Mesure et évaluation en éducation

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

Gilles Raîche, professeur et Jean-Guy Blais, professeur

Volume 24, numéro 2-3, 2001

URI : https://id.erudit.org/iderudit/1091168ar
DOI : https://doi.org/10.7202/1091168ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Raîche, G. & Blais, J.-G. (2001). 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), 23–39. https://doi.org/10.7202/1091168ar

Résumé de l'article

Cet article s'intéresse à l'application des modélisations issues de la théorie de la réponse à l'item au testing adaptatif par ordinateur. Plus spécifiquement, il s'intéresse à l'impact de la variation des critères retenus pour la règle d'arrêt sur la distribution de probabilité de certaines statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif. Les règles d'arrêt considérées sont de deux types : selon l'erreur-type de l'estimateur du niveau d'habileté et selon le nombre d'items administrés.

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

Gilles Raîche, professeur
Université de Moncton

Jean-Guy Blais, professeur
Université de Montréal

MOTS-CLÉS : *Testing* adaptatif, théorie de la réponse à l'item, règle d'arrêt, espérance *a posteriori*, distribution d'échantillonnage, asymétrie, kurtose, estimateur du niveau d'habileté

Cet article s'intéresse à l'application des modélisations issues de la théorie de la réponse à l'item au testing adaptatif par ordinateur. Plus spécifiquement, il s'intéresse à l'impact de la variation des critères retenus pour la règle d'arrêt sur la distribution de probabilité de certaines statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en testing adaptatif. Les règles d'arrêt considérées sont de deux types : selon l'erreur-type de l'estimateur du niveau d'habileté et selon le nombre d'items administrés.

KEY WORDS: Adaptive testing, item response theory, stopping rule, esperance a posteriori, sampling distribution, asymetry, kurtosis, proficiency estimate

This article presents results concerning the application to adaptive testing of modelisations coming from item response theory. More specifically, interests are on the impact of the variation of the criterias retained for the stopping rules on the probability distribution of certain statistics associated with the sampling distribution of the proficiency estimate in adaptive testing. Stopping rules considered are of two kinds : according to the standard error of the proficiency estimate and according to the number of administered items.

Notes des auteurs. Toute correspondance peut être adressée comme suit : Gilles Raîche, professeur, Université de Moncton, Faculté des sciences de l'éducation, Pavillon Jeanne-de-Valois, Moncton, N.-B., E1A 1E9 ; tél : (506) 858-4425 ; télécopieur : (506) 858-4317 ; courriel : raicheg@umoncton.ca – Jean-Guy Blais, professeur, Université de Montréal, Faculté des sciences de l'éducation, case postale 6128, succursale Centre-Ville, Montréal, QC, H3C 3J7 ; tél : (514) 343-7527 ; courriel : blaisjg@scedu.umontreal.ca.

Introduction

Depuis que les examens et les tests ont été introduits pour le recrutement et la sélection des officiers et bureaucrates du gouvernement en Chine en 210 AC, on a produit très peu de façons de tester, d'examiner ou d'apprécier les apprentissages, les connaissances, les compétences, les habiletés ou les performances des individus. Pendant une bonne partie du XX^{e} siècle, l'approche des tests papier-crayon a été la stratégie privilégiée pour recueillir certaines données auprès d'individus, de sujets ou de candidats. Un test est ici défini comme *une épreuve psychotechnique normalisée impliquant une tâche à accomplir, identique pour tous les sujets examinés, avec technique précise pour l'évaluation du succès ou de l'échec, ou par la notation numérique de la réussite* (Legendre, 1988, pp. 604-605).

Lorsque tous les items d'un test sont identiques pour chaque candidat, tant selon le contenu des items que selon le nombre d'items, on dit que le test est fixe et invariable. Lorsque le nombre d'items varie et que ceux-ci peuvent être différents pour chacune des personnes, le test entre alors dans la catégorie des tests dits adaptatifs, c'est-à-dire dans la catégorie des tests qui sont sur mesure en fonction des problèmes rencontrés par le sujet à chaque item qui lui est présenté. La séquence d'items est ainsi dépendante du point de départ et des réponses données à chaque item présenté.

Le développement accéléré des ordinateurs ces vingt dernières années a aussi permis une évolution fulgurante des tests adaptatifs. En effet, alors que les modèles et les stratégies de branchement pour le *testing* adaptatif existent depuis près de cinquante ans, ce n'est que récemment, grâce à la puissance de calcul et de stockage des ordinateurs personnels, que cette stratégie de *testing* a pu déployer pleinement les avantages qu'elle possède sur les stratégies conventionnelles. Les résultats présentés dans ce texte font suite à une recherche qui s'est intéressée à certaines caractéristiques du *testing* adaptatif par ordinateur lorsque les réponses des candidats sont notées de façon dichotomique (0, 1) et lorsque les réponses sont modélisées avec la théorie de la réponse à l'item (TRI).

Testing adaptatif

Le *testing* adaptatif offre plusieurs avantages par rapport aux tests papier-crayon fixes et invariables. L'une des caractéristiques les plus importantes du *testing* adaptatif est de permettre l'administration d'items dont le niveau de difficulté correspond au niveau d'habileté de la personne passant le test. À l'opposé des tests papier-crayon fixes et invariables, où tous les items du test sont administrés sans égard au niveau d'habileté de la personne, le *testing* adaptatif permet l'administration de tests sur mesure, de façon à ce que le niveau de difficulté des items de ce test ne soit ni trop élevé, ni trop faible (Weiss, 1983, p. 5). Hambleton, Swaminathan et Rogers (1991, p. 145) soulignent que le nombre d'items administrés, ainsi que la durée de l'administration, sont ainsi réduits par rapport à une version papier-crayon du test, sans que la précision de l'estimateur du niveau d'habileté diminue proportionnellement. Selon Lord (1980, p. 201), le *testing* adaptatif devrait d'ailleurs permettre d'obtenir un estimateur plus précis du niveau d'habileté, plus spécifiquement lorsque le niveau d'habileté est faible ou élevé.

En *testing* adaptatif, chaque personne peut recevoir une version du test dont les items ont un niveau de difficulté adapté à son niveau d'habileté et dont la séquence des items peut varier d'une personne à une autre. Toutefois, cette caractéristique du *testing* adaptatif fait en sorte que le nombre de bonnes réponses au test ne permet plus de comparer les personnes entre elles puisqu'elles obtiennent toutes, selon certains auteurs (Weiss, 1985, p. 776), environ le même pourcentage de bonnes réponses aux items. Il serait plus approprié d'estimer le niveau d'habileté indépendamment du choix particulier des items d'une version du test (Hambleton & Swaminathan, 1987, p. 296).

Des propositions de modélisation de la réponse à l'item, telles que celles décrites par Goldstein et Wood (1989) ou par Thissen et Steinberg (1986), ont facilité l'utilisation du *testing* adaptatif en permettant justement d'estimer le niveau d'habileté indépendamment du choix particulier des items d'une version du test. Toutefois, les calculs exigés par les différentes modélisations mathématiques proposées ne permettaient pas, jusqu'à tout récemment, l'application du *testing* adaptatif à des situations réalistes, pendant des opérations d'inscription scolaire, par exemple. L'accessibilité à un ordinateur central ou à un mini-ordinateur n'était pas toujours possible en raison à la fois des coûts d'utilisation et de la disponibilité physique des appareils. Les micro-ordinateurs offrent maintenant une puissance de calcul suffisante pour supporter ces propositions de modélisation, et à un coût abordable.

Problématique

Plusieurs caractéristiques des tests adaptatifs ont reçu une attention particulière et ont été conséquemment l'objet d'études. Ainsi, certains auteurs ont effectué des comparaisons entre l'estimateur du niveau d'habileté obtenu à partir de différentes propositions de modélisation de la réponse à l'item (Dodd, de Ayala & Koch, 1995) et de différentes méthodes d'estimation (Chen, Hou & Dodd, 1998). D'autres ont étudié l'influence de la dimensionnalité de la banque d'items (de Ayala, 1992), la conformité au postulat d'indépendance locale (Mislevy & Chang, 2000) ou des caractéristiques de la banque d'items et des différentes règles d'arrêt sur l'estimateur du niveau d'habileté (Dodd, Koch & de Ayala, 1993). La comparaison de certaines règles de sélection des items a été effectuée (Chang & Ying, 1999) comme celle des méthodes pour évaluer le fonctionnement différentiel des items (*differential item functionning*, DIF) (Zwick, 1997) ou les indices d'ajustement de l'estimateur du niveau d'habileté (*person fit*) (van Krimpen-Stoop & Meijer, 1999). Cependant plusieurs aspects du *testing* adaptatif par ordinateur restent à étudier.

Parmi les aspects du *testing* adaptatif qui restent à étudier, nous avons choisi d'examiner les caractéristiques des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté. Ces caractéristiques permettent de se prononcer sur le sens à donner à l'estimateur du niveau d'habileté obtenu en *testing* adaptatif et, conséquemment, à définir les conditions d'application du *testing* adaptatif. Les caractéristiques des second (erreur-type), troisième (asymétrie) et quatrième (kurtose) moments centrés associés à cette distribution d'échantillonnage sont analysés.

Si la distribution d'échantillonnage de l'estimateur du niveau d'habileté se distribue selon une loi de probabilité normale, la précision de cet estimateur, lorsqu'elle est mesurée par son erreur-type, permet de déterminer un intervalle de confiance autour de l'estimateur du niveau d'habileté. La détermination de cet intervalle de confiance n'est toutefois valide que lorsque la distribution d'échantillonnage de l'estimateur du niveau d'habileté est symétrique et qu'elle est mésokurtique, soit ni surélevée ni aplatie. Raîche (2000, p. 148-156) a étudié l'impact des coefficients d'asymétrie (g1) et de kurtose (g2) sur l'intervalle de confiance à 68 % autour de la moyenne. Selon ses résultats, un coefficient de kurtose de -0,40, associé à un coefficient d'asymétrie de 0,50, augmente de 10% l'intervalle de confiance autour de la moyenne. À ce moment, une erreur-type de 0,30 correspondrait plutôt à une erreur-type de 0,33. Lorsque le coefficient d'asymétrie est égal à 0,75, l'augmentation de l'intervalle de

confiance est de 20 % et la même erreur-type correspondrait alors à 0,36. De plus, la différence entre la moyenne et la médiane augmente rapidement avec l'augmentation du coefficient d'asymétrie. À titre d'illustration, aux mêmes valeurs précitées de la kurtose et de l'asymétrie, la différence entre la moyenne et la médiane correspond respectivement à 10 % et à 17 % de l'erreur-type. Raîche indique qu'il faut ainsi se méfier quant aux interprétations reliées à une distribution normale lorsque les valeurs de l'asymétrie et de la kurtose dépassent 0,40 en valeur absolue. Quelles valeurs de l'asymétrie et de la kurtose retrouve-t-on dans un test adaptatif? À notre connaissance, aucune recherche ne s'est intéressée à cette question.

L'objectif de cette recherche est précisément d'étudier les caractéristiques des second, troisième et quatrième moments centrés associés à la distribution d'échantillonnage de l'estimateur du niveau d'habileté en *testing* adaptatif. Ces caractéristiques seront étudiées en fonction de deux règles d'arrêt d'utilisation fréquente *:* selon le nombre d'items administrés et selon la détermination *a priori* de l'erreur-type de l'estimateur du niveau d'habileté.

Méthode

Dans le but d'exercer un contrôle strict de la situation de *testing* et pour qu'un nombre important d'observations soit disponible, nous proposons une simulation informatisée. Ensuite, puisque les résultats obtenus peuvent varier lorsque les caractéristiques des items composant la banque d'items sont modifiées, nous effectuerons une analyse sur des données qui ne tiennent pas compte de la nature de la banque d'items de manière à neutraliser l'effet de sa composition. La banque d'items sera ainsi composée de façon à ce que le choix des valeurs des paramètres du modèle privilégié (par exemple, la difficulté, la discrimination ou la pseudo-chance) dépende uniquement de la règle de sélection. Urry propose une règle de sélection qui satisfait cette condition (1970, p. 82*;* Thissen & Mislevy, 1990, p. 111*)*. Selon cette règle, le prochain item administré correspond à un item dont le niveau de difficulté est égal à la valeur de l'estimateur du niveau d'habileté obtenue après l'administration de l'item précédent. Le résultat de cette stratégie n'est d'ailleurs pas incompatible avec des situations réelles de *testing* adaptatif lorsque sont considérées des modélisations permettant la génération de tous les items d'un univers, telles que le proposent Bejar (1993) ou Embretson (1999).

De plus, pour éviter l'impact du paramètre de discrimination sur la sélection des items à administrer décrit par Thissen et Mislevy (1990, pp. 112-113), on utilisera une modélisation logistique à un paramètre de la réponse à l'item. Cela nous permettra aussi de réaliser l'étude sans que les valeurs des paramètres de discrimination et de pseudo-chance n'aient à être contrôlées. Les résultats seront donc principalement généralisables aux tests où une modélisation logistique à un paramètre de la réponse à l'item (modèle de Rasch) est appliquée.

Déroulement de la simulation

Simulation des réponses aux items

Pour les fins de ce travail, les valeurs du niveau d'habileté utilisées pour effectuer les simulations d'un test adaptatif sont aléatoires et sont tirées d'une distribution de probabilité qui suit une loi normale N(0,1). Un échantillon de taille 2 000 est généré de façon aléatoire. C'est une taille d'échantillon convenable considérant que les tailles d'échantillon utilisées dans la documentation consultée varient entre 100 et 10 000, avec une valeur médiane de 500.

La simulation d'un test adaptatif est appliquée à chacune des 2 000 valeurs aléatoires du niveau d'habileté. Chaque simulation est réalisée selon une méthode de génération d'une réponse aux items fréquemment utilisée à l'intérieur des recherches sur la théorie de la réponse à l'item (Nicewander & Thomasson, 1999, p. 244). Selon cette méthode, pour chaque valeur du niveau d'habileté générée au hasard, on obtient la réponse à chacun des items en calculant la probabilité d'obtenir une bonne réponse à l'item, $P(r = 1|\theta)$, en tenant compte du paramètre de difficulté de l'item ainsi que de la valeur du niveau d'habileté. La façon de déterminer la valeur du paramètre de difficulté de l'item est expliquée à la section traitant de la règle de départ et de la sélection des items. Cette probabilité est ensuite comparée à un nombre aléatoire x, compris entre 0 et 1, tiré d'une distribution de probabilité uniforme U(0,1). Si la probabilité d'obtenir une bonne réponse à l'item $P(r = 1|\theta)$ est supérieure au nombre aléatoire x, la réponse à l'item prend la valeur 1, soit une bonne réponse. Sinon, la réponse à l'item prend la valeur 0, soit une mauvaise réponse. Ainsi,

$$\text{si } P(r = 1 \mid \theta) \geq x, \text{ alors } r = 1, \text{ sinon } r = 0 \qquad [1]$$

où $P(r = 1|q)$ correspond à la fonction logistique à un paramètre :

$$P(r = 1 \mid \theta) = \frac{1}{1 + e^{-D(\theta - b)}} \qquad [2]$$

Une constante D, égale à 1,07, est utilisée dans cette fonction pour faire en sorte que les valeurs obtenues se rapprochent le plus possible de celles qui proviendraient d'une fonction de modélisation de la réponse à l'item basée sur la loi normale (Baker, 1992, p. 16)*;* la différence est d'au plus 0,01.

Règles de départ et de suite

À tous les niveaux d'habileté simulés, le test débute par l'administration d'un item dont le niveau de difficulté, b_1, est égal à 0, soit la moyenne de la distribution *a priori*. L'utilisation d'un niveau de difficulté, b_1, constant à tous les niveaux d'habileté simulé permet de s'assurer que l'estimateur du niveau d'habileté obtenu ne varie pas en fonction du niveau de difficulté du premier item administré.

Nous utilisons la méthode de Urry (Thissen & Mislevy, 1990, p. 111*;* Urry, 1970, p. 82) pour sélectionner le prochain item. Selon cette méthode, le prochain item à administrer, b_{j+1}, correspond à un item dont le niveau de difficulté est égal à l'estimateur provisoire du niveau d'habileté obtenu selon la méthode de l'espérance *a posteriori*, $EAP_j(\theta)$, après l'administration de l'item j.

$$b_{j+1} = EAP_j(\theta) \qquad [3]$$

Cette règle de sélection des items, lorsque le modèle à un paramètre est utilisé, permet d'obtenir le prochain item qui fournit l'information maximale*;* elle est donc équivalente à la stratégie de maximisation de l'information. De plus, elle permet de faire en sorte que le choix des valeurs du paramètre de difficulté dépende uniquement de la règle de sélection de façon que la composition de la banque d'items ne puisse affecter les valeurs de l'estimateur du niveau d'habileté.

Méthode d'estimation provisoire du niveau d'habileté

L'estimateur provisoire du niveau d'habileté est calculé selon la méthode de l'espérance a posteriori. L'estimateur a posteriori du niveau d'habileté, $EAP_j(\theta)$, est calculé pour chaque valeur j du nombre d'items administrés selon une approximation de l'intégrale correspondante (Baker, 1992, p. 211) :

$$EAP_j(\theta) = \frac{\sum_{k=1}^{q} X_k L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \qquad [4]$$

où X_k est un des q points de quadrature équidistants compris entre θ = -4 et θ = 4 (q = 40), $A(X_k)$ est la pondération associée à chacun des points de quadrature selon une loi de probabilité N(0,1) et

$$L_j(\theta) = \prod_{i=1}^{j} P(r_i \mid \theta, b_i)^{r_i} Q(r_i \mid \theta, b_i)^{1-r_i} \qquad [5]$$

est la vraisemblance (*likelihood*) du patron de réponses, $R = \{r_1 \ldots r_j\}$ après l'administration de j items. Pour faire en sorte que la somme des probabilités soit égale à 1,00 la contrainte suivante est de plus imposée*:*

$$\sum_{k=1}^{q} A(X_k) = 1 \qquad [6]$$

L'intégration est ainsi réalisée selon la méthode de l'histogramme de Mislevy (Baker, 1992, p. 187), avec 40 points de quadrature dont la pondération est égale à la probabilité *a priori* à ces points (Bock & Mislevy, 1982, p. 433*;* de Ayala, Schafer & Sava-Bolesta, 1995, p. 387).

L'erreur-type de l'estimateur du niveau d'habileté est calculée selon :

$$S_{EAP_j(\theta)} = \left[\frac{\sum_{k=1}^{q} (X_k - EAP_j(\theta))^2 L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \right]^{1/2} \qquad [7]$$

Les équations 9 et 10, en concordance avec le calcul des moments centrés proposé par Spiegel (1961, p. 90), sont respectivement utilisées pour réaliser le calcul des estimateurs de l'asymétrie, $g1_{EAP(\theta),}$ et de la

$$g1_{EAP_j(\theta)} = \left[\frac{\sum_{k=1}^{q} (X_k - EAP_j(\theta))^3 L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \right] / S^3_{EAP_j(\theta)} \qquad [8]$$

kurtose, $g2_{EAP(\theta),}$ de la distribution d'échantillonnage de l'estimateur du niveau d'habileté.

$$g2_{EAP_j(\theta)} = -3 + \left[\frac{\sum_{k=1}^{q} (X_k - EAP_j(\theta))^4 L_j(X_k) A(X_k)}{\sum_{k=1}^{q} L_j(X_k) A(X_k)} \right] / S^4_{EAP_j(\theta)} \quad [9]$$

On remarquera que nous n'effectuons pas le calcul de l'erreur-type, de l'asymétrie et de la kurtose sur la distribution d'échantillonnage empirique de l'estimateur du niveau d'habileté. Nous avons plutôt choisi de réaliser le calcul à partir des valeurs théoriques de ces statistiques selon les estimations obtenues par les formules 7, 8 et 9, puisque dans la pratique de l'administration des tests adaptatifs les valeurs empiriques ne sont pas disponibles à chaque estimation du niveau d'habileté.

Règles d'arrêt et méthode d'estimation finale du niveau d'habileté

Dans la simulation, tous les tests se terminent après l'administration de 60 items. Toutefois, la disponibilité des résultats intermédiaires (de 1 à 60 items administrés) permet de connaître la valeur de l'estimateur du niveau d'habileté et de son erreur-type après l'administration de chacun des 60 items. Sont aussi disponibles, conséquemment, les résultats en ce qui concerne cet estimateur après l'atteinte d'un niveau prédéterminé de l'erreur-type de l'estimateur du niveau d'habileté. L'estimateur final du niveau d'habileté, après l'administration de j items, est égal à l'estimateur provisoire du niveau d'habileté au j^{e} item administré. Le tableau 1 présente sommairement le déroulement des tests.

Tableau 1
Algorithme décrivant le déroulement des tests adaptatifs utilisés dans la simulation (d'après Raîche, 2000, p. 134)

RÈGLE	*ACTION*
1. Règle de départ	Administrer un item dont le niveau de difficulté est égal à $\mu(\theta) = 0{,}00$
2. Règle de suite	Administrer un item dont le niveau de difficulté est égal à la valeur de l'estimateur provisoire du niveau d'habileté
	L'estimateur provisoire du niveau d'habileté est calculé selon la méthode de l'espérance *a posteriori*
3. Règle d'arrêt	Terminer les tests après l'administration d'un nombre prédéterminé d'items variant entre 1 et 60 ou lorsqu'une erreur-type prédéterminée de l'estimateur du niveau d'habileté variant entre 0,20 et 0,85 est obtenue
	L'estimateur final du niveau d'habileté est calculé selon la méthode de l'espérance a posteriori

Résultats et discussion

Règle d'arrêt selon l'erreur-type

La figure 1 et le tableau 2 présentent les valeurs théoriques obtenues en fonction de diverses valeurs retenues pour la règle d'arrêt selon l'erreur-type. On peut observer à la figure 1 a) et au tableau 2 que plus l'erreur-type retenue pour la règle d'arrêt est petite, plus les maximums et minimums de l'estimateur du niveau d'habileté affichent des valeurs qui tendent à couvrir l'intervalle d'intégration utilisé, soit [-4,00 à 4,00]. Toutefois, c'est seulement lorsque l'erreur-type retenue pour la règle d'arrêt est égale ou inférieure à 0,40 que ces maximums et minimums se rapprochent des valeurs utilisées pour l'intervalle d'intégration, sans jamais les atteindre. Cela implique que lorsque le niveau d'habileté affiche des valeurs extrêmes, la méthode d'estimation de l'espérance a posteriori n'est pas appropriée.

L'erreur-type de l'estimateur du niveau d'habileté est toujours inférieure ou égale à l'erreur-type retenue pour la règle d'arrêt. Elle peut toutefois s'éloigner de plus de 0,05 de l'erreur-type retenue pour la règle d'arrêt quand l'erreur-type retenue est supérieure à 0,50. Selon nous, c'est donc uniquement lorsque l'erreur-type retenue est égale ou supérieure à 0,50 que nous pouvons obtenir une précision constante de l'estimateur du niveau d'habileté sur tout le continuum du niveau d'habileté. Cette caractéristique ne présente cependant des avantages que lorsqu'on désire réaliser des analyses statistiques nécessitant l'homogénéité de la variance. C'est le cas, notamment, lorsqu'on désire comparer des moyennes à l'aide de l'analyse de la variance ou d'un test t de Student. Ces tests reposent sur le postulat d'homogénéité de la variance. Toutefois, quand on ne cherche qu'à estimer le niveau d'habileté d'un individu, il est inutile de se préoccuper de l'homogénéité de la variance.

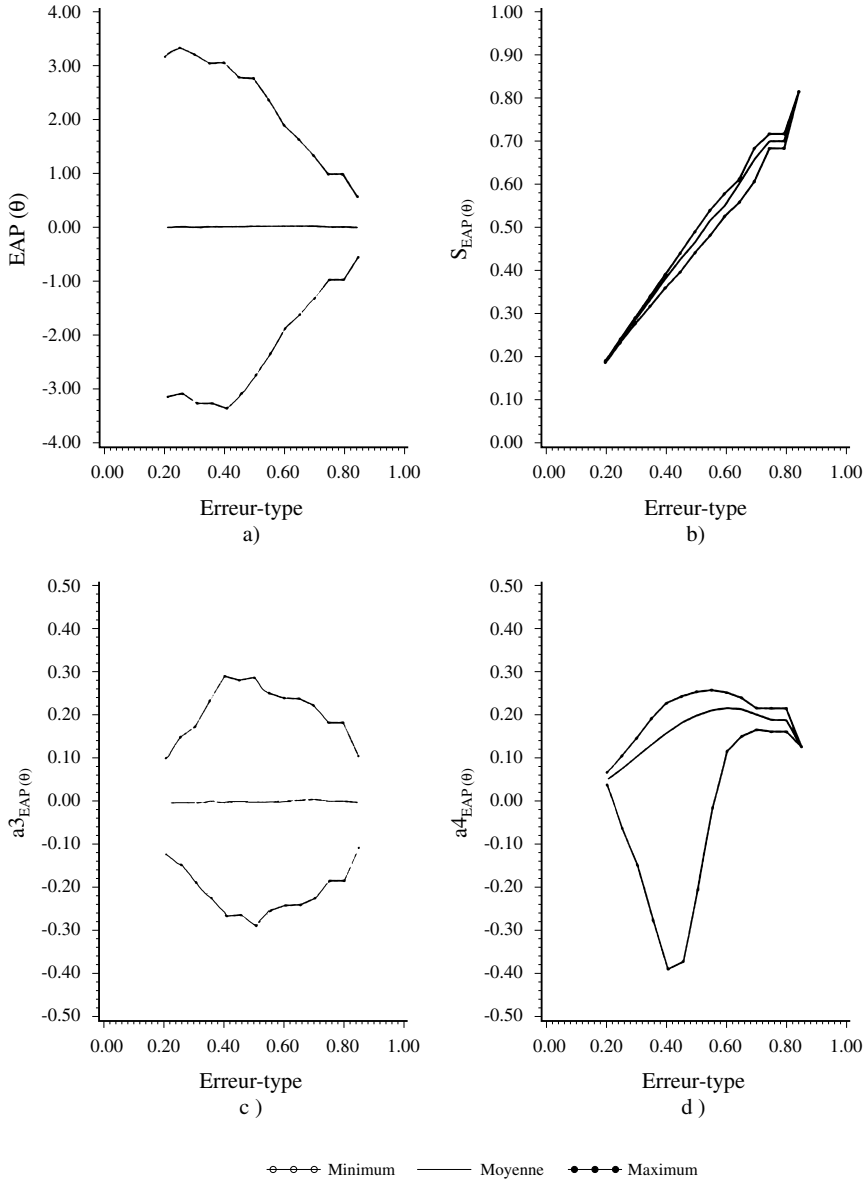

Figure 1. ***Caractéristiques des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque la règle d'arrêt selon l'erreur-type est utilisée (d'après Raîche, 2000)***

Tableau 2
Minimums (↓) et maximums (↑) des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque la règle d'arrêt selon l'erreur-type est utilisée (d'après Raîche, 2000)

S	$EAP(\theta)$		$S_{EAP(\theta)}$		$g1_{EAP(\theta)}$		$g2_{EAP(\theta)}$	
	↓	↑	↓	↑	↓	↑	↓	↑
0,85	-0,56	0,56	0,82	0,82	-0,11	0,11	0,13	0,13
0,80	-0,99	0,99	0,69	0,73	-0,18	0,18	0,16	0,22
0,75	-0,99	0,99	0,69	0,73	-0,18	0,18	0,16	0,22
0,70	-1,33	1,33	0,62	0,69	-0,22	0,22	0,17	0,22
0,65	-1,63	1,63	0,57	0,62	-0,24	0,24	0,15	0,24
0,60	-1,90	1,90	0,54	0,59	-0,24	0,24	0,12	0,25
0,55	-2,36	2,36	0,49	0,55	-0,25	0,25	-0,01	0,26
0,50	-2,76	2,76	0,45	0,50	-0,29	0,29	-0,20	0,26
0,45	-3,10	2,78	0,41	0,45	-0,26	0,28	-0,37	0,24
0,40	-3,38	3,05	0,37	0,40	-0,27	0,29	-0,39	0,16
0,35	-3,28	3,04	0,33	0,35	-0,22	0,23	-0,28	0,16
0,30	-3,28	3,21	0,29	0,30	-0,19	0,17	-0,15	0,17
0,25	-3,10	3,34	0,24	0,25	-0,15	0,15	-0,06	0,15
0,20	-3,16	3,16	0,20	0,20	-0,12	0,10	0,04	0,12

Au tableau 2, nous remarquons aussi que l'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté n'est jamais inférieure à -0,29 et jamais supérieure à 0,29. On peut donc considérer cette distribution comme symétrique puisque ces nombres ne dépassent pas en valeur absolue la valeur critique de 0,40 suggérée par Raîche (2000). Quant à la kurtose de la distribution d'échantillonnage de l'estimateur du niveau d'habileté, elle est n'est jamais inférieure à -0,39 et jamais supérieure à 0,26*;* des valeurs encore de peu d'importance. Nous pouvons donc affirmer que les interprétations relatives à l'erreur-type et à la moyenne de la distribution d'échantillonnage de l'estimateur du niveau d'habileté peuvent être faites, à toutes fins utiles, en fonction d'une distribution de probabilité normale $N(EAP(\theta), S_{EAP(\theta)})$ pourvu que l'erreur-type retenue pour la règle d'arrêt permette à l'estimateur du niveau d'habileté de couvrir raisonnablement l'intervalle d'intégration. C'est le cas lorsque l'erreur-type retenue pour la règle d'arrêt est égale ou inférieure à 0,40.

Règle d'arrêt selon le nombre d'items administrés

À la figure 2 et au tableau 3, les valeurs théoriques des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté obtenues en fonction de diverses valeurs retenues pour la règle d'arrêt selon le nombre d'items administrés sont présentées. Selon la figure 2 a) et le tableau 3, c'est autour de l'administration du dixième item que l'estimateur du niveau d'habileté affiche des minimums et maximums qui tendent à couvrir l'intervalle d'intégration utilisé. Si on considère que lorsque le nombre d'items administrés est égal à neuf, dix et onze l'erreur-type moyenne est respectivement égale à 0,40, 0,38 et 0,36, cette situation est identique à celle que nous avons observée en lien avec la règle d'arrêt selon l'erreur-type. La méthode d'estimation de l'espérance *a posteriori* est alors appropriée seulement quand dix items ou plus sont administrés.

Nous présentons la relation entre le nombre d'items administrés et l'erreur-type de l'estimateur du niveau d'habileté à la figure 2 b) et au tableau 3. La relation, comme on devait s'y attendre est curvilinéaire. Pour satisfaire notre curiosité, nous avons appliqué une adaptation de la formule de prophétie de Spearman-Brown à partir de la valeur obtenue après l'administration du premier item (0,82) : la différence entre la valeur obtenue de l'erreur-type de l'estimateur du niveau d'habileté et la valeur prédite par la formule, quel que soit le nombre n d'items administrés, est toujours positive et ne dépasse jamais 0,04. La formule adaptée est la suivante :

$$S_{EAP_n(\theta)}\ \text{prédite} = \sqrt{1 - \frac{n * (1 - S^2_{EAP_n(\theta)})}{1 + (n-1) * (1 - S^2_{EAP_n(\theta)})}} \qquad [10]$$

L'asymétrie de la distribution d'échantillonnage de l'estimateur du niveau d'habileté présente des valeurs qui varient entre -0,29 et 0,29 tandis que la kurtose est toujours d'au moins -0,41 et ne dépasse jamais 0,39. C'est seulement lorsque douze items sont administrés que la kurtose affiche une valeur extrême de -0,41. Les valeurs obtenues sont peu importantes selon nous et, comme nous l'avons indiqué à la section traitant de la règle d'arrêt selon l'erreur-type, les interprétations relatives à une distribution normale $N(EAP(\theta), S_{EAP(\theta)})$ sont applicables pourvu que l'estimateur du niveau d'habileté couvre raisonnablement l'intervalle d'intégration. Nous avons souligné plus haut que cette condition est assurée par l'administration d'au moins dix items.

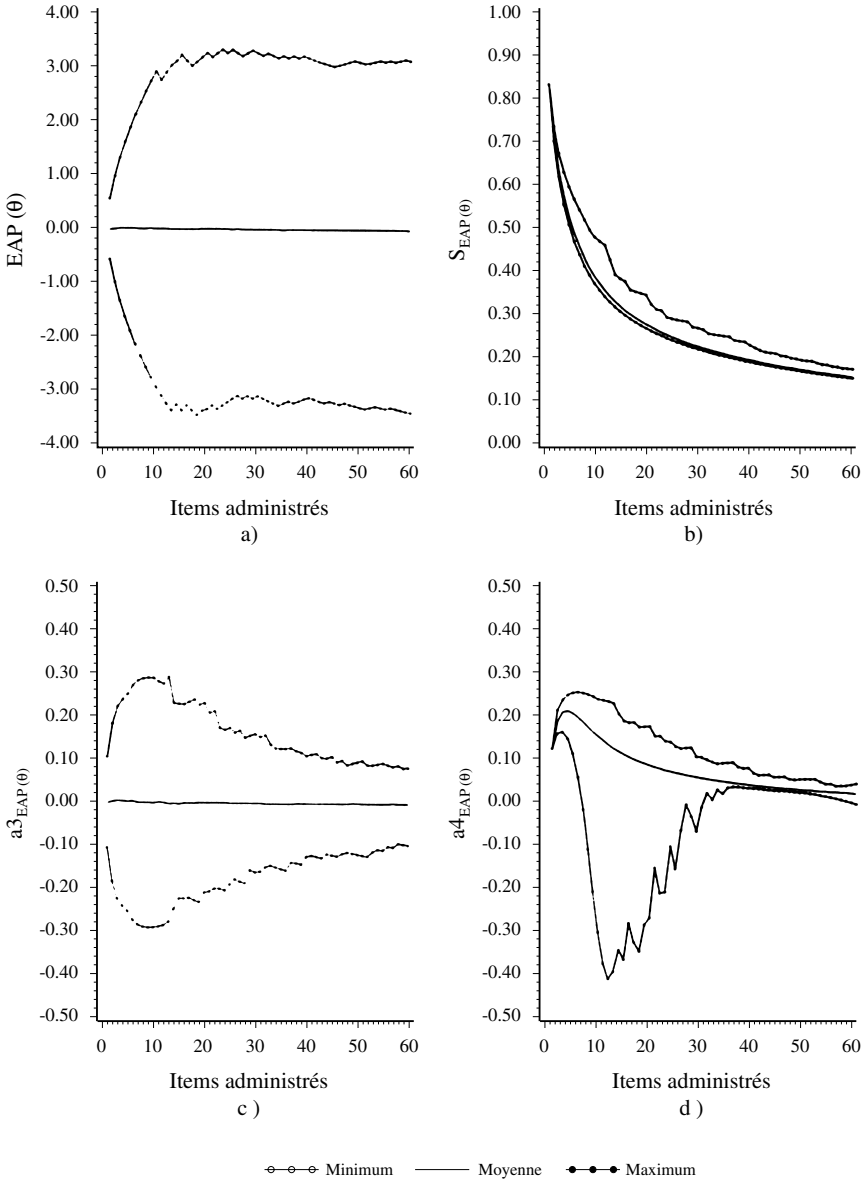

Figure 2. ***Caractéristiques des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est utilisée (d'après Raîche, 2000)***

Tableau 3

Minimums (↓) et maximums (↑) des statistiques associées à la distribution d'échantillonnage de l'estimateur du niveau d'habileté lorsque la règle d'arrêt selon le nombre d'items administrés est utilisée (d'après Raîche, 2000)

Items	$EAP(\theta)$ ↓	↑	$S_{EAP(\theta)}$ ↓	↑	$g1_{EAP(\theta)}$ ↓	↑	$g2_{EAP(\theta)}$ ↓	↑
1	-0,56	0,56	0,82	0,82	-0,11	0,11	0,13	0,13
2	-0,99	0,99	0,69	0,73	-0,18	0,18	0,16	0,22
3	-1,33	1,33	0,61	0,67	-0,22	0,22	0,17	0,24
4	-1,63	1,63	0,55	0,62	-0,24	0,24	0,15	0,25
5	-1,90	1,90	0,50	0,59	-0,25	0,25	0,12	0,26
6	-2,14	2,14	0,46	0,56	-0,27	0,27	0,06	0,26
7	-2,36	2,36	0,43	0,53	-0,28	0,28	-0,01	0,26
8	-2,57	2,57	0,40	0,51	-0,29	0,29	-0,10	0,26
9	-2,76	2,76	0,38	0,49	-0,29	0,29	-0,20	0,25
10	-2,94	2,94	0,36	0,47	-0,29	0,29	-0,30	0,24
11	-3,10	2,78	0,35	0,46	-0,29	0,28	-0,37	0,24
12	-3,25	2,92	0,33	0,45	-0,29	0,28	-0,41	0,24
13	-3,38	3,05	0,32	0,42	-0,28	0,29	-0,39	0,24
14	-3,27	3,13	0,31	0,38	-0,25	0,23	-0,34	0,21
15	-3,38	3,25	0,30	0,38	-0,22	0,23	-0,36	0,39
20	-3,35	3,29	0,26	0,34	-0,21	0,23	-0,26	0,18
25	-3,16	3,35	0,23	0,28	-0,19	0,17	-0,15	0,14
30	-3,10	3,29	0,21	0,26	-0,16	0,16	0,00	0,11
40	-3,13	3,20	0,19	0,23	-0,12	0,11	0,04	0,08
60	-3,40	3,16	0,15	0,17	-0,10	0,08	0,01	0,06

Conclusion

Nous avons pu observer les caractéristiques de différentes statistiques associées à la distribution d'échantillonnage théorique de l'estimateur du niveau d'habileté en *testing* adaptatif. Les résultats obtenus nous suggèrent qu'il est possible d'appliquer les interprétations relatives à une distribution de probabilité normale lorsque l'erreur-type retenue pour la règle d'arrêt est d'au plus 0,40 ou que le nombre d'items administrés retenu pour la règle d'arrêt est d'au moins dix.

Ces résultats sont toutefois valides seulement pour la distribution d'échantillonnage théorique que nous avons étudiée. Ils ont néanmoins l'avantage de nous fixer sur les valeurs que nous pouvons obtenir à partir des estimateurs théoriques des second, troisième et quatrième moments centrés. Ce sont d'ailleurs ces seuls estimateurs théoriques que nous pouvons obtenir lorsque nous procédons au calcul de l'estimateur du niveau d'habileté dans un test adaptatif. Il est clair que ces résultats doivent être interprétés strictement dans leur contexte de réalisation car les valeurs que nous avons obtenues de l'asymétrie et la kurtose n'atteignent pas les valeurs limites qu'on obtient réellement dans les simulations empiriques de distribution normales. Si nous avions étudié la distribution d'échantillonnage empirique de l'asymétrie et de la kurtose, les maximums et minimums auraient été beaucoup plus importants. Pour cette raison, il serait approprié de refaire cette recherche à partir de la distribution d'échantillonnage empirique de l'estimateur du niveau d'habileté à des valeurs fixes du niveau d'habileté.

RÉFÉRENCES

Baker, F.B. (1992). *Item response theory : parameter estimation techniques*. New York : Marcel Dekker.

Bejar, I.I. (1993). A generative approach to psychological and educational measurement. In N. Fredericksen, R.J. Mislevy & I.I. Bejar (éds), *Test theory for a new generation of tests*. Hillsdale : Lawrence Erlbaum Associates.

Bock, R.D., & Mislevy, R.J. (1982). Adaptive EAP estimation of ability in a micro computer environment. *Applied Psychological Measurement, 6*(4), 431-444.

Chang, H.H., & Ying, Z. (1999). α-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222.

Chen, S.K., Hou, L., & Dodd, B.G. (1998). A comparison of maximum likelihood estimation and expected a posteriori estimation in CAT using the partial credit model. *Educational and Psychological Measurement, 58*(4), 569-595.

de Ayala, R.J. (1992). The influence of dimensionality on CAT ability estimation. *Educational and Psychological Measurement, 52*(3), 513-528.

de Ayala, R.J., Schafer, W.D., & Sava-Bolesta, M. (1995). An investigation of the standard errors of expected a posteriori ability estimates. *British Journal of Mathematical and Statistical Psychology, 48*(2), 385-405.

Dodd, B.G., de Ayala, R.J., & Koch, W.R. (1995). Computerized adaptive testing with polytomous items. *Applied psychological measurement, 19*(1), 5-22.

Dodd, B.G., Koch, W.R., & de Ayala, R.J. (1993). Computerized adaptive testing using the partial credit model effects of item pool characteristics and different stopping rules. *Educational and Psychological Measurement, 53*(1), 61-77.

Embretson, S.E. (1999). Generating items during testing : Psychometric issues and models. *Psychometrika, 64*(4), 407-433.

Goldstein, H., & Wood, R. (1989). Five decades of item response modelling. *British Journal of Mathematical and Statistical Psychology, 42*, 139-167.

Hambleton, R.K., & Swaminathan, H. (1987). *Item response theory : principles and applications*. Boston : Kluwer.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of item response theory*. Newbury Park : Sage Publications.

Legendre, R. (1988). *Dictionnaire actuel de l'éducation*. Montréal : Guérin.

Lord, F.M. (1980). Some how and which for practical tailored testing. In L.J.T. van der Kamp, W.F. Langerak & D.N.M. de Gruijter (éds), *Psychometrics for educational debates*. New York : John Wiley and Sons.

Mislevy**,** R.J., & Chang, H.H. (2000). Does adaptive testing violate local independence ? *Psychometrika, 65*(2), 149-156.

Nicewander, W.A., & Thomasson, G.L. (1999). Some reliability estimates for computerized adaptive tests. *Applied Psychological Measurement, 23*(3), 239-247.

Raîche, G. (2000). *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, Université de Montréal.

Spiegel, M.R. (1961). *Theory and problems of statistics*. New York : McGraw-Hill.

Thissen, D., & Mislevy, R.J. (1990). Testing algorithms. In H. Wainer, N.J. Dorans, R. Flaugher, B.F. Green, R.J. Mislevy, L. Steinberg & D. Thissen (éds), *Computerized adaptive testing - A primer*. Hillsdale : Lawrence Erlbaum Associates.

Thissen, D., & Steinberg, L. (1986). A taxonomy of item response models. *Psychometrika, 51*(4), 567-577.

Urry, V.W. (1970). *A Monte Carlo investigation of logistic mental models*. Thèse de doctorat non publiée, Purdue University, West Lafayette.

van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*(4), 327-345.

Weiss, D.J. (1983). *New horizons in testing - Latent trait test theory and computerized adaptive testing.* New York : Academic Press.

Weiss, D.J. (1985). Adaptive testing by computer. *Journal of Consulting and Clinical Psychology, 53*(6), 774-789.

Zwick, R. (1997). The effect of adaptive administration on the variability of the Mantel-Haenszel measure of differential item functioning. *Educational and Psychological Measurement, 57*(3), 412-421.