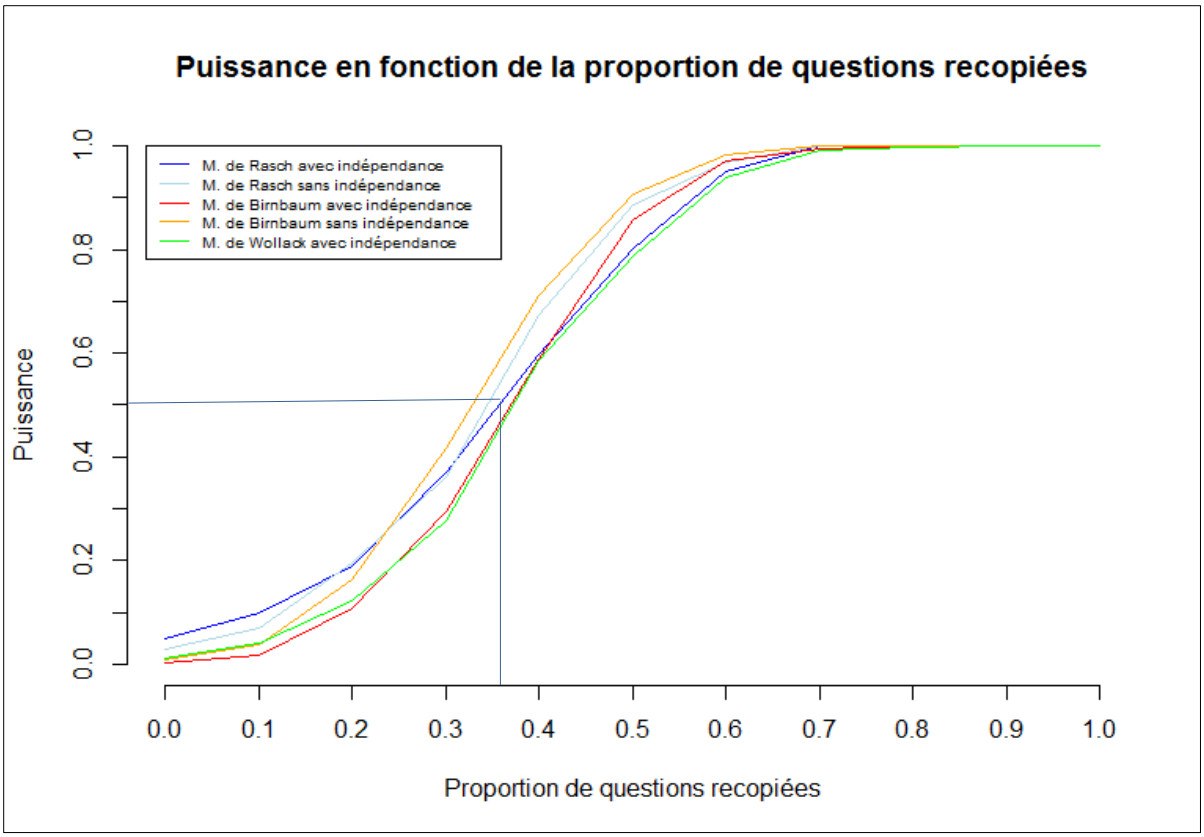

Figure 2 : Puissance des statistiques de test en fonction de la proportion de questions copiées

Conclusion

Les analyses statistiques menées sur les différents indices de détection de comportements suspects nous permettent d'identifier des moyens automatisables de prévenir et détecter la fraude dans les épreuves QCM du Test d'évaluation de français (TEF). Un certain nombre de paramètres, tels que la connaissance des items potentiellement dévoilés, entrent cependant en jeu et leur exploitation contribue à une meilleure efficacité des outils mis en œuvre. Différentes analyses sur des données réelles ont ainsi permis d'identifier les questionnaires présentant une plus grande fréquence de comportements aberrants. Leur distribution a été suspendue et les items les plus fréquemment concernés ont été retirés de la banque d'items. Ces analyses ont également permis de déterminer les zones qui concentraient le plus de cas de fraudes. Des actions peuvent ainsi être entreprises avec les centres agréés concernés pour renforcer la sécurité des tests.

Mots-clés

Intégrité des tests, patrons de réponses aberrants, français langue étrangère

Références bibliographiques

Artner, R. (2016) A simulation study of person-fit in the Rasch model. *Psychological Test and Assessment Modeling, Volume 58* (3), 531-563

Béland, S., Raîche, G., Magis, D., Riopel, M. (2016). Étude de nouveaux indices de détection de la réponse au hasard et de l'inattention selon différentes valeurs de l'habileté dans le contexte de la modélisation de Rasch. *Mesure et évaluation en éducation. 39*(1), 95-118.

Bertrand, R. & Blais, J.-G. (2004). *Modèle de mesure : l'apport de la théorie de la réponse aux items*. Sainte-Foy, Québec : Presses de l'Université du Québec.

Huang, T.-W. (2012). Aberrance Detection Powers of the BW and Person-Fit Indices. *Educational Technology & Society, 15* (1), 28–37.

Magis, D., Béland, S., & Raîche, G. (2012). Snijders's correction of Infit and Outfit indexes with estimated ability level: An analysis with the Rasch model. *Journal of Applied Measurement*, *15*, 82-93.

Magis, D., Raîche, G., & Béland, S. (2011). A didactic presentation of Snijders' *lz** index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics*, *37*, 57-81.

Raîche, G. (2002). Le dépistage du sous-classement aux tests de classement en anglais, langue seconde, au collégial. Gatineau, Québec : Collège de l'Outaouais.

Raîche, G. & Blais, J.-G. (2003). Efficacité du dépistage des étudiantes et des étudiants qui cherchent à obtenir un résultat faible au test de classement en anglais, langue seconde, au collégial. Dans J.-G. Blais et G. Raîche (dir.), *Regards sur la modélisation de la mesure en éducation et en sciences sociales* (pp. 73-90). Saint-Nicolas, Québec : Presses de l'Université Laval.

Romero Mauricio R., Alvaro R. & Diego J. (2015). On the optimality of answer-copying indices: theory and practice. *Journal of Educational and Behavioral Statistics, 40(5)*, pp. 435-453.

Tendeiro, J. N., Meijer, R. R., & Niessen, A. S. M. N. (2016). PerFit: An R package for person-fit analysis in IRT. *Journal of Statistical Software, 74*(5). DOI: 10.18637/jss.v074.i05