

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

**Gilles Raïche, Université de Moncton
Jean-Guy Blais, Université de Montréal**

Texte proposé au LABRIPROF

16 décembre 2002

Le personnel enseignant des départements de langues des cégeps et collèges du réseau collégial québécois est à la recherche, depuis longtemps, de moyens simples et efficaces pour, sinon contrer, du moins amenuiser le comportement de sous-classement aux tests de classement en anglais langue seconde de la part des étudiants. Plusieurs enseignants ont cru que la solution au problème du sous-classement résidait dans l'élaboration d'un test de classement plus efficace dans l'appréciation du niveau d'habileté de l'étudiant. Malheureusement, comme le démontrent les résultats consécutifs à l'élaboration par Laurier, Froio, Pearo et Fournier (1998) d'un test de classement en anglais langue seconde amélioré, le TCALS II, les étudiants du réseau collégial continuent à tenter de se sous-classer à ce test. D'autres solutions doivent donc être envisagées. Nous sommes à leur quête depuis plusieurs années (Raïche, 2000, 2002a).

Nous présentons ici les étapes, ainsi que les résultats, d'un projet de recherche qui propose une solution au problème (Raïche, 2002b). Le projet a été soutenu financièrement par le programme d'aide à la recherche sur l'enseignement et l'apprentissage du ministère de l'Éducation du Québec et a été réalisé au Collège de l'Outaouais.

Objectif spécifique du projet

Tenant compte du grand nombre d'administrations de test de classement dans chacun des collèges, soit plus de mille par an à l'intérieur d'un collège de taille moyenne, il est

nécessaire d'utiliser une formule simple pour dépister les tentatives de sous-classement. Selon nous, l'analyse de la cohérence des réponses à chacune des questions permettrait de détecter des patrons de réponses plutôt étranges et ainsi peu probables. Certains auteurs se sont intéressés à cette problématique et ont proposé des indices numériques de scores plutôt suspects à un test (Cizek, 1999; Reise et Flannery, 1996). C'est l'une des pistes des plus prometteuses à envisager pour pallier ce problème. Maintes mesures d'ajustement inadéquat ont été proposées. Ces mesures ne s'adressent toutefois pas au problème spécifique du sous-classement. Habituellement, elles visent plutôt à déterminer de façon globale le mauvais ajustement des réponses au test sans se soucier de la nature du problème : copie, réponse au hasard, mauvaise utilisation de la feuille de réponses, etc. Il semblait donc nécessaire de développer un indice, ou des indices, d'ajustement inadéquat spécifiques au dépistage du sous-classement. C'était l'objectif spécifique de ce projet de recherche.

Pour nous permettre l'atteinte de cet objectif, nous devions, au départ, identifier les stratégies utilisées par les étudiants pour tenter de se sous-classer à au test de classement en anglais, langue seconde, utilisé à l'intérieur du réseau collégial québécois. À cette fin, nous avons organisé une rencontre avec des étudiants du Collège de l'Outaouais. Par la suite, nous avons sélectionné et développés des indices de détection de patrons de réponses suspects qui seront mis à l'essai à partir d'une simulation par ordinateur. Pour les fins du texte, même si nous décrivons plus loin divers indices propices à la détection de patrons de réponses suspects, nous limiterons la présentation des résultats de l'efficacité de la détection à l'indice le plus performant, soit I_z .

Les stratégies de sous-classement des élèves

Seize étudiants inscrits à un cours d'anglais langue seconde de niveau 102 au Collège de l'Outaouais (604-EWG-03) du même groupe-classe ont participé à la rencontre. Il s'agissait d'un cours adapté à la famille des programmes rattachés aux sciences sociales et aux techniques humaines, regroupement, bien sûr, propre au Collège de l'Outaouais.

De ces 16 étudiants, six étaient de sexe masculin et dix, de sexe féminin. Six étaient inscrits dans un programme technique, soit Techniques administratives (410.12), tandis que les dix autres étaient inscrits dans un programme pré-universitaire, soit Sciences humaines (300.12 et 300.13). Enfin, quelques étudiants ont indiqué qu'ils n'ont pas participé à l'administration du TCALS en 2001-2002.

La rencontre, d'une heure, s'est déroulée à la fin de la période d'un cours de niveau 102 le jeudi 14 mars 2002, la semaine suivant la semaine de relâche au Collège de l'Outaouais. Nous avons pris le temps de bien expliquer le contexte et les objectifs de la recherche aux étudiants. Nous nous sommes assurées que les étudiants se souvenaient bien du type de questions dont est composé le TCALS et nous avons répondu aux questions des étudiants pour ensuite présenter le déroulement de la cueillette des informations. La cueillette consistait à obtenir par écrit, dans un premier temps, la réponse à la question suivante : *si vous désiriez vous sous-classer au test de classement en anglais, langue seconde, quelle stratégie, ou quelles stratégies, utiliseriez-vous?* On a bien insisté sur le caractère confidentiel des commentaires écrits. Les étudiants avaient au plus 10 minutes pour répondre par écrit à cette question. On a ensuite recueilli les commentaires des étudiants et, par la suite, on a continué la discussion avec les étudiants, principalement quant aux stratégies éventuelles de sous-classement volontaire

Considérant la nature exploratoire de cette étape de la recherche, l'analyse des résultats se limite au recueil et à l'interprétation des commentaires reçus de la part des étudiants. Par 14 fois, les étudiants ont indiqué qu'ils répondraient au hasard au test. Choisir la mauvaise réponse, avec diverses variantes, a été proposé par dix étudiants : c'est ce que nous nommons, plus loin, la stratégie de réponses inversées. Plusieurs ont souligné qu'ils utiliseraient cette dernière stratégie seulement s'il connaissait la bonne réponse. Malheureusement, puisque nous avons recueilli ces informations par écrit, nous n'avons pas pu découvrir si ces étudiants adopteraient une autre stratégie lorsqu'ils ne connaissent pas la bonne réponse. Six étudiants opteraient pour omettre de répondre à certaines questions, tandis que trois choisiraient de donner la bonne réponse aux questions faciles et la mauvaise réponse aux questions difficiles. Dans ce dernier cas, on peut se demander

comment les étudiants peuvent juger du niveau de difficulté des questions. Certains ont indiqué qu'ils donneraient le même choix de réponse à plusieurs questions du test.

Certaines stratégies proposées ne nécessitent pas l'application d'une procédure bien complexe pour dépister le comportement de sous-classement. C'est le cas notamment des stratégies où l'étudiant omet sciemment de répondre à des questions. Seules quelques recommandations aux collègues, faciles à appliquer, sont alors nécessaires. Toutefois, les stratégies qui ont été le plus fréquemment identifiées, telles que de répondre au hasard ou de choisir la mauvaise réponse, nécessitent la mise en œuvre de procédures de dépistage plus sophistiquées.

Les procédures de dépistage proposées

Les suites à cette étude ont consisté à élaborer des procédures de dépistage de ces deux dernières stratégies : répondre au hasard ou choisir la mauvaise réponse. Ces procédures sont basées sur les modélisations issues de la théorie de la réponse à l'item (Hambleton, Swaminathan et Rogers, 1991). Pour les soins de cette recherche, nous avons retenu la modélisation logistique à un paramètre selon laquelle la probabilité d'obtenir une bonne réponse à un item g , $X_g = 1$, ne dépend que du niveau d'habileté de l'étudiant, θ , et du niveau de difficulté de l'item g , b_g . La fonction correspond à

$$P(X_g = 1 | \boldsymbol{q}) = \frac{1}{1 + e^{-1.7(\theta - b_g)}} \quad \text{(Équation 1)}$$

La probabilité d'obtenir un patron de réponses, X , est égale à

$$P(X | \boldsymbol{q}) = \prod_{g=1}^n P(X_g = 1 | \boldsymbol{q})^{X_g} (1 - P(X_g = 1 | \boldsymbol{q}))^{1 - X_g} \quad \text{(Équation 2)}$$

La figure 1 illustre une courbe caractéristique d'item du modèle à un paramètre. Plus la valeur de b_g est élevée, plus le niveau de difficulté de l'item correspondant est élevé. À la figure 1, le niveau de difficulté de l'item a été fixé à 0, soit un niveau de difficulté moyen. Cette courbe permet d'observer la probabilité d'obtenir une bonne réponse à un item en fonction du niveau d'habileté. Selon le modèle à un paramètre, la probabilité d'obtenir une bonne réponse est nulle lorsque le niveau d'habileté est très faible. Elle est certaine lorsque le niveau d'habileté est élevé.

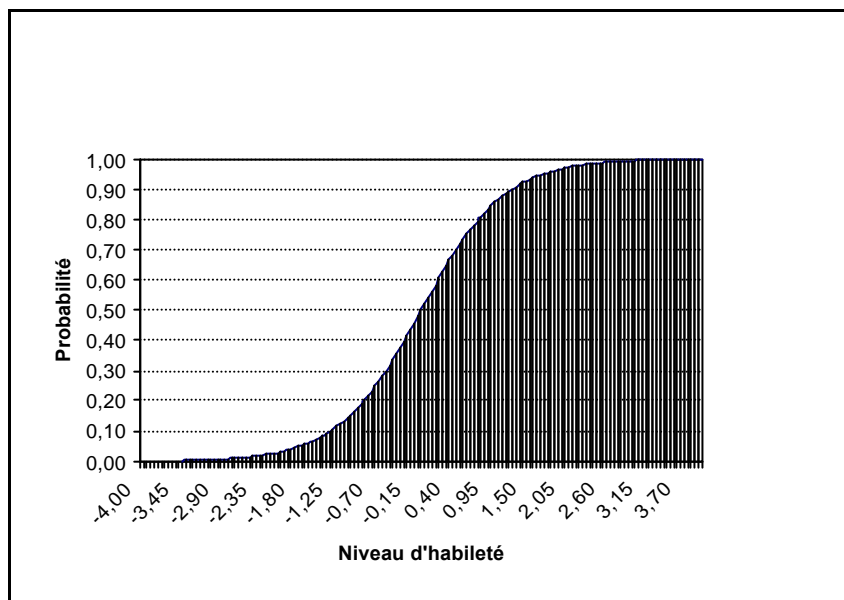


Figure 1 Courbe caractéristique d'item de la modélisation logistique à un paramètre ($b_g = 0$)

La méthode d'estimation du niveau d'habileté qui a été retenue est la méthode de l'espérance a posteriori (EAP). La méthode de l'espérance a posteriori utilise une moyenne comme estimateur du niveau d'habileté (Raïche, 2001a, p. 66). L'estimateur est calculé selon :

$$\hat{\epsilon} = \frac{\int_{q=-\infty}^{\infty} \mathbf{q} f(\mathbf{q}) P(X_g = 1 | \mathbf{q})^{X_g} P(X_g = 0 | \mathbf{q})^{1-X_g} dq}{\int_{q=-\infty}^{\infty} f(\mathbf{q}) P(X_g = 1 | \mathbf{q})^{X_g} P(X_g = 0 | \mathbf{q})^{1-X_g} dq} \quad (\text{Équation 3})$$

avec comme variance

$$S_{\hat{\epsilon}} = \frac{\int_{q=-\infty}^{\infty} (\mathbf{q} - \hat{\epsilon})^2 f(\mathbf{q}) P(X_g = 1 | \mathbf{q})^{X_g} P(X_g = 0 | \mathbf{q})^{1-X_g} dq}{\int_{q=-\infty}^{\infty} f(\mathbf{q}) P(X_g = 1 | \mathbf{q})^{X_g} P(X_g = 0 | \mathbf{q})^{1-X_g} dq}, \quad (\text{Équation 4})$$

où $f(\mathbf{q})$ correspond à la distribution de probabilité a priori du niveau d'habileté. Le calcul est réalisé à partir d'une approximation numérique (Baker, 1992, p. 210-211; Bock et Mislevy, 1982; Raïche, 2001a, p. 130-131).

Généralement, la distribution de probabilité a priori retenue est normale et centrée réduite ($N(0, 1)$). Toutefois, nous avons pu constater que le biais de l'estimateur du niveau d'habileté peut être assez important lorsqu'une distribution $N(0, 1)$ est retenue et que le niveau d'habileté s'éloigne de 0. Récemment, Raïche et Blais (2002a; 2002b) ont proposé une amélioration à la méthode de l'espérance a posteriori. Ils ont ainsi suggéré d'utiliser une distribution a priori $N(0, 1)$ seulement pour le premier item administré. Pour les autres items, la moyenne de la distribution a priori est égale à l'estimateur du niveau d'habileté obtenu à l'item précédent, soit $f(\hat{\epsilon}_{g-1})$. La variance est, pour sa part, maintenue à 1.

Cette stratégie, dans un contexte de testing adaptatif, a permis, à toutes fins pratiques, de rendre le biais nul quel que soit le niveau d'habileté du sujet. Nous avons décidé d'adopter cet ajustement adaptatif à la méthode de l'espérance a posteriori. Nous tenons toutefois à souligner que nous avons remarqué, dans cette recherche, que l'estimateur de

niveau d'habileté ne maintient plus la propriété de statistique suffisante (Rost, 1995, p.37-38; Wright, 1997, p. 35-36), caractéristique prisée de la modélisation logistique à un paramètre de Rasch. Cette propriété permettait à l'estimateur du niveau d'habileté d'avoir une correspondance parfaite avec le score total au test, soit le nombre de bonnes réponses. Il n'est donc plus certain que l'utilisation de la modélisation logistique à un paramètre de Rasch permet à l'estimateur du niveau d'habileté de se retrouver sur une échelle d'intervalle, comme le soutiennent certains auteurs (Fischer, 1995, p. 37-38; Michell, 2002, p. 13-14; Wright, 1997, p. 37-38).

Indices non spécifiques à un comportement particulier

Des indices de détection de patrons de réponses aberrants non spécifiques à un comportement de sous-classement ont été identifiés à travers la littérature. Les indices retenus sont l_z , W , $Zeta$ et c_B^2 . Les auteurs les ont généralement considérés plutôt efficaces à détecter des individus dont la qualité des réponses à des tests ou à des questionnaires est douteuse (Meijer et Sijtsma, 2001).

L'indice standardisé l_o a été proposé par Levine et Rubin en 1979 et a été mis en application à plusieurs reprises par la suite (Drasgow et Guertler, 1987; Drasgow et Levine, 1986; Drasgow, Levine et McLaughlin, 1987, 1991; Drasgow, Levine et Williams, 1985; Levine et Drasgow, 1982, 1983). Cet indice était toutefois difficile à interpréter car il est dépendant de la valeur du niveau d'habileté estimé. Pour pallier ce problème, certains auteurs (Drasgow, Levine et Williams, 1985) lui ont apporté une modification et ont ainsi proposé le nouvel indice standardisé l_z . Cet indice serait distribué selon une distribution de probabilité $N(0,1)$, ce qui lui confère des propriétés très intéressantes au point de vue statistique ainsi qu'au plan pratique. Pour un test constitué de n items, lorsque la probabilité d'obtenir une bonne réponse à chacun des items $P(X_g = 1 | \mathbf{q})$ est obtenue selon la modélisation logistique à un paramètre présentée auparavant, il est égal à

$$l_z = \frac{l_o - E(l_o)}{[Var(l_o)]^{1/2}} \quad \text{(Équation 5)}$$

où

$$l_o = \sum_{g=1}^n [X_g \ln P(X_g = 1 | \mathbf{q}) + (1 - X_g) \ln P(X_g = 0 | \mathbf{q})], \quad \text{(Équation 6)}$$

$$E(l_o) = \sum_{g=1}^n [P(X_g = 1 | \mathbf{q}) \ln P(X_g = 1 | \mathbf{q}) + P(X_g = 0 | \mathbf{q}) \ln P(X_g = 0 | \mathbf{q})] \quad \text{(Équation 7)}$$

et

$$Var(l_o) = \sum_{g=1}^n \left\{ P(X_g = 1 | \mathbf{q}) P(X_g = 0 | \mathbf{q}) \left[\ln \frac{P(X_g = 1 | \mathbf{q})}{P(X_g = 0 | \mathbf{q})} \right] \right\}. \quad \text{(Équation 8)}$$

Lorsque l_z affiche une valeur inférieure ou égale à $-1,65$, le patron de réponses obtenu est considéré comme peu probable avec une probabilité d'erreur inférieure ou égale à $0,05$. Toute valeur supérieure à $-1,65$ est jugée convenable.

De leur côté, Wright et Stone (1979; Wright et Master, 1982; Bond, et Fox, 2001; Li, Olejnik et Bashaw, 1991), ainsi que Smith (1991, 2002, Smith et Suh, 2002), ont proposé un indice légèrement différent, W . Celui-ci est égal à

$$W = \frac{\sum_{g=1}^n [X_g - P(X_g = 1 | \mathbf{q})]^2}{\sum_{g=1}^n [P(X_g = 1 | \mathbf{q})P(X_g = 0 | \mathbf{q})]} \quad \text{(Équation 9)}$$

Les logiciels destinés exclusivement à la modélisation logistique à un paramètre utilisent fréquemment l'indice W . Lorsque la valeur affichée par W est sensiblement supérieure à 1, le patron de réponses est jugé aberrant. La distribution de probabilité de W a toutefois été peu étudiée. Les auteurs supposent que la moyenne de W est égale à 1 tandis que peu d'entre eux s'entendent sur la variance qui lui est associée. C'est pourquoi il est difficile de statuer à quel moment la valeur affichée par W est sensiblement supérieure à 1.

Enfin, Tatsuoka (1996, Birenbaum, Kelly et Tatsuoka, 1992a, 1992b) a proposé un indice de détection des patrons de réponses aberrants dans le but de diagnostiquer des problèmes d'apprentissage chez les étudiants, plus particulièrement des problèmes de conceptions erronées en mathématiques. Cet indice, *Zeta*, s'est toutefois avéré intéressant quant à la détection de toutes formes de patrons de réponses aberrants. Il est calculé comme suit :

$$Zeta = \frac{\sum_{g=1}^n [P(X_g = 1 | \mathbf{q}) - X_g] [P(X_g = 1 | \mathbf{q}) - T(\mathbf{q})]}{\sum_{g=1}^n \{ [P(X_g = 1 | \mathbf{q})P(X_g = 0 | \mathbf{q})] [P(X_g = 1 | \mathbf{q}) - T(\mathbf{q})]^2 \}^{1/2}}, \quad \text{(Équation 10)}$$

où

$$T(\mathbf{q}) = \frac{\sum_{g=1}^n P(X_g = 1 | \mathbf{q})}{n} \quad \text{(Équation 11)}$$

Zeta afficherait une distribution de probabilité $N(0,1)$ et une valeur sensiblement supérieure à 0 ($Zeta \geq 1,65$) indiquerait que le patron de réponses est peu probable avec une probabilité d'erreur inférieure ou égale à 0,05.

L'indice c_B^2 , utilisé à l'intérieur du logiciel Bilog (Mislevy et Bock, 1986; Mislevy et Stocking, 1987) repose sur une logique différente. Il vise à déterminer si le patron de réponses d'un individu s'ajuste bien à la modélisation utilisée en comparant cet ajustement avec une modélisation où on a ajouté un paramètre d'item supplémentaire. Dans notre recherche, cet indice compare la probabilité d'obtenir un patron de réponses telle que calculée selon une modélisation logistique à un paramètre à la probabilité d'obtenir de patron de réponses si on avait appliqué la modélisation logistique à deux paramètres. Une valeur trop élevée de c_B^2 signifierait que l'indice de discrimination d'une étudiante ou d'un étudiant particulier n'est pas constant tout au long du test. L'indice c_B^2 pourrait ainsi détecter l'étudiante ou l'étudiant qui donne des mauvaises réponses plutôt que des bonnes réponses puisque pour celui-ci l'indice de discrimination ne correspond pas à celui de la modélisation logistique à un paramètre. L'indice c_B^2 est égal à

$$c_B^2 = \{-2 * \ln P(X_g = 1 | \mathbf{q})_{1PL}\} - \{-2 * \ln P(X_g = 1 | \mathbf{q})_{2PL}\}, \quad \text{(Équation 12)}$$

où $P(X_g = 1 | \mathbf{q})_{1PL}$ correspond à l'équation de la modélisation logistique à un paramètre, tandis que $P(X_g = 1 | \mathbf{q})_{2PL}$ correspond à l'équation associée à la modélisation logistique à deux paramètres. L'indice c_B^2 se distribuerait selon la distribution de probabilité du c^2 avec un nombre de degrés de liberté égal au nombre d'items du test, ici 85.

Indices spécifiques au comportement de sous-classement

Trois indices spécifiques au comportement de sous-classement ont aussi été élaborés selon l'approche d'identification optimale préconisée par Drasgow, Levine et Zickar (1996). Le premier de ces indices, I_{hasard} , permettrait de détecter un étudiant qui répond au hasard au test, tandis que le second, $I_{inverse}$, permettrait d'identifier un étudiant qui chercherait à donner des mauvaises réponses, soit des réponses inversées. Un troisième indice, I_{sous} , combine en un seul les propriétés des deux indices précédents

Levine et Drasgow (1982, 1988; Drasgow, Levine et Zickar, 1996) suggèrent que la détection de patrons de réponses aberrants gagnerait à être effectuée en ciblant spécifiquement les comportements qui sont directement la cause du mauvais ajustement du patron de réponses. Il suffirait alors de calculer le rapport entre la vraisemblance marginale d'un patron de réponses fourni selon le comportement aberrant spécifique et la vraisemblance marginale d'un patron de réponses obtenu selon un comportement dit normal au test. L'indice serait alors égal à

$$I_{aberrant} = \frac{P_{aberrant}}{P_{normal}}. \quad \text{(Équation 13)}$$

Si $I_{aberrant}$ est supérieur à une valeur critique, le patron de réponses est alors considéré suspect.

Sachant que la probabilité conditionnelle au niveau d'habileté q d'obtenir un patron de réponses normal est calculée à partir d'une des modélisations logistiques de la réponse à l'item,

$$P_{normal|\mathbf{q}} = P(X = x | \mathbf{q}) = \prod_{g=1}^n P(X_g = 1 | \mathbf{q})^{X_g} P(X_g = 0 | \mathbf{q})^{1-X_g}. \quad (\text{Équation 14})$$

et que la probabilité marginale (non conditionnelle) d'obtenir un patron de réponses normal est calculée selon :

$$P_{normal} = \int_{\mathbf{q}=-\infty}^{\infty} P_{normal|\mathbf{q}} d\mathbf{q} \quad (\text{Équation 15})$$

La probabilité conditionnelle d'obtenir un patron de réponses au hasard dépend du nombre de choix de réponses, k , à chacun des items et est pour sa part égale à

$$P_{hasard|\mathbf{q}} = \prod_{g=1}^n \left(\frac{1}{k}\right)^{X_g} \left(\frac{k-1}{k}\right)^{1-X_g}, \quad (\text{Équation 16})$$

tandis que la probabilité marginale d'obtenir un patron de réponses au hasard est égale à

$$P_{hasard} = \int_{\mathbf{q}=-\infty}^{\infty} P_{hasard|\mathbf{q}} d\mathbf{q}. \quad (\text{Équation 17})$$

L'indice de détection d'un patron de réponses obtenu au hasard est alors égal à

$$I_{hasard} = \frac{P_{hasard}}{P_{normal}}. \quad (\text{Équation 18})$$

Puisque les étudiants nous ont signalé qu'ils utiliseraient la stratégie de donner la mauvaise réponse quant ils connaissent la bonne réponse, la vraisemblance conditionnelle de production du patron de réponses associé est égale à

$$P_{inverse\mathbf{q}} = \prod_{g=1}^n P(X_g = 1 | \mathbf{q})^{1-X_g} P(X_g = 0 | \mathbf{q})^{X_g}, \quad (\text{Équation 19})$$

tandis que la vraisemblance marginale correspond à

$$P_{inverse} = \int_{\mathbf{q}=-\infty}^{\infty} P_{inverse\mathbf{q}} d\mathbf{q}. \quad (\text{Équation 20})$$

L'indice associé correspond alors à

$$I_{inverse} = \frac{P_{inverse}}{P_{normal}}. \quad (\text{Équation 21})$$

Enfin, nous avons cru utile d'améliorer la détection par l'utilisation d'un indice de sous-classement global qui permettrait de combiner les stratégies de réponses au hasard et de réponses inversées. En fait, il s'agit de la vraisemblance conditionnelle de la production de réponses au hasard ou de la production de réponses inversées, soit :

$$P_{sous} = \prod_{g=1}^n [P_{hasard} + P_{inverse}] = \prod_{g=1}^n \left[\left(\frac{1}{k} \right)^{X_g} \left(\frac{k-1}{k} \right)^{1-X_g} P(X_g = 1 | \mathbf{q})^{1-X_g} P(X_g = 0 | \mathbf{q})^{X_g} \right]$$

(Équation 22)

La vraisemblance marginale est alors égale à

$$P_{sous} = \int_{q=-\infty}^{\infty} P_{sousq} d\mathbf{q} .$$

(Équation 23)

L'indice associé devient alors égal à

$$I_{sous} = \frac{P_{sous}}{P_{normal}} .$$

(Équation 24)

Une simulation d'étudiantes et d'étudiants qui tentent de se sous-classer

Puisque à partir des données réelles d'une cohorte étudiante nous ne pouvons pas connaître pas à l'avance quels sont les étudiants qui ont cherché à se sous-classer, il est impossible de mettre à l'épreuve les indices de détection directement avec ces étudiants. Seulement une simulation des patrons de réponses normaux et aberrants nous permet de vérifier le pouvoir de détection des différents indices.

Par ordinateur, nous avons simulé 1000 patrons de réponses dans 85 conditions différentes, soit 85 000 patrons de réponses différents. Plus précisément, nous avons produit au hasard 1000 patrons de réponses à 17 niveaux d'habileté, variant, sur une échelle de score z , entre -2,00 et 2,00 par saut de 0,25, et selon cinq conditions : 1) des

patrons de réponses totalement normaux; 2) des patrons de réponses où 10 % des premières questions reçoivent des réponses produites au hasard; 3) des patrons de réponses où 20 % des premières questions reçoivent des réponses produites au hasard; 4) des patrons de réponses où 10 % des premières questions reçoivent des réponses inversées; 5) des patrons de réponses où 20 % des premières questions reçoivent des réponses inversées.

Puisqu'il est nécessaire de détecter un étudiant qui cherche à se sous-classer en ne répondant au hasard ou en ne donnant des mauvaises réponses qu'à quelques questions, seulement 10 % et 20 % des réponses au test sont simulées de cette façon. Les autres questions sont simulées selon un patron de réponse normal. De plus, nous croyons que l'étudiant, en cherchant à contrôler le nombre de mauvaises réponses, n'applique la stratégie de réponses au hasard ou la stratégie de réponses inversées qu'au début du test. Ce sont ainsi les premières questions du test qui sont produites au hasard.

Dans les simulations, de tous les indices de détection évalués, quelle que soit la stratégie de sous-classement adoptée et peu importe le pourcentage de réponses aberrantes, c'est l'indice I_z qui s'est montré le plus efficace à identifier les cas de sous-classement. C'est pourquoi, à l'intérieur de ce texte, nous nous limiterons à la présentation des résultats du dépistage à partir de l'indice I_z . Pour être en mesure de bien évaluer l'efficacité de l'indice I_z , on peut observer les résultats de détection aux figures 2 et 3. La figure 2 nous renseigne sur l'efficacité de l'indice I_z à identifier les cas de sous-classement lorsque 0 %, 10 % et 20 % des réponses sont au hasard : 0 % correspondant à la situation où un étudiant répondrait de manière honnête au test.

Quand aussi peu que 10 % des réponses sont données au hasard, le taux de détection des cas réels de sous-classement est égal ou supérieur à 97 % pour les étudiants des niveaux 3 et 4. L'indice I_z se montre donc très efficace en ce qui a trait aux étudiants dont le niveau d'habileté en anglais, langue seconde, est élevé. C'est justement ces étudiants que nous espérons pouvoir identifier par cette procédure. Le résultat est supérieur à ce que nous avons espéré. Lorsque la stratégie de réponses au hasard est utilisée, au niveau 2, on peut

tout de même détecter un peu plus de la moitié des cas (52 %), tandis que seulement 26 % des cas peuvent être dépistés au niveau 1.

Lorsque 20 % des réponses sont données au hasard, ces pourcentages augmentent considérablement. Même au niveau 1, au moins la moitié des cas de sous-classement sont détectés, tandis qu'aux niveaux 2, 3 et 4 le taux de détection est toujours supérieur ou égal à 87 % quelle que soit la stratégie adoptée.

Cependant, lorsque la stratégie de réponses inversées est considérée, les taux de détection augmentent considérablement. C'est ce que présente la figure 3. Ainsi, lorsque 10 % des réponses sont inversées, le taux de détection est égal à 100 % aux niveaux 3 et 4. Aux niveaux 1 et 2, les taux de détection sont respectivement de 30 % et 66 %. À ce moment, avec aussi peu que 10 % de réponses aberrantes, l'efficacité de la procédure d'identification est encore plus importante qu'avec la stratégie de réponses au hasard. Quand 20 % des réponses sont inversées le taux de détection est égal ou supérieur à 99 % aux niveaux 2, 3 et 4, tandis qu'il est de 88 % au niveau 1.

L'indice I_z constitue donc un indicateur très efficace pour nous permettre de détecter l'utilisation d'une stratégie de réponses au hasard ou inversées. De plus, son efficacité est à son maximum chez les étudiants qui, selon nous, risquent le plus souvent de désirer se sous-classer, soit ceux dont le niveau d'habileté en anglais est élevé (niveaux 3 et 4)

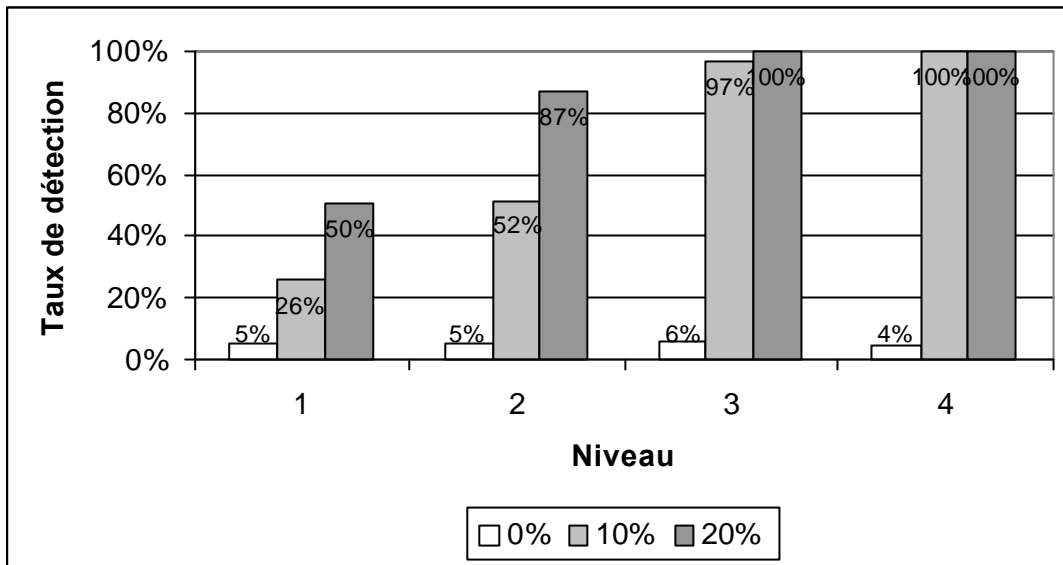


Figure 2. Taux de détection des patrons de réponses aberrants lorsque les réponses sont données au hasard en fonction du niveau de classement réel et de la stratégie adoptée

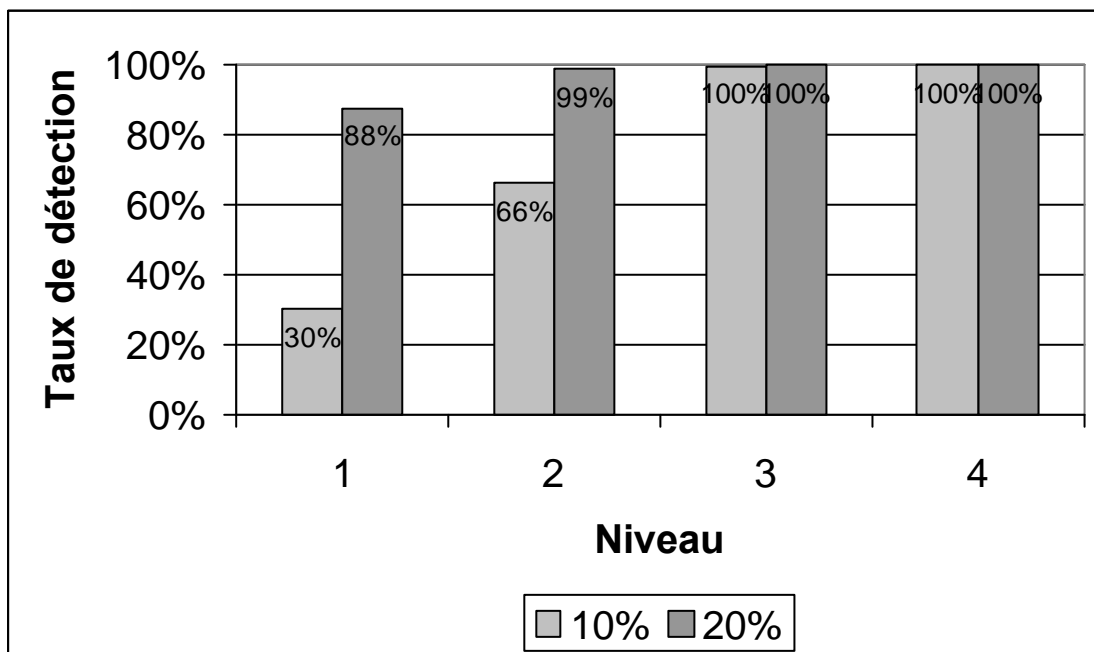


Figure 3. Taux de détection des patrons de réponses aberrants lorsque des réponses incorrectes sont données en fonction du niveau de classement réel et de la stratégie adoptée

Recommandations

Quelques recommandations s'imposent. Premièrement, il est clair que nous jugeons approprié d'utiliser l'indice I_z pour détecter les cas de sous-classement intentionnels de la part des étudiants. Il ne faut toutefois pas oublier que celui-ci ne détecte pas spécifiquement les cas de sous-classement. Il ne faut donc pas prendre pour acquis que l'identification d'un étudiant par cet indice signifie nécessairement que celle-ci ou celui-ci a tenté de sous-classer. Le patron de réponses peut provenir d'un étudiant qui n'était tout simplement pas disposé à répondre au test ce jour-là. Il se peut que l'administration du test se soit déroulée dans des conditions peu usuelles. Cependant, à chaque fois que l'indice I_z présente une valeur suspecte, il est totalement justifié de mettre en doute le résultat au test et d'investiguer plus à fond ce qui s'est passé. Des procédures appropriées, qui selon nos observations au Collège de l'Outaouais ne devraient toucher au plus que 10 % des nouvelles admissions, adaptées aux réalités de l'institution d'enseignement, doivent être planifiées.

En second lieu, il faut prévoir un processus de détection des omissions et des répétitions consécutives du même choix de réponse aux questions du TCALS II. Un étudiant qui omet de répondre ou répète consécutivement un même choix de réponse à plus de 10 % des questions du test, devrait attirer l'attention des intervenants.

En dernier lieu, il est nécessaire d'appliquer un mécanisme de suivi à l'identification des comportements suspects au test. On pourrait, par exemple, rappeler les étudiantes et les étudiants détectés et leur administrer une composition écrite ou une entrevue orale. D'ailleurs, le seul fait de rappeler des étudiants pour une seconde procédure de classement envoie un message assez clair à ceux-ci : il est possible de détecter les tentatives de sous-classement intentionnel. Selon les échos que nous avons reçus en Outaouais, l'application de cette procédure a fait jaser beaucoup dans la région. À elle

seule, cette pratique pourrait éventuellement être suffisamment dissuasive pour diminuer la fréquence des tentatives de sous-classement.

Les suites au projet

Pour assurer la réalisation de nos travaux, nous avons dû élaborer un outil informatique spécialisé. Toutefois, le programme actuel est actuellement un peu lourd à utiliser et nécessitera quelques améliorations pour permettre son utilisation dans la plupart des collèges. Nous espérons, d'ici peu, développer une version adaptée du logiciel pour les intervenants des services informatiques du réseau collégial.

Nous désirons aussi continuer les travaux suites à l'administration au Collège de l'Outaouais d'une composition écrite et vérifier quels ont été les cas de sous-classement détectés par l'administration de cette composition écrite. Il nous sera alors possible d'identifier formellement des étudiants sous-performants, mais surtout d'étudier les patrons de réponses de celles-ci et de ceux-ci. Nous croyons pouvoir ainsi percer plus à fond les stratégies utilisées par eux.

Conjugué à nos recherches antérieures sur le testing adaptatif par ordinateur, ce projet de recherche prépare aussi le terrain à des travaux d'élaboration d'une procédure adaptative qui permettrait d'administrer un test dont le niveau de difficulté des items varierait non seulement avec l'estimateur du niveau d'habileté de l'étudiant, mais aussi avec le niveau d'aberrance du patron de réponses. Actuellement, à notre connaissance, aucun travail en ce sens n'a été entrepris. Nous espérons, alors, appliquer ces résultats au développement d'un environnement informatique destiné à soutenir une version adaptative du TCALS II. Cette version adaptative pourrait alors intégrer des mécanismes de dépistage des tentatives de sous classement au test.

Références

- Baker, F.B. (1992). Item response theory : Parameter estimation techniques. New York : Marcel Dekker.
- Birenbaum, M., Kelly, A.E., Tatsuoka, K.K. (1992a). *Diagnosing knowledge states in algebra using the rule space model*. Research Report RR-92-57-ONR, Princeton : Educational Testing Service.
- Birenbaum, M., Kelly, A.E., Tatsuoka, K.K. (1992b). *Toward a stable diagnostic representation of students' errors in algebra*. Research Report RR-92-58-ONR, Princeton : Educational Testing Service.
- Bock, R.D. et Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431-444.
- Bond, T.G. et Fox, C.M. (2001). *Applying the Rasch model : Fundamental measurement in the human sciences*. Mahwah, NJ : Laurence Erlbaum Associates.
- Cizek, G.J. (1999), *Cheating on test: how to do it, detect it, and prevent it*. Mahwah : Laurence Erlbaum Associates.
- Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, 6(3), 297-308.
- Drasgow, F. et Guertler, E. (1987). A decision-theoretic approach to the use of appropriateness measurement for detecting invalid test and scale scores. *Journal of Applied Psychology*, 72(1), 10-18.
- Drasgow, F. et Levine, M.V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement*, 10(1), 59-67.

- Drasgow, F., Levine, M.V. et McLaughlin, M.E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement*, 11(1), 59-79.
- Drasgow, F., Levine, M. et McLaughlin, M.E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement*, 15(2), 171-191.
- Drasgow, F. Levine, M.V. et Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38, 67-86.
- Drasgow, F., Levine, M.V. et Zickar, M.J. (1996). Optimal identification of mismeasured individuals. *Applied Measurement in Education*, 9(1), 47-64.
- Fischer, G.H. (1995). Derivations of the Rasch model. Dans G.H. Fischer et I.W. Molenaar (Éds) : *Rasch models – Foundations, recent developments, and applications*. New York : Springer-Verlag.
- Fournier, P. (1992). *Pour un tests incontestable : rapport de recherche sur les tests de classement en anglais (langue seconde) au collégial*. Québec : ministère de l'Éducation, Direction générale de l'enseignement collégial.
- Hambleton, R.K., Swaminathan, H. et Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park : Sage.
- Laurier, M., Froio, L., Paero, C. et Fournier, M. (1998). *Test de classement d'anglais langue seconde au collégial*. Montréal : Collège de Maisonneuve.

- Levine, M.V. et Drasgow, F. (1982). Appropriateness measurement : Review, critique and validating studies. . *British Journal of Mathematical and Statistical Psychology*, 35, 42-56.
- Levine, M.V. et Drasgow, F. (1983). Appropriateness measurement : Validating studies and variable ability models. Dans D.J. Weiss (Éd.) : *New horizons in testing – Latent trait test theory and computerized adaptive testing*. New York : Academic Press.
- Levine, M.V. et Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, 53(2), 161-176.
- Li, M.N.F. et Olejnik, S. (1997). The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement*, 21(3), 215-231.
- Meijer, R.R. et Sijtsma, K. (2001). Methodology review : Evaluating person fit. *Applied psychological Measurement*, 25(2), 107-135.
- Michell, J. (2002). *Measurement: A beginner's guide*. Texte présenté au congrès annuel de l'American Educational Research Association tenu à New Orleans.
- Mislevy, R.J., Bock, R.D. (1986). *PC-BILOG : Item analysis and test scoring with binary logistic models*. Mooresville, Indiana : Scientific Software Inc.
- Mislevy, R.J., Stocking, M.L. (1987). *A consumer's guide to LOGIST and BILOG*. Research report RR-87-4. Princeton, NJ : Educational testing service.
- Raïche, G. (2000). *Le sous-classement des élèves aux tests de classement en anglais langue seconde au Collège de l'Outaouais*. Gatineau : Collège de l'Outaouais.
- Raïche, G. (2001a). *La distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt : selon l'erreur type*

et selon le nombre d'items administrés. Thèse de doctorat inédite. Montréal : Université de Montréal.

Raïche, G. (2002a). Pour dépister les étudiants qui cherchent à sous-performer au test de classement en anglais langue seconde dans le réseau collégial québécois. *Bulletin de l'Association pour la recherche au collégial*, 15(3), 8-9.

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

Raïche, G. et Blais, J.-G. (2002a). Étude de la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif en fonction de deux règles d'arrêt dans le contexte de l'application du modèle de Rasch. *Mesure et évaluation en éducation*, 24(2-3).

Raïche, G. et Blais, J.-G. (2002b). *Practical considerations about expected a posteriori estimation in adaptive testing : Adaptive a priori, adaptive correction for bias, and adaptive integration interval.* Texte présenté lors du 11^e Biennial International Objective Measurement Workshop tenu à New Orleans. [À paraître sur ERIC]

Reise, S.P. et Flannerey, W.P. (1996). Assessing person-fit on measures of typical performance. *Applied Measurement in Education*, 9(1), 9-26.

Rost, J. (1995). The growing family of Rasch models. Dans A. Boomsma, M.A.J. van Duijn et T.A.B. Snijders (Éds) : *Essays on item response theory.* New York : Springer-Verlag.

Smith, R.M. (1991). The distributional properties of Rasch item fit statistics. *Educational and Psychological Measurement*, 51(3), 541-565.

Smith, R.M. (2002). *The family approach to assessing fit in Rasch measurement*. Texte présenté au 11^e International Biennial Objective Measurement Workshop tenu à New Orleans.

Smith, R. et Suh, K.K. (2002). *Rasch fit statistics as a direct test of the invariance parameter estimation*. Texte présenté au 11^e International Biennial Objective Measurement Workshop tenu à New Orleans.

Tatsuoka, K. (1996). Use of generalized person-fit indexes, zetas for statistical pattern classification. *Applied Measurement in Education*, 9(1), 65-76.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement : Issues and Practice*, 16(4), 33-45 et 52.

Wright, B.D. et Masters, G.N. (1982). *Rating scale analysis : Rasch measurement*. Chicago : MESA Press.

Wright, B.D. et Stone, M.H. (1979). *Best test design : Rasch measurement*. Chicago : MESA Press.