Revue des sciences de l'éducation

L'utilisation du facteur de Bayes pour identifier les étudiants qui répondent au hasard
The use of the Bayes factor to identify students who guess at random
La utilización del factor de Bayes para identificar a los estudiantes que responden al azar

Sébastien Béland, Gilles Raîche and David Magis

Volume 41, Number 3, 2015

URI: https://id.erudit.org/iderudit/1035310ar
DOI: https://doi.org/10.7202/1035310ar

See table of contents

Publisher(s)

Revue des sciences de l'éducation

ISSN

0318-479X (print)
1705-0065 (digital)

Explore this journal

Cite this article

Béland, S., Raîche, G. & Magis, D. (2015). L'utilisation du facteur de Bayes pour identifier les étudiants qui répondent au hasard. *Revue des sciences de l'éducation*, *41*(3), 385–407. https://doi.org/10.7202/1035310ar

Article abstract

The available methods that permit detecting students who guess at random in learning assessment tests present many limits. For example, person-fit indexes need generally large data matrices and can be used only to detect if a student responds in accordance to a measurement model (e.g., Rasch models).

In this paper, we will present a new approach to identify students who guess at random in learning assessment tests. After discussing limits of some existing methods, we will expose the technical details of the use of the Bayes factor to evaluate a number of informative hypotheses (Hoijtink, 2012 ; Hoijtink, Klugkist and Boelen, 2008). Next, we will apply this Bayes factor to a simulation study and real data sets for illustration purposes. Our result shows that the Bayes factor is a promising way to detect students who guess at random in learning assessment tests.

L'utilisation du facteur de Bayes pour identifier les étudiants qui répondent au hasard *

Sébastien Béland
Professeur
Université de Montréal

Gilles Raîche
Professeur
Université du Québec à Montréal

David Magis
Chargé de recherches FNRS
Université de Liège

RÉSUMÉ • Les méthodes permettant de détecter les réponses au hasard dans l'évaluation des apprentissages présentent quelques limites. Par exemple, les indices de détection de patrons de réponses inappropriés (*person-fit indexes*) nécessitent généralement d'énormes bases de données et permettent seulement de dire si un étudiant répond en accord ou non avec un modèle de mesure (par exemple, le modèle de Rasch). Dans le cadre de cet article, nous présentons une nouvelle approche permettant d'identifier les étudiants qui répondent au hasard lors d'épreuves d'évaluation des apprentissages. Après avoir discuté des limites des principales approches existantes, nous exposons les détails techniques de l'utilisation du facteur de Bayes pour évaluer un nombre fini d'hypothèses informatives. Ensuite, nous appliquons le facteur de Bayes à des données simulées et des données réelles obtenues à des fins d'illustration. Les résultats permettent de voir que le facteur de Bayes est une méthode prometteuse pour détecter le comportement de réponse au hasard.

* Nous tenons à remercier Herbert Hoijtink pour son soutien lors d'un stage doctoral effectué à l'Université d'Utrecht (Hollande). Nous remercions aussi les trois évaluateurs anonymes de cet article : ils ont grandement contribué à l'amélioration de celui-ci. Cette étude a été soutenue financièrement par le Conseil national de recherche en sciences humaines (CRSH).

MOTS-CLÉS • Réponse au hasard, facteur de Bayes, évaluation des apprentissages, patrons de réponses inappropriés, hypothèses informatives.

1. Introduction et problématique

Certains étudiants peuvent répondre au hasard dans le cadre d'épreuves d'évaluation à visée certificative ou sommative. Par exemple, il est connu que les étudiants évalués au moyen du *Law school admission test* (LSAT) ou au moyen du test de classement en anglais, langue seconde, (TCALS-II) du réseau collégial québécois francophone ont parfois intérêt à répondre au hasard. Dans le premier cas, cette stratégie est utilisée par des étudiants pour éviter de laisser des questions sans réponses et ainsi tenter d'obtenir un meilleur résultat. Dans le second cas, ceux-ci peuvent plutôt tenter de sous-performer intentionnellement pour être ensuite classés dans un cours plus facile (Raîche, 2002). Ce problème de réponses au hasard peut alors nuire à l'interprétation des résultats par une surévaluation du niveau d'habileté des étudiants, dans le premier cas, et par une sous-évaluation de ce niveau d'habileté, dans le second. Mentionnons, de plus, qu'Angoff (1989) a montré que les étudiants les plus habiles sont avantagés par la réponse au hasard, alors que les étudiants moins habiles sont désavantagés par la réponse au hasard. Enfin, ce type de réponse peut biaiser l'estimation des paramètres des modèles de réponses à l'item (Waller, 1983) et le calcul de la corrélation bisériale (Ashler, 1979), qui sont toutes deux des approches permettant d'obtenir des informations importantes sur les qualités métriques d'un test.

1.1 Identifier la nature de la réponse au hasard : un défi

Selon Brassard (2011), la réponse au hasard consiste à répondre à un questionnaire sans choisir, sans réfléchir. Dans ce cas-ci, l'étudiant décide volontairement de ne pas mettre son niveau d'habileté réel à contribution lors d'un test.

Ce type de réponse intéresse les chercheurs depuis très longtemps (Slakter, 1968 ; Votaw, 1936). Par exemple, Cronbach (1946) parlait déjà de la réponse au hasard (*tendency to gamble*) pour référer à l'un des plus importants problèmes susceptibles de survenir lors d'une situation d'évaluation en éducation. Plus récemment, Brassard (2011) demandait à des étudiants de répondre à une épreuve de classement, en anglais langue seconde, en utilisant plusieurs stratégies, dont celle de la réponse au hasard. L'objectif de Brassard était non seulement d'identifier les différents comportements utilisés par les étudiants pour tenter de se sous-classer intentionnellement mais aussi de leur associer des patrons de réponses représentatifs. Elle demandait, dans un premier temps, à des étudiants de niveau collégial (entre 17 et 19 ans) de sous-performer intentionnellement au test de classement. Ensuite, ceux-ci devaient décrire la stratégie qu'ils avaient appliquée pour sous-performer. Une analyse de contenu a été effectuée pour classer les diverses stratégies décrites par les étudiants, et une analyse de régression logistique pour données nominales a ensuite été appliquée pour tenter de prédire

l'utilisation de ces stratégies à partir des patrons de réponses observés. Malheureusement, ces analyses n'ont pas permis d'identifier de patrons de réponse type représentant un comportement de réponses au hasard. Il reste donc beaucoup de travail à faire afin de mieux comprendre la nature de la réponse au hasard dans les épreuves d'évaluation.

1.2 Détecter le hasard : un défi méthodologique

Selon Lanning (1989), il est important de comprendre que les réponses inappropriées telles que la réponse au hasard ne sont pas des événements fréquents. Malheureusement, la détection d'événements rares est très difficile. À ce jour, les chercheurs ont développé relativement peu d'outils permettant de détecter les étudiants qui auraient présenté un patron de réponses au hasard. Ainsi, les auteurs se sont surtout concentrés à développer des stratégies visant à détecter si un étudiant répond en conformité avec un modèle de mesure (Karabatsos, 2003 ; Meijer et Sijtsma, 2001) ou encore si un étudiant semble avoir copié sur ses voisins (Angoff, 1974 ; Sotaridona et Meijer, 2002 ; Wollack, 1997 ; Wollack et Cohen, 1998). Malheureusement, ces stratégies ne permettent pas d'identifier spécifiquement un comportement de réponse au hasard.

Dans le cadre de cet article, nous aurons comme objectif de présenter et d'illustrer une toute nouvelle méthode permettant de détecter les étudiants qui répondent au hasard dans les épreuves d'évaluation des apprentissages en éducation. Pour ce faire, nous nous inspirerons des travaux de Hoijtink (2012) et de Hoijtink, Klugkist et Boelen (2008) en utilisant le facteur de Bayes afin d'évaluer un nombre fini d'hypothèses informatives sur les réponses offertes par les étudiants dans les épreuves d'évaluation en éducation. L'intérêt d'une telle méthode est qu'elle permet d'analyser des données recueillies avec des échantillons relativement limités et de tester des hypothèses de différente nature.

Le contenu du présent texte se divise en six grandes sections. Dans la deuxième section, nous présentons le contexte théorique. Nous exposons la méthodologie dans la troisième section. Dans la quatrième section, nous présentons les résultats avant de continuer par une discussion. Le texte se terminera par une conclusion générale, à la section six.

2. Contexte théorique

2.1 Principales stratégies psychométriques permettant de détecter le comportement de réponse au hasard

Nous présentons deux approches permettant de soutenir la détection de la réponse au hasard : les indices de détection de patrons de réponses inappropriés (*person-fit indexes*) et le modèle multidimensionnel à quatre paramètres de personnes de Raîche, Magis, Blais et Brochu (2012). Notez qu'il est possible de trouver quelques méthodes peu usuelles telles que l'indice de sabotage de Cattell (*Cattell's sabotage index*), mais O'Dell (1971) a déjà montré que cette méthode n'offrait pas toujours

des résultats convaincants. De plus, il est extrêmement difficile de retracer des écrits portant sur ces approches, car elles ont été utilisées de façon relativement marginale dans le passé.

Selon Waller (1973), la réponse au hasard peut biaiser l'estimation des paramètres des modèles de réponse à l'item. Pour contourner ce problème, cet auteur a développé le modèle *Ability Removing Random Guessing*, qui utilise un modèle de réponse à l'item permettant d'obtenir des estimations moins biaisées des paramètres en présence de réponses au hasard. Bien que cette approche offre un certain intérêt, nous ne l'avons pas retenue puisqu'elle présente un problème de taille : elle ne permet pas de détecter formellement la réponse au hasard.

2.1.1 Les indices de détection de patrons de réponses inappropriés (*person-fit indexes*)

Les indices de détection de patrons de réponses inappropriés permettent de détecter les réponses qui ne respectent pas un modèle de mesure précis. Selon Meijer et Sijtsma (2001), il existe deux grandes catégories d'indices de détection de patrons de réponses inappropriés : les indices paramétriques et les indices non paramétriques. D'une part, les indices paramétriques reposent sur l'utilisation d'un modèle de réponse à l'item (Bertrand et Blais, 2004 ; Hambleton et Swaminathan, 1985) qui permet de calculer la probabilité π_j qu'un étudiant obtienne une bonne réponse à l'item j ($j = 1,\ldots, J$). Ainsi, le modèle de Rasch peut s'écrire comme suit :

$$\pi_j = \frac{e^{\theta - b_j}}{1 + e^{\theta - b_j}} \tag{1}$$

où θ est un paramètre de personne correspondant au niveau d'habileté de celle-ci et b_j est un paramètre de difficulté de l'item. Il est à noter que d'autres modèles tels que le modèle à deux paramètres et le modèle à trois paramètres ont aussi été proposés (Hambleton et Swaminathan, 1985).

Le modèle présenté à l'équation (1), comme ses déclinaisons à deux ou à trois paramètres, est le point de départ permettant de calculer l'indice l_z (Drasgow, Levine et Williams, 1985), qui est fort probablement le plus connu de tous. D'abord, il faut calculer la somme du logarithme (…) de la vraisemblance du niveau d'habileté à chacun des items d'une épreuve d'évaluation contenant des items à réponses dichotomiques :

$$l_0 = \sum_{j=1}^{J} \{x_j log \pi_j + (1 - x_j) log(1 - \pi_j)\} \tag{2}$$

où x_j est la réponse d'un étudiant à l'item j, codée 1 (correspondant à une bonne réponse) ou 0 (correspondant à une mauvaise réponse). Cette probabilité est généralement calculée selon une des modélisations pour réponses dichotomiques issues de la théorie de la réponse à l'item que nous avons présentée précédemment

à l'aide de l'équation (1). Ensuite, puisque l'indice l_0 n'est pas standardisé, celui-ci doit être transformé en scores z pour faciliter son interprétation et le rendre comparable, quelle que soit la valeur du niveau d'habileté de l'étudiant :

$$l_z = \frac{l_0 - E(l_0)}{V(l_0)^{1/2}} \tag{3}$$

où

$$E(l_0) = \sum_{j=1}^{J}\{\pi_j log\pi_j + (1-\pi_j)\log(1-\pi_j)\} \tag{4}$$

et

$$V(l_0) = \sum_{j=1}^{J}\{\pi_j(1-\pi_j)\left(\log\frac{\pi_j}{1-\pi_j}\right)^2\} \tag{5}$$

sont respectivement la moyenne (espérance mathématique) et la variance de l_0. Puisque l'indice l_z devrait se distribuer asymptotiquement selon une loi normale, on dira qu'un patron de réponses est inapproprié lorsqu'il respecte la condition suivante

$$l_z \leq z_\alpha \tag{6}$$

où z_α correspond au quantile d'une loi normale centrée réduite. Par exemple, au seuil de détection α de 0,01, un étudiant présentant un score l_z plus petit ou égal à -2,33 sera considéré comme présentant un patron de réponses inapproprié. Les valeurs positives de cet indice indiquent, au contraire, que le patron de réponses est approprié. Enfin, il est pertinent de noter que Lee, Stark et Chernyshenko (2014) ont démontré que l'indice l_z est efficace pour détecter la réponse au hasard.

D'autre part, les indices non paramétriques reposent plutôt sur la logique du vecteur parfait de Guttman (1950). Imaginons une épreuve d'évaluation contenant six items l'étudiant peut obtenir une bonne réponse (symbolisée par 1) ou une mauvaise réponse (symbolisée par 0). Une fois tous les items classés en ordre croissant de difficulté, les réponses 111000 représentent un ensemble de réponses parfait : l'étudiant donne des bonnes réponses aux items faciles et des mauvaises aux items difficiles. À l'opposé, le patron 010011 semblerait inapproprié, car l'étudiant aurait obtenu de bonnes réponses aux deux items les plus difficiles et une seule bonne réponse aux quatre items les plus faciles.

Il existe deux familles d'indices qui sont utilisés pour comparer le patron parfait au patron observé. La première est basée sur le nombre de fois où le patron observé ne correspond pas au patron parfait. C'est cette famille qu'on rencontre le plus souvent. La seconde est basée sur le calcul d'un coefficient de corrélation entre le score obtenu à chacun des items et un indice associé à chacun des items du patron parfait : il s'agit généralement du rang ou du niveau de difficulté calculé selon la proportion de bonnes réponses dans le groupe de référence.

Ces indices, qu'ils soient paramétriques ou non paramétriques, présentent de nombreuses limites. Premièrement, puisqu'ils ne vérifient que l'ajustement global du modèle, ils ne permettent pas de détecter directement le comportement spécifique de réponse au hasard. Au mieux, on doit s'inspirer de patrons de réponses-modèles et voir si un indice est capable de bien les détecter. Deuxièmement, les indices de détection ont une interprétation dichotomique : à partir d'un modèle de mesure tel que celui présenté à l'équation (1), on ne peut que dire si les réponses d'un étudiant sont appropriées ou inappropriées. Or, la réalité comporte beaucoup plus de nuances, surtout si un chercheur souhaite détecter un comportement aussi précis que la réponse au hasard.

2.1.2 L'indice de pseudo-chance personnelle *C*

Une deuxième approche est inspirée d'une modélisation multidimensionnelle de la théorie de la réponse aux items (Reckase, 1985, 1997, 2009). Raîche et ses collaborateurs (2012) ont élaboré un modèle probabiliste comprenant un indice de personne de pseudo-chance. Dans le cadre de cette modélisation, la probabilité d'obtenir une bonne réponse à un item correspond maintenant à :

$$\pi_j = (C + c_j) + \frac{1-(C+c_j)}{1+e^{-a_j(\theta-b_j)}} \tag{7}$$

où θ est un paramètre de personne correspondant au niveau d'habileté de celle-ci, a_j est un paramètre de discrimination de l'item, b_j est un paramètre de difficulté de l'item et c_j est un paramètre de pseudo-chance de l'item. Dans ce cas-ci, C est un paramètre de pseudo-chance propre à chacune des personnes plutôt qu'à l'item. Raîche et al., (2012) ont déjà démontré que le modèle présenté à l'équation (7) semble suffisamment efficace pour corriger de façon appréciable le niveau d'habileté de l'individu. De plus, cette modélisation, comme d'autres variations de celle-ci, n'engendre pas trop de biais ni d'augmentation importante de l'erreur type dans l'estimation du niveau d'habileté lorsque le patron de réponses est approprié. Nous suggérons au lecteur intéressé de consulter Raîche et al., (2012) pour obtenir plus de détails techniques sur cette approche.

L'indice *C* permettrait de détecter le comportement de réponse au hasard. Par contre, son utilisation présente toujours des limites importantes Raîche, Béland, Magis, Blais et Brochu, (2010). Premièrement, l'estimation de *C* impose l'utilisation d'une source de données importante, puisqu'il repose sur la calibration préalable des items d'un test ou d'une banque d'items. Il est alors généralement difficile d'utiliser cet indice pour analyser des résultats au sein d'un seul groupe-classe. Deuxièmement, à ce jour, peu d'études ont été effectuées sur ce modèle et on en connaît encore trop peu les caractéristiques et, donc, les limites. Il faudrait, entre autres, investiguer plus en détail au sujet de la distribution de cet indice pour en faciliter l'interprétation.

2.2 Évaluation d'hypothèses informatives à l'aide du facteur de Bayes

Nous avons vu que les modèles de mesure permettant de détecter la réponse au hasard présentent certaines limites : ils nécessitent en général le recours à de vastes banques de données et ne permettent pas de détecter directement le comportement de réponse au hasard. Ainsi, il serait utile de développer de nouvelles méthodes permettant de détecter plus adéquatement le comportement de réponse au hasard.

Dans le cadre de cet article, nous nous inspirerons de l'utilisation du facteur de Bayes dans le contexte d'hypothèses dites informatives (Hoijtink, 2012 ; Hoijtink, Klugkist et Boelen, 2008) pour détecter les étudiants qui répondent au hasard dans les épreuves d'évaluation en éducation. Pour utiliser cette méthode, nous devons tout d'abord :

1) définir les hypothèses à évaluer ;
2) calculer le facteur de Bayes et
3) interpréter le facteur de Bayes.

Les sections suivantes expliquent ces étapes en détail.

2.2.1 Définir les hypothèses à évaluer

L'utilisation du facteur de Bayes nécessite l'emploi de modèles à comparer (Kass et Raftery, 1995). Dans le cadre de cette étude, nous parlerons plutôt d'hypothèses (informatives) à évaluer, afin de rester cohérents avec la terminologie employée par Hoijtink (2012) et Hoijtink, Klugkist et Boelen (2008). Avec un test d'hypothèses classique, nous poserions l'hypothèse nulle (notée H_0) selon laquelle un ensemble de statistiques sont égales, par exemple des probabilités notées de $j = 1$ à J :

$$H_0 : \pi_1 = \pi_2 = \ldots = \pi_J.$$

Dans ce cas-ci, l'hypothèse H_0 est considérée comme une hypothèse informative, car elle donne une information très précise sur le lien existant entre les probabilités π_j : elles sont toutes égales. À l'opposé, l'hypothèse alternative classique (notée H_1) est plutôt une hypothèse informative non contrainte, car les paramètres π_j sont tous libres :

$$H_1 : \text{rejet de } H_0.$$

Évidemment, comparer H_0 à H_1 paraît un peu simpliste pour le chercheur qui a des hypothèses précises à évaluer : par exemple, dans une situation où nous souhaitons tester la présence d'un comportement de réponse au hasard. Dans le cas où un étudiant répondrait à une série de questions ou d'items contenant chacun quatre choix de réponses, nous pourrions poser l'hypothèse suivante :

$$H_{\text{hasard}} : 0{,}25 - \zeta < \pi_j < 0{,}25 + \zeta \text{ pour les items } j = 1,\ldots, J \qquad (8)$$

où l'étudiant obtient une probabilité π égale à 0,25 d'obtenir une bonne réponse pour chacun des *j* items (soit une chance sur quatre). Le symbole ζ permet de borner la probabilité π selon les valeurs désirées. Par exemple, dans une situation où ζ = 0,05, l'hypothèse $H_{hasard} = 0{,}20 < \pi_j < 0{,}30$. L'avantage de cette notation repose donc sur sa grande flexibilité.

Un autre système d'hypothèses d'intérêt consisterait à évaluer si les réponses d'un étudiant respectent le principe de monotonicité. Cela implique que les réponses d'un étudiant doivent être ordonnées en ordre croissant de difficulté pour tous les items ou questions $j = 1, \ldots, J$ d'une épreuve d'évaluation. Par exemple, $\pi_1 > \ldots > \pi_j$ où π_j est la probabilité qu'un étudiant obtienne une bonne réponse à l'item *j*. Notez que les π_j peuvent être estimés à partir de n'importe quel modèle d'intérêt tel que le modèle de Rasch ou à partir de la proportion de bonnes réponses pour chacun des items ou questions. Aussi, la monotonicité implique que l'ordre de difficulté soit respecté autant pour les items que pour les étudiants :

$$H_{ordre}: \pi_1 > \ldots > \pi_j. \tag{9}$$

2.2.2 Calculer le facteur de Bayes

Nous l'avons déjà dit : le facteur de Bayes permet de comparer plusieurs hypothèses entre elles. Il existe deux grandes composantes pour calculer le dit facteur : l'adéquation et la complexité, qui sont respectivement générées à partir d'une distribution a priori et d'une distribution a posteriori. Dans le cadre de cet article, nous nous en servirons pour évaluer quelle hypothèse est la plus représentative des réponses fournies par chacun des étudiants à une épreuve d'évaluation.

Imaginons que nous souhaitons comparer l'hypothèse H_m à l'hypothèse $H_{m'}$ en utilisant les données *X*. La fonction de probabilité associée à l'hypothèse *m* peut être définie comme étant égale à :

$$p_m(X) = \int f_m(X|\pi) h_m(\pi) \partial\pi \tag{10}$$

où π est un paramètre de probabilité, $h_m(\pi)$ est la distribution a priori de ce paramètre et $f_m(X|\pi)$ la fonction de densité des données selon l'hypothèse *m*. De là, on peut calculer le facteur de Bayes entre ces deux hypothèses en utilisant la forme suivante :

$$FB_{mm'} = \frac{p_m(X)}{p_{m'}(X)} = \frac{\int f_m(X|\pi) h_m(\pi) \partial\pi}{\int f_{m'}(X|\pi) h_{m'}(\pi) \partial\pi} \tag{11}$$

qui est la portion de la distribution a posteriori en accord avec l'hypothèse m divisée par la portion de la distribution a posteriori en accord avec l'hypothèse *m'*. Ainsi, dans une situation où l'évaluateur souhaite évaluer une hypothèse informative (la contrainte entre les π_j est précise) et une hypothèse non informative, Mulder, Hoijtink et Klugkist (2010) ont simplifié le calcul du facteur bayésien de la façon suivante :

$$FB_{mm'} = {f_m}/{c_m} \tag{12}$$

où c_m est la complexité (*complexity*) et f_m est l'adéquation (*fit*) des hypothèses m et m'. Dans une situation où deux hypothèses informatives sont évaluées, Hoijtink (2012) a démontré que $FB_{mm'}$ devient.

$$FB_{mm'} = {\frac{f_m}{c_m}}\Big/{\frac{f_{m'}}{c_{m'}}}. \tag{13}$$

Les lignes qui suivent donnent plus de détails sur la façon de calculer les éléments contenus dans les équations (12) et (13).

Pour la présente étude, nous utilisons la distribution bêta lors de l'évaluation d'hypothèses non contraintes

$$h(\pi) = \prod_{j=1} B\hat{e}ta(\pi_j|1,\ 1) = 1 \tag{14}$$

où les paramètres {1, 1} la rendent équivalente à la loi uniforme. Par contre, lorsque les hypothèses sont contraintes selon un ordre du type $\pi_j > \pi_{jr}$ ou $\pi_j < \pi_{jr}$, la complexité n'est pas dépendante de cette loi de distribution. De plus, la proportion de la distribution a priori en accord avec l'hypothèse contrainte H_m peut être réécrite comme suit :

$$c_m = \int_{\pi \in H_m} h(\pi)\partial\pi \tag{15}$$

où l'élément $\pi \in H_m$ détermine que la probabilité est soutenue par l'hypothèse H_m. Cette fonction représente la complexité. En ce qui a trait à la génération des données à partir de la distribution a posteriori, la fonction de densité des données est définie par :

$$f(x|\pi) = \prod_{j=1} \pi_j^{x_j}(1-\pi_j)^{1-x_j} \tag{16}$$

et elle stipule que les x_j sont indépendants. Ainsi, l'adéquation f_m est définie comme la proportion de la distribution a posteriori en accord avec l'hypothèse H_m où :

$$f_{m=} \int_{\pi \in H_m} \prod_{j=1}^{J}(-2(x_j - 1) + 2\pi_j(2x_j - 1))\partial\pi. \tag{17}$$

2.2.3 Échantillonnage des distributions de probabilité a priori et a posteriori

Selon Jeffreys (1961), les modèles bayésiens peuvent être simplifiés sous la forme suivante :

$$\text{Distribution a posteriori} \propto \text{Distribution des données} \times \text{Distribution a priori} \tag{18}$$

Plus explicitement, l'équation (18) souligne que la distribution a posteriori est proportionnelle ($\propto$) au produit de la distribution des données et de la distribution a priori. Dans ce cas-ci, la distribution a posteriori représente l'information obtenue après avoir pondéré les données à l'aide de la distribution a priori. Comme le lecteur le comprendra, c'est surtout la fonction de densité des données telles qu'elles ont été observées qui est pertinente. Ensuite, la distribution a priori est l'information que le chercheur possède avant l'observation des données. Cette information est généralement tirée d'études antérieures ou d'hypothèses théoriques pertinentes. Elle peut prendre la forme d'une distribution de probabilité précise : par exemple, la loi normale pour exprimer l'étalonnage du niveau d'habileté des étudiants (nous pensons à la moyenne et à l'écart type obtenus à une enquête internationale en science, tel que le *Trends in international mathematics and science study* – TIMSS).

Quelques propriétés doivent être mises en relief pour bien comprendre la portée technique de cette méthode. Premièrement, c'est la distribution de probabilité bêta qui sera utilisée comme distribution a priori. Le choix de cette loi est lié au fait qu'elle permet de générer des probabilités pour des données qui sont bornées dans l'intervalle [0, 1] avec une probabilité uniforme. Deuxièmement, le fait de fixer les paramètres $bêta(\pi_j|1,1$ rend cette distribution de probabilité a priori assez vague, et donc non informative, de telle façon qu'elle influence peu la distribution a posteriori. Du même coup, cela permet à la distribution a posteriori d'être complètement déterminée par les données observées, ce qui confère un caractère dit *objectif* à cette dernière distribution de probabilité (Hoijtink, Klugkist et Boelen, 2008).

Dans le cas présent, nous utilisons la méthode de Gibbs pour échantillonner les données des distributions de probabilité a priori et a posteriori pour calculer la complexité ainsi que l'adéquation du facteur de Bayes. Rappelons rapidement que l'échantillonnage de Gibbs est une méthode numérique permettant de générer des données selon des distributions de probabilités conditionnelles complexes à partir de fonctions de distribution de probabilités connues et plus simples. Dans le contexte des hypothèses que nous désirons vérifier ici, son fonctionnement peut être synthétisé en trois grandes étapes :

1) spécification des valeurs initiales des paramètres π_j ;
2) pour chacun des échantillons, génération aléatoire de nouveaux paramètres π_j^k pour les valeurs de π_j générées à la k^e itération ;
3) répétition des étapes i) et ii) le nombre de fois désiré en utilisant toujours comme point de départ les paramètres obtenus à l'étape précédente.

C'est la loi *Bêta* qui est utilisée pour générer les π_j. Le lecteur intéressé à obtenir plus d'information sur cette approche est invité à consulter Casella et George (1992) ou Jackman (2009).

2.2.4 La complexité c_m et l'adéquation f_m

Dans un premier temps, la complexité est définie comme le rapport de la distribution de probabilité a priori en accord avec l'hypothèse H_m et la distribution de probabilité a priori en accord avec une hypothèse de rechange. Par exemple, pour l'hypothèse H_{ordre}: $\pi_1 > \ldots > \pi_J$, nous calculerions la complexité en utilisant $H_{ordre} = 1/J!$ Sachant que nous comparons trois items:

$$H_{ordre}: \pi_1 > \pi_2 > \pi_3$$

la valeur de c_m serait de un sur six, puisque l'on peut dériver six autres configurations de H_{ordre}; par exemple, H_{ordre}: $\pi_3 > \pi_2 > \pi_1$ ou H_{ordre}: $\pi_2 > \pi_1 > \pi_3$. Un autre exemple pertinent est l'hypothèse H_{hasard}. Dans ce cas particulier, on calculerait H_{hasard}: $0{,}25 - \zeta < \pi_j < 0{,}25 + \zeta$ pour les trois items de la façon suivante: $c_{hasard} = (2\zeta)^{-3}$.

Dans un deuxième temps, l'adéquation (*fit*) est le rapport qui consiste à comparer la distribution de probabilité a posteriori des données D en accord avec l'hypothèse $H_{m:}$ $P(D, H_m)$ et la distribution de probabilité a posteriori des données en accord avec une hypothèse de rechange $P(D, H_a)$. Dans ce cas-ci, nous vérifions si les données confirment bien une hypothèse précise. Par exemple, imaginons les patrons de réponses, celles-ci mises en ordre croissant de difficulté, $\boldsymbol{x}_1 = [111010]$ et $\boldsymbol{x}_2 = [010101]$. Si nous souhaitons tester l'hypothèse H_{ordre}: $\pi_1 > \ldots > \pi_J$, nous voyons que x_1 a plus de chances d'être en accord avec H_{ordre} que x_2. Dans le cas de l'hypothèse H_{hasard}: $0{,}25 - 0{,}05 < \pi_j < 0{,}25 + 0{,}05$, c'est plutôt le patron de réponse x_2 qui respecte le mieux l'hypothèse évaluée. Ainsi, f_m sera élevé si les données permettent bien de vérifier l'hypothèse à évaluer.

2.2.5 Interprétation du facteur de Bayes et nomenclature

Le tableau 1 permet aussi d'obtenir plus d'informations sur l'interprétation du facteur de Bayes. Sachant que les réponses ont été ordonnées selon l'ordre croissant du niveau de difficulté des items, nous pouvons remarquer que les réponses du haut du tableau 1 sont clairement en accord avec l'hypothèse du respect de l'ordre de la double monotonicité. À l'opposé, nous pouvons remarquer que les deux patrons de réponses au bas du tableau penchent plus en faveur de l'hypothèse d'un comportement de réponses au hasard.

Il est à noter qu'il existe plusieurs règles pour interpréter le facteur de Bayes. Par exemple, Kass et Raftery (1995) ont produit une nomenclature visant à faciliter l'interprétation de celui-ci. Les correspondances interprétatives vont comme suit: pour un facteur de Bayes de 1-3, *peu important*; de 3-20, *présence de preuves positives*; de 20-50, *présence de preuves fortes* et de plus de 50, *présence de preuves convaincantes*.

Tableau 1
Données théoriques et $FB_{ordre,hasard}$

J	Données	$FB_{ordre/hasard}$
	Monotonicité	
6	110100	8,86
12	111011010000	102,74
	Réponses au hasard	
6	000100	0,10
12	100010000010	0,09

2.3 Un exemple

Imaginons une situation où nous souhaiterions analyser les données suivantes :

$$x_j = [11101000]$$

et tester l'hypothèse H_{ordre}: $\pi_1 > \ldots > \pi_j$ contre l'hypothèse H_{hasard}: $0,25 - 0,05 < \pi_j < 0,25 + 0,05$. En nous inspirant des étapes présentées à la section précédente, le calcul du facteur de Bayes peut être décomposé comme suit.

2.3.1 Étape 1 : Calculer les probabilités

Le chercheur doit fournir le vecteur des probabilités π_j pour chacun des *j* items. Ces probabilités sont généralement obtenues à l'aide d'un modèle probabiliste tel que celui présenté à l'équation (1) ou sont fournies à partir de toute autre information à la disposition de l'analyste. Dans cet exemple, nous déterminons le vecteur des probabilités suivant :

$$P(x_j) = [0,8, 0,9, 0,7, 0,5, 0,9, 0,4, 0,6, 0,3].$$

Selon ce patron de réponses, un étudiant a une probabilité de 0,8 d'obtenir une bonne réponse à l'item 1 et de 0,3 au dernier item.

2.3.2 Étape 2 : Calcul de la complexité c_m

Pour calculer la complexité c_m, il faut et il suffit de générer le nombre de probabilités désiré à partir de la distribution de probabilité choisie, ici une distribution *Bêta*[1, 1]. Par exemple, pour 1 000 000 itérations, nous avons obtenu les résultats suivants présentés au tableau 2.

Les deux dernières colonnes de ce tableau indiquent si les probabilités générées à chacune des itérations sont en accord avec les hypothèses testées. Ainsi, l'itération cinq est en accord avec l'hypothèse H_{ordre}, alors que l'itération trois est en accord avec l'hypothèse H_{hasard}.

Tableau 2
Itérations générées pour le calcul de c_m

Itérations	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	H_o	H_h
1	0,83	0,88	0,91	0,67	0,82	0,46	0,33	0,35	0	0
2	0,70	0,81	0,84	0,67	0,78	0,33	0,67	0,45	0	0
3	0,22	0,26	0,28	0,21	0,29	0,28	0,21	0,22	0	1
4	0,79	0,77	0,86	0,81	0,99	0,61	0,46	0,50	0	0
5	0,99	0,84	0,82	0,79	0,73	0,67	0,61	0,55	1	0
…										
1000000	0,89	0,81	0,74	0,38	0,88	0,48	0,50	0,31	0	0
Total									20	1
Proportion									0,00002	0,000001

2.3.3 Étape 3 : calcul de l'adéquation f_m

Il est nécessaire d'apporter quelques nuances pour calculer l'adéquation f_m. Dans ce cas-ci, nous devons utiliser l'échantillonnage de Gibbs pour pondérer le vecteur des probabilités $P(x_j)$ à l'aide de la distribution *bêta*[1, 1]. Pour 1 000 000 d'itérations, nous présentons les résultats au tableau 3 (sensiblement différents de ceux présentés au tableau 2).

Tableau 3
Itérations générées pour le calcul de f_m

Itérations	π_1	π_2	π_3	π_4	π_5	π_6	π_7	π_8	H_o	H_h
1	0,87	0,82	0,79	0,67	0,62	0,46	0,35	0,27	1	0
2	0,75	0,68	0,81	0,77	0,74	0,43	0,57	0,45	0	0
3	0,22	0,25	0,28	0,28	0,25	0,21	0,26	0,23	0	1
4	0,31	0,84	0,76	0,44	0,49	0,55	0,42	0,60	0	0
5	0,93	0,87	0,81	0,79	0,77	0,61	0,59	0,55	1	0
…										
1000000	0,89	0,71	0,74	0,48	0,88	0,58	0,58	0,21	0	0
Total									600	6
Proportion									0,0006	0,000006

Dans cet exemple, nous remarquons que les probabilités générées aux itérations un et cinq sont en accord avec l'hypothèse $H_{ordre,}$ alors que les probabilités générées à l'itération trois sont en accord avec l'hypothèse H_{hasard}.

2.3.4 Étape 4 : Calcul du facteur de Bayes FB

La dernière étape consiste à utiliser les résultats obtenus aux tableaux 2 et 3 pour calculer le facteur de Bayes. Puisque nous avons calculé, c_{ordre} = 20 / 1000000 =

0,00002, f_{ordre} = 600 / 1000000 = 0,0006, c_{hasard} = 1 / 1000000 = 0,000001 et f_{hasard} = 6 / 1000000 = 0,000006, cela nous permet d'obtenir un facteur de Bayes $FB_{ordre,hasard}$ = [0,0006 / 0,00002] / [0,000006 / 0,000001] = 5 (voir l'équation 13), ce qui indique ainsi que l'hypothèse H_{ordre} est 5 fois plus fréquentes selon les données que l'hypothèse H_{hasard}.

3. Méthodologie

La méthode de détection d'un comportement de réponse au hasard selon le facteur de Bayes présentée ci-haut sera appliquée à l'aide d'une simulation informatique et à partir des réponses d'étudiants à une épreuve de classement en anglais langue seconde, au collégial. Il est à noter que nous souhaitons uniquement illustrer le potentiel de cette méthode.

3.1 Étude de simulation

3.1.1 Génération des données

Dans un premier temps, nous allons utiliser le modèle de Rasch, qui a été présenté à l'équation (1), pour générer des données pour 3 longueurs de test (6, 12 et 18 items). Dans le cadre de cette étude, le paramètre d'habileté θ sera généré à l'aide de la loi normale N(0, 1) et les difficultés b_i seront fixées et distribuées uniformément entre les valeurs -2,2 à 2,2. Ensuite, 1000 patrons de réponses seront générés de la façon suivante : 1) nous créerons l'échantillon des habiletés θ, 2) nous générerons J nombres aléatoires (pour 6, 12 et 18 items) à partir de la loi uniforme U[0, 1] et 3) pour chacun des j = 1, ... J items, la réponse x_j à un item sera égale à un si la valeur tirée à partir de la loi uniforme U[0, 1] est plus petite que la probabilité π_j calculée à partir du modèle de Rasch. Si ce n'est pas le cas, x_j sera égal à zéro.

Dans un deuxième temps, nous générerons des données au hasard pour trois longueurs de test : 6, 12 et 18 items. Dans ce cas-ci, la procédure suivra les étapes ci-dessous. 1) Nous générerons J nombres aléatoires pour 6, 12 et 18 items à partir de la loi uniforme U[0, 1]. 2) Pour chacun des j = 1, ... J items, la réponse x_j à un item sera égale à un si la valeur tirée de la loi uniforme U[0, 1] est plus petite que la probabilité π_j = 0,25. Si cette valeur est plus petite que π_j = 0,25, x_j est égal à zéro. Cette procédure sera répétée 1000 fois.

3.1.2 Méthode d'analyse des données

Nous allons présenter les résultats pour l'hypothèse

$$H_{ordre}: \pi_1 > \ldots > \pi_j \text{ contre } H_{non\text{-}ordre}: \text{pas } H_{ordre}$$

où $H_{non\text{-}ordre}$ est une hypothèse non contrainte. De plus, nous allons analyser les données simulées à l'aide de l'hypothèse

$$H_{hasard}: 0{,}25 - 0{,}05 < \pi_j < 0{,}25 + 0{,}05 \text{ contre } H_{non\text{-}hasard}: \text{pas } H_{hasard}$$

où $H_{non\text{-}hasard}$ est une hypothèse non contrainte. Dans tous les cas, nous indiquerons le pourcentage d'erreur de FB < 1.

3.2 Analyse d'un test en anglais langue seconde

La loi des grands nombres stipule que pour une épreuve d'évaluation contenant quatre choix de réponse, un étudiant répondant au hasard aurait une chance sur quatre de répondre correctement. Pour cette deuxième étude, nous utiliserons une épreuve de classification contenant des items qui présentent chacun quatre choix de réponses pour vérifier si les étudiants qui obtiennent entre 20 % et 30 % de bonnes réponses ont effectivement répondu au hasard.

3.2.1 Sujets

En 1998, 1373 étudiants du Cégep de l'Outaouais (749 femmes et 624 hommes) ont été soumis à cette épreuve obligatoire pour tous lors de la période d'inscription aux cours d'anglais. Dans le cadre de cette étude, nous analyserons seulement les résultats des 19 étudiants des deux sexes qui ont obtenu entre 20 % et 30 % de bonnes réponses à cette épreuve.

3.2.2 Instrumentation

Une grande partie des étudiants francophones nouvellement admis dans certains cégeps francophones de la province de Québec (Canada) passent le test de classement en anglais langue seconde, de niveau collégial (TCALS-II). Cette épreuve comprend de 85 items, à quatre choix de réponse, répartis en deux domaines (oral et écrit). On l'utilise pour classer les étudiants dans un groupe-classe adapté à leur niveau de maîtrise de l'anglais langue seconde. Il est à noter que les qualités métriques de ce test ont déjà été vérifiées par Raîche (2002) ainsi que par Laurier, Froio, Pearo et Fournier (1998). Selon ces auteurs, il est possible de postuler l'unidimensionnalité du construit, et la fidélité du test est égale à 0,96, donc une valeur assez importante. Notons, de plus, que cette épreuve d'évaluation présente principalement des items relativement faciles et aucun item difficile : la proportion moyenne de bonnes réponses est égale à 0,78 (s = 0,15).

3.2.3 Déroulement

Les étudiants doivent répondre aux questions du TCALS-II en moins de 90 minutes. Ce test comporte deux grandes sections. Dans la première moitié, ils écoutent une bande audio afin que soit analysée leur compréhension auditive. Dans la seconde moitié, c'est la compréhension écrite et la compréhension de la lecture qui sont évaluées. Il est à noter que le test se fait de façon individuelle et en silence. Enfin, tout le matériel est récupéré à la fin de l'épreuve d'évaluation.

3.2.4 Méthode d'analyse des données

Pour cette deuxième étude, nous testerons l'hypothèse selon laquelle l'étudiant a répondu de façon ordonnée contre l'hypothèse selon laquelle il a répondu au hasard :

$$H_{ordre}: \pi_1 > \ldots > \pi_j \text{ contre } H_{hasard}: 0{,}25 - 0{,}05 < \pi_j < 0{,}25 + 0{,}05.$$

Rappelons que chacun des items du test de classement en anglais, langue seconde, comportait quatre choix de réponse. Ainsi, un étudiant qui répondrait au hasard aurait 25 % de chance d'avoir une bonne réponse à un item. Pour être moins limités, nous avons sélectionné uniquement les répondants qui ont obtenu entre 20 % et 30 % de bonnes réponses, soit une approximation du nombre théorique de bonnes réponses en présence d'items contenant quatre choix de réponses. Ensuite, nous avons utilisé le code en langage FORTRAN produit par Hoijtink afin de calculer les facteurs de Bayes des réponses de chacun de ces étudiants. Finalement, nous interpréterons les facteurs de Bayes afin de mettre en relief l'hypothèse qui permettra le mieux de vérifier les données.

3.2.5 Éthique

Il s'agit de données secondaires qui avaient déjà été utilisées dans une recherche antérieure et obtenues dans le cadre d'une opération administrative au Collège de l'Outaouais. Il n'a donc pas été nécessaire de prendre en considération des aspects éthiques. Toutefois, les résultats de recherche ont été communiqués dans un rapport (Raîche, 2002) adressé au personnel du Collège impliqué dans l'administration du test.

4. Résultats

4.1 Étude de simulation

Les résultats sont présentés au tableau 4.

Observons la première colonne. Si les données sont générées à l'aide du modèle de Rasch, le pourcentage d'erreurs (choisir $H_{non\text{-}ordre}$ plutôt que H_{ordre}) est faible et

Tableau 4
Probabilités d'erreurs (en %)

	$FB_{ordre,non\text{-}ordre}<1$	$FB_{hasard,non\text{-}hasard}<1$
Rasch $J = 6$	16 %	62 %
Rasch $J = 12$	16 %	71 %
Rasch $J = 18$	14 %	75 %
Hasard $J = 6$	69 %	17 %
Hasard $J = 12$	82 %	17 %
Hasard $J = 18$	87 %	15 %

borné entre 14 % et 16 %. De plus, le nombre d'erreurs est relativement indépendant du nombre d'items. En ce qui concerne les données au hasard, nous observons que le pourcentage d'erreurs est beaucoup plus élevé (il est borné entre 69 % et 87 %). De plus, ce pourcentage augmente au fur et à mesure que le nombre d'items augmente.

Observons maintenant la deuxième colonne. Lorsque les données sont générées à partir du modèle de Rasch, (choisir $H_{non\text{-}hasard}$ plutôt que H_{hasard}) le pourcentage d'erreurs est plus élevé et borné entre 62 % et 75 %. De plus, nous observons que le nombre d'erreurs augmente lorsque *J* augmente. En ce qui concerne la génération de données au hasard, nous observons que le pourcentage d'erreurs est beaucoup moins élevé et indépendant du nombre d'items.

4.2 Étude des données en anglais, langue seconde

Le Tableau 5 présente les résultats des 19 étudiants qui ont obtenu entre 20 % et 30 % de bonnes réponses au TCALS-II

Selon les résultats illustrés au Tableau 5, c'est l'hypothèse d'un comportement de réponses au hasard H_{hasard} qui est la mieux validée par les patrons de réponses de ces étudiants. Ainsi, pour tous les patrons de réponses analysés, les $FB_{ordre/hasard}$ obtenus tendent vers zéro ou sont égaux à 0,04, ce qui est une manifestation de soutien des données à l'hypothèse H_{hasard}. Encore une fois, l'utilisation de la nomenclature de Kass et Raftery (1995) nous permet de classifier les résultats à cette épreuve d'évaluation.

5. Discussion des résultats

Les résultats présentés à la section précédente sont encourageants et permettent de comprendre un peu mieux le comportement du facteur de Bayes pour réaliser la sélection d'hypothèses visant la détection d'un comportement de réponses au hasard chez les étudiants. En effet, l'étude de simulation a montré que le facteur de Bayes semblait détecter adéquatement les réponses au hasard.

Il est important de comprendre que nos résultats avaient surtout comme objectif d'illustrer une méthode plutôt que d'étudier le fonctionnement de celle-ci de façon approfondie. Dans ce contexte, il n'est pas facile de faire des liens avec les autres approches qui concernent à la réponse au hasard. En effet, on ne connaît pas encore la distribution des scores de l'indice de pseudo-chance personnelle *C*, et les indices de détection tels que l_z sont des approches fondées sur l'adéquation des données à un modèle de mesure. Ils ne sont pas construits pour répondre spécifiquement au problème de la réponse au hasard.

Quelques limites doivent cependant être soulignées à propos des analyses effectuées à l'intérieur de cette recherche. Premièrement, les hypothèses émises pour détecter le comportement de réponses au hasard doivent être considérées comme une représentation fondée sur des données théoriques, et non pas empiriques, de la réalité. Il va de soi qu'analyser une série d'items où la probabilité de

Tableau 5
Résultats de 19 étudiants ayant obtenu entre 20 % et 30 % de bonnes réponses au TCALS-II

Proportion de bonnes réponses	Réponses	$FB_{ordre/hasard}$
0,29	1 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 1 0 1 0 0 0 0 1 0 0 0 1 1 0 0 1 1 0 1 0 0 0 0 1 0 0 1 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 0	0,00
0,27	1 1 1 1 0 0 1 1 0 1 0 0 1 0 0 0 0 0 1 1 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 1 0 0 0 0 0 1 0 0	0,00
0,27	0 1 1 0 0 1 1 0 1 1 1 0 1 1 0 0 0 0 0 0 0 1 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 1 0 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0	0,00
0,29	1 0 0 1 1 0 1 1 0 1 1 1 1 0 1 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 1	0,00
0,29	1 1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 1 0 1 1 0 0 0 0 0 1 1 1 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0	0,00
0,29	0 1 1 0 0 1 1 1 0 0 1 0 1 0 0 0 1 0 1 1 0 0 1 0 1 0 0 0 0 1 1 0 1 0 0 0 1 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0 1 0 0 0 0 0 0	0,00
0,26	0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 1 1 0 1 1 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0	0,00
0,25	1 1 0 0 1 1 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 0 1 1 1 0 0 0 1 0 0 1 1 1 1 1 0 1 1 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0	0,00
0,24	0 0 0 1 1 0 1 1 1 0 1 1 0 0 0 0 1 0 0 0 0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0	0,00
0,28	0 1 0 0 1 1 0 1 1 0 1 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 1 0 1 0 0 1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0	0,00
0,26	1 1 1 0 0 1 1 1 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 1 0	0,00
0,28	1 1 1 1 1 1 1 1 1 0 0 1 0 0 0 0 0 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0	0,00
0,25	0 0 1 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1 1 1 0 0 1 0 1 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 0 0 0 1 1 0 0 0 0 0 0 0	0,00
0,27	1 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 1 1 0 1 0 1 0 0 1 1 1 0 1 0 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 1 0 1 0 0 0 0 0 0 0 0 0	0,00
0,26	1 1 1 0 0 1 0 1 0 1 1 0 1 1 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 1 0 0 1 0 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0	0,04
0,24	1 1 0 0 1 1 0 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 1 0 0 0 0 0	0,00
0,25	1 0 0 0 0 0 1 0 1 1 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 1 1 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0 1 0 1 0 0 0 0	0,00
0,27	1 0 0 0 0 0 1 0 1 1 1 0 0 0 0 1 0 0 1 0 0 0 0 1 0 1 1 0 1 0 0 0 0 1 0 0 0 1 0 0 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 1 0 0 0 0 0 0	0,00
0,29	1 1 1 1 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 1 1 0 0 0 1 1 1 0 0 0 0 0 0 1 1 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0	0,00

donner un score de 1 est égale à 0,25 peut sembler quelque peu artificiel. Par exemple, nous savons que des étudiants répondant au hasard pourraient être plus chanceux dans une section d'une épreuve d'évaluation que dans une autre. Considérer que cette probabilité est fixée et ainsi constante pour tout le test peut, certes, être contestable.

Deuxièmement, la qualité des items peut certainement influencer le calcul du facteur de Bayes. Puisque celui-ci repose sur des analyses probabilistes, il est important de procéder à un calcul adéquat de ces probabilités. Par exemple, les postulats d'unidimensionnalité des données et d'indépendance locale des items doivent être respectés si le modèle de Rasch est utilisé.

Troisièmement, le fait d'utiliser une distribution a priori spécifique plutôt qu'une autre peut être critiquable. Bien que la stratégie présentée dans cet article vise à prioriser une distribution vague, donc non informative, il ne faut pas perdre de vue que le choix de cette approche fait en sorte qu'il y a tout de même une part de subjectivité dans ces analyses.

6. Conclusions

Le survol des écrits de recherche est explicite : les administrateurs de tests manquent d'outils pour évaluer adéquatement si un étudiant a répondu au hasard lors de l'administration d'épreuves d'évaluation en éducation. Le but de cet article était de présenter une nouvelle approche permettant de pallier ce problème. Cette méthode est fondée sur l'utilisation du facteur de Bayes pour procéder à l'évaluation d'hypothèses informatives (Hoijtink, 2012 ; Hoijtink, Klugkist et Boelen, 2008). Grâce aux résultats obtenus à l'étude de simulation, nous constatons que le facteur de Bayes permet de bien détecter la réponse au hasard. L'analyse de données provenant du test de classement en anglais langue seconde nous permet aussi d'être optimistes et de croire que le facteur de Bayes peut être utile pour détecter un comportement de réponses au hasard. En effet, nous avons été en mesure de détecter toutes les réponses fondées sur une probabilité de 0,25 d'obtenir une bonne réponse.

Bien qu'il reste encore beaucoup de travail pour comprendre le comportement du facteur de Bayes, nous pouvons affirmer que cette approche présente déjà de nombreux avantages. Premièrement, le facteur de Bayes est fondé sur des hypothèses simples. Deuxièmement, cette approche peut être interprétée selon certains critères (par exemple, la nomenclature proposée par Kass et Raftery en 1995), même si ces critères pourraient être éventuellement mieux reliés à une réalité pratique. Troisièmement, l'approche peut être adoptée pour convenir à des ensembles de données plus petits que ceux généralement utilisés avec d'autres approches.

Certes, d'autres recherches devront être entreprises pour étendre l'applicabilité de l'approche et optimiser l'utilisation du facteur de Bayes. Premièrement, un devis mixte serait pertinent pour examiner plus spécifiquement les patrons de

réponses des sujets qui admettent avoir répondu au hasard et de détecter ces individus avec la méthode présentée dans cet article. Deuxièmement, il serait important de comparer les résultats de la détection à l'aide du facteur de Bayes à ceux d'indices de détection de patrons de réponse inappropriés paramétriques et non paramétriques. Cela n'a pas été fait ici, car nous désirions plutôt présenter la méthode et non pas comparer son efficacité avec d'autres approches. Troisièmement, il serait pertinent d'étendre l'applicabilité de cette méthode à d'autres types de données. Par exemple, il est possible d'envisager une application à des réponses polytomiques. Enfin, il serait pertinent d'utiliser d'autres densités de distribution telles que la distribution de Jeffreys afin d'observer si celles-ci peuvent potentiellement changer la portée de nos résultats.

ENGLISH TITLE • The use of the Bayes factor to identify students who guess at random

SUMMARY • The available methods that permit detecting students who guess at random in learning assessment tests present many limits. For example, person-fit indexes need generally large data matrices and can be used only to detect if a student responds in accordance to a measurement model (e.g., Rasch models).

In this paper, we will present a new approach to identify students who guess at random in learning assessment tests. After discussing limits of some existing methods, we will expose the technical details of the use of the Bayes factor to evaluate a number of informative hypotheses (Hoijtink, 2012 ; Hoijtink, Klugkist and Boelen, 2008). Next, we will apply this Bayes factor to a simulation study and real data sets for illustration purposes. Our result shows that the Bayes factor is a promising way to detect students who guess at random in learning assessment tests.

KEYWORDS • Answer guessing, Bayes factor, learning assessment.

TÍTULO • La utilización del factor de Bayes para identificar a los estudiantes que responden al azar

RESUMEN • Los métodos que permiten detectar las respuestas dadas al azar en la evaluación de aprendizajes presentan algunos límites. Por ejemplo, los índices de detección de patrones de respuestas inapropiadas (*person-fit indexes*) necesitan generalmente grandes bases de datos y solamente permiten decir si un estudiante responde de acuerdo o no con un modelo de medida (por ejemplo, el modelo de Rasch).

En este artículo presentamos un nuevo enfoque que permite identificar a los estudiantes que responden al azar en las pruebas de evaluación de aprendizajes. Después de discutir los límites de los principales enfoques existentes, exponemos los detalles técnicos de la utilización del factor de Bayes para evaluar un número finito de hipótesis informativas. Posteriormente, aplicamos el factor de Bayes a datos simulados y a datos reales obtenidos para fines de ilustración. Los resultados permiten ver que el factor de Bayes es un método prometedor para detectar el comportamiento de respuesta aleatoria.

PALABRAS CLAVE • Respuesta al azar, factor de Bayes, evaluación de aprendizajes.

7. Références

Angoff, W. H. (1974). The development of statistical indices for detecting cheaters. *Journal of the American statistical association*, *69*, 44-49.

Angoff, W. H. (1989). Does guessing really help ? *Journal of educational measurement*, *26*, 323-336.

Ashler, D. (1979). Biserial estimators in the presence of guessing. *Journal of educational and behavioral statistics*, *4*, 325-355.

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

Brassard, P. D. (2011). *Identification des stratégies de sous-classement intentionnel aux tests de classement en anglais, langue seconde, au collégial* (Mémoire de maîtrise non publié). Université du Québec à Montréal.

Casella, G. and George, E. I. (1992). Explaining the Gibbs sampler. *American statistician*, *46*, 167-174.

Cronbach, L. J. (1946). Response set and test validity. *Educational and psychological measurement*, *6*, 475-494.

Drasgow, F., Levine, M. V. and Williams, E. A. (1985). Appropriateness measurement with polytomous item response models and standardized indices. *British journal of mathematical and statistical psychology*, *38*, 67–86.

Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer, L. Guttman, E. A. Suschman, P. F. Lazarsfeld, S. A. Star and J. A. Clausen (eds), *Measurement and prediction* (p. 60-90). Princeton, New Jersey : Princeton University press.

Hambleton, R. K. and Swaminathan, H. (1985). *Item response theory : principles and applications*. Boston, Massachusetts : Kluwer.

Hoijtink, H. J. A. (2012). *Informative hypotheses : theory and practice for behavioral and social scientists*. London, United Kingdom : Chapman and Hall.

Hoijtink, H. J. A., Klugkist, I. and Boelen, P. A. (eds) (2008). *Bayesian evaluation of informative hypotheses*. New York, New York : Springer.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester, United-Kingdom : Wiley.

Jeffreys, H. (1961). *Theory of probability* (3rd edition). Oxford, United Kingdom : Oxford University Press.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied measurement in education*, *16*, 277-298.

Kass, R. E. and Raftery, A. E. (1995). Bayes factor. *Journal of the American statistical association*, *90*, 773-795.

Lanning, K. (1989). Detection of invalid response patterns on the California. Psychological inventory. *Applied psychological measurement*, 13, 45-56.

Laurier, M., Froio, L., Parro, C. et Fournier, P. (1998). *L'élaboration d'un test provincial pour le classement des étudiants en anglais, langue seconde, au collégial*. Québec, Québec : Direction générale de l'enseignement collégial, ministère de l'Éducation du Québec.

Lee, P., Stark, S. and Chernyshenko, O. S. (2014). Detecting aberrant responding on unidimensional pairwise preference tests : an application of lz based on the Zinnes-Griggs Ideal Point IRT Model. *Applied psychological measurement*, *38*, 391-403.

Meijer, R. R. and Sijtsma, K. (2001). Methodology review : evaluating person fit. *Applied psychological measurement*, *25*, 107-135.

Mulder, J., Hoijtink, H. J. A. and Klugkist, I. (2010). Equality and inequality constrained multivariate linear models : Objective model selection using constrained posterior priors. *Journal of statistical planning and inference*, *140*, 887-906.

O'Dell, J. W. (1971). Method for detecting random answer on personality questionnaire. *Journal of applied psychology*, *55*, 380-383.

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., Béland, S., Magis, D., Blais, J.-G. et Brochu, P. (2010). La modélisation des patrons de réponses atypiques à partir de modèles paramétriques multidimensionnels. Communication présentée au Congrès des sciences humaines- XXXVIII. Université Concordia (Montréal).

Raîche, G., Magis, D., Blais, J.-G. et Brochu, P. (2012). Taking atypical response patterns into account. In M. Simon, K. Ercikan et M. Rousseau (eds), Improving large scale assessment in education : theory, issues and practice (p. 238-259). New York, New York : Taylor and Francis.

Reckase, M. D. (1985). The difficulty of test items that measure more than one ability. *Applied psychological measurement*, *9*, 401-412.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous items response data. In W. J. van der Linden and R. K. Hambleton (eds.), *Handbook of modern item response theory* (p. 271-286). New York, New York : Springer.

Reckase, M. D. (2009). *Multidimensional item response theory.* New York, New York : Springer.

Slakter, M. J. (1968). The effect of guessing on objective test scores. *Journal of educational measurement*, *5*, 217-221.

Sotaridona, L. S. and Meijer, R. R. (2002) Statistical properties of the K-index for detecting answer copying. *Journal of educational measurement*, *39*, 115-132.

Votaw, D. F. (1936). The effect of do-not-guess directions on the validity of true-false or multiple-choice tests. *Journal of educational psychology*, *27*, 698-703.

Waller, M. I. (1973). *Removing the effects of random guessing from latent trait ability estimates* (Unpublished doctoral thesis). University of Chicago, Chicago, Illinois.

Waller, M. I. (1983). Modeling guessing behavior : A comparison of two IRT models. *Applied psychological measurement*, *13*, 233-243.

Wollack, J. A. (1997). A nominal response model approach to detect answer copying. *Applied psychological measurement*, *21*, 307-320.

Wollack, J. A. and Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied psychological measurement*, *22*, 144-152.

Correspondance
sebastien.beland@umontreal.ca
raiche.gilles@uqam.ca
david.magis@ulg.ac.be

Contribution des auteurs
Sébastien Béland : 60 %
Gilles Raîche : 20 %
David Magis : 20 %

Ce texte a été révisé par Christophe Chénier.

Texte reçu le : 30 septembre 2013
Version finale reçue le : 20 avril 2015
Accepté le : 22 avril 2015