Mesure et évaluation en éducation

Détection des biais d'items et de personnes en *testing* adaptatif

Richard Bertrand, professeur

Volume 24, Number 2-3, 2001

URI: https://id.erudit.org/iderudit/1091167ar
DOI: https://doi.org/10.7202/1091167ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Bertrand, R. (2001). Détection des biais d'items et de personnes en *testing* adaptatif. *Mesure et évaluation en éducation*, *24*(2-3), 1–22. https://doi.org/10.7202/1091167ar

Article abstract

Computerized adaptative testing (CAT) still have a host of advantages but also some drawbacks. For example, as Wainer (2000, p. xxiii) stated, high stake tests like exams are not good candidates for computerized tests. Also, as the number of items is limited in CAT, the process of detecting biased (DIF) items is vital. This paper suggests a DIF method specially aimed at identifying biased items in a CAT context. A person-fit index for adaptive testing will also be proposed.

Détection des biais d'items et de personnes en *testing* adaptatif

Richard Bertrand, professeur
Université Laval

MOTS-CLÉS : *Testing* adaptatif par ordinateur (TAO), fonctionnement différentiel d'item (FDI), items biaisés, indice de patrons de réponse atypique, théorie des réponses aux items

Si le testing adaptatif par ordinateur (TAO) possède des avantages reconnus depuis plusieurs décennies, il recèle également quelques inconvénients. Par exemple, tel que l'a déjà souligné Wainer (2000, p. xxiii), il ne serait pas très approprié d'utiliser la stratégie du TAO pour les tests à enjeux critiques (high stake tests) comme les examens. De même, puisque le nombre d'items des tests administrés selon la stratégie du TAO est limité, il est impératif de relever et, au besoin, d'éliminer les items comportant un fonctionnement différentiel (FDI). Cet article propose une procédure pour découvrir les items comportant un FDI dans le contexte spécifique du TAO. Une méthode pour distinguer les patrons de réponses atypiques dans le contexte du TAO est aussi suggérée.

KEY WORDS : Computerized adaptive testing (CAT), differential item functionning (DIF), biased items, person-fit index, item response theory

Computerized adaptative testing (CAT) still have a host of advantages but also some drawbacks. For example, as Wainer (2000, p. xxiii) stated, high stake tests like exams are not good candidates for computerized tests. Also, as the number of items is limited in CAT, the process of detecting biased (DIF) items is vital. This paper suggests a DIF method specially aimed at identifying biased items in a CAT context. A person-fit index for adaptive testing will also be proposed.

Introduction

Le *testing* adaptatif par ordinateur comporte certes des avantages qui ont souvent été évoqués par le passé et qui le seront encore longtemps. Déjà en 1980, Fred Lord, un des pilliers de la recherche à «*l'Educational Testing Service*» (ETS), affirmait[1] : «Il est vraisemblable que dans un avenir pas trop éloigné plusieurs tests soient administrés et corrigés par ordinateur.» (Lord, 1980, p. 150.) Lord était même en deça de la vérité puisqu'il ne devait sûrement pas se douter qu'au tournant du millénaire, ETS aurait administré son millionnième test adaptatif par ordinateur (Gitomer, 2000, p. xiii).

Loin de nous inscrire en faux contre ce discours parfois empreint, il faut bien l'admettre, d'un certain prosélytisme, nous reconnaissons que le *testing* adaptatif constitue une modalité de *testing* qui a toutes les chances de s'imposer... un jour dans certains créneaux du *testing*. Car, selon Wainer (2000, p. xxii), il ne serait pas approprié d'administrer tous les tests selon la modalité du *testing* adaptatif par ordinateur (TAO). Cette modalité est encore plus coûteuse que la modalité papier-crayon (PC) et ne convient tout simplement pas à certaines catégories de tests comme ceux comportant des conséquences cruciales pour les sujets examinés, les soit-disants «high stake tests», qui ne sont administrés qu'une fois par année. Au contraire, là où la modalité TAO est la plus appropriée, c'est lorsque le test doit être accessible tout au long de l'année où encore lorsque le test mesure un concept non propice à la modalité PC, par exemple, une simulation de vol d'avion.

Visant l'optimisation de la mesure, le *testing* adaptatif est au cœur même d'une idée remontant à près d'un siècle maintenant et qui a pour objectif essentiellement d'individualiser le *testing*. Quelles sont donc les grandes qualités d'un test adaptatif? Il est (en principe!) plus court qu'un test conventionnel papier-crayon, du moins à précision égale, il comporte des items dont la difficulté tend à s'apparier à l'habileté du sujet testé et il permet d'obtenir un score très rapidement, soit dès la fin du test. Ceci dit, le *testing* adaptatif a aussi les défauts de ses qualités. Par exemple, avec un très petit nombre d'items, comment assurer la validité des interprétations des scores au test, d'autant que ce nombre d'items varie d'un sujet testé à l'autre? Il est, bien sûr, possible d'utiliser un critère d'arrêt de test exigeant un nombre égal d'items par sujet testé. Il est encore possible d'utiliser une méthode de sélection des items qui, tout en répondant à des impératifs psychométriques de maximisation d'information, prenne appui sur une stratification des domaines (ex. algèbre, géométrie) associés au contenu mesuré par le test.

Si ces stratégies liées au critère d'arrêt du test ou au critère de sélection des items sont de nature à améliorer la validité des interprétations des scores émanant d'un test adaptatif, il n'en demeure pas moins que le faible nombre d'items administrés lors d'un TAO rend encore plus vital le processus de détection de biais d'items, biais qui, comme on le sait, limitent également la validité des interprétations. Plusieurs méthodes de détection de biais d'item, appelés FDI pour fonctionnement différentiel d'item lorsqu'il s'agit de considérer spécifiquement la procédure statistique en elle-même, ont vu le jour, plus particulièrement au cours des vingt dernières années. Certaines de ces méthodes, nous le verrons, sont plus adaptées à la détection de biais associés à de tests courts, comme on en rencontre souvent en *testing* adaptatif.

La validité des interprétations des scores au test dépend également de la pertinence du patron de réponses des sujets testés, c'est-à-dire de l'ajustement de ce patron au modèle choisi de réponse aux items. En fait, il faut vérifier que le patron de réponses est minimalement vraisemblable : il serait par exemple très peu vraisemblable d'observer un patron de réponses pour lequel le sujet testé aurait réussi la plupart des items difficiles tout en manquant la plupart des items faciles. Le score d'habileté émanant de ce patron invraisemblable serait peu valable et pourrait être la conséquence d'un comportement atypique de la part du sujet testé, que ce soit la tricherie, la nonchalance extrême ou tout autre comportement.

Après avoir distingué la notion de fonctionnement différentiel d'item (FDI) de celle de biais et présenté les résultats d'une analyse utilisant quelques-unes des méthodes les plus prometteuses de détection de FDI, nous en arriverons à suggérer une méthode plus particulièrement adaptée au contexte du *testing* adaptatif. De même, une fois présenté le concept de patron atypique ainsi que quelques indices, qu'ils soient heuritisques ou fondés sur la théorie des réponses aux items, permettant de détecter de tels patrons, nous en venons à suggérer un indice pertinent au TAO et à faire quelques recommandations et limites.

Biais associés à la nature des items : le fonctionnement différentiel d'item (FDI)

Le contexte de mondialisation des marchés et des échanges qui ne peut que s'étendre à la plupart des activités, incluant les échanges de nature scientifique comme ceux que l'on voit dans le cadre des enquêtes internationales (ex. TIMSS, PISA), nous force à nous occuper au premier chef de la traduction et de l'adaptation culturelle des items d'un test. Or, en traduisant ou en adaptant un test, le sens d'un mot ou d'une phrase peut ne pas être équivalent dans les deux langues ou les deux cultures, risquant ainsi de défavoriser un des groupes testés. En outre, plusieurs items risquent de contenir des stéréotypes qui défavorisent un sous-groupe de la population par rapport à un autre.

Lors de la rédaction des items, il faut donc être très attentif de façon à ne pas changer le sens d'une question traduite ou adaptée d'une autre culture ou encore à ne pas ajouter de stéréotypes. Mais comme il y a loin de la coupe aux lèvres et que les vieux démons (déterminismes sociaux) reviennent souvent à notre insu, il convient de mener une étude empirique afin de limiter au minimum la portée des biais d'items.

Cette étude de détection de biais est encore plus cruciale dans le cas d'un test adaptatif puisque, en principe du moins, un test adaptatif informatisé contiendra moins d'items qu'un test PC de même précision. De plus, la stratégie visant à éliminer *a posteriori* un item biaisé n'aide pas vraiment lors d'un TAO puisque c'est le fait de réussir ou d'échouer un item qui guide le choix et donc l'administration d'un autre item.

Il existe toute une ribambelle de procédures de détection de biais d'item que nous allons maintenant rappeler. Nous allons ensuite proposer une procédure qui s'adapte le mieux au TAO. Au préalable, il convient de distinguer le concept de biais d'item de celui de fonctionnement différentiel d'item (FDI).

Un item sera considéré comme présentant un FDI si deux sujets d'habileté égale mais appartenant à des groupes distincts ont une probabilité différente de réussir l'item. Le FDI est dit uniforme si, comme à la figure 1, la différence de probabilité de réussite favorise le même groupe tout le long de l'échelle d'habileté. Autrement, le FDI est dit non uniforme (figure 2). Lors d'une étude visant à relever des items présentant un FDI, il est d'usage de distinguer les deux groupes comparés en les qualifiant de groupe de référence (ex. anglophones) et de groupe focal (ex. francophones).

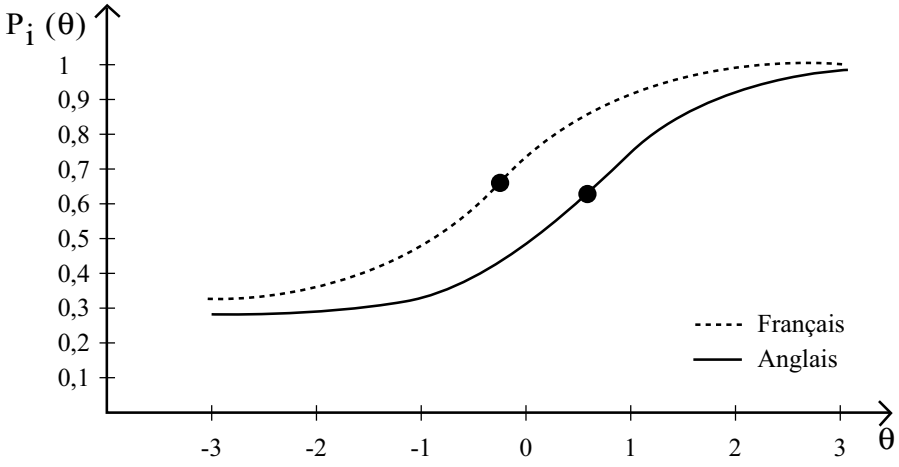

Figure 1. ***Chacune des deux courbes représente le même item. La courbe en pointillé concerne les francophones alors que la courbe en trait plein se rapporte aux anglophones. Le parallèlisme (relatif) entre les deux courbes est signe que le FDI est uniforme.***

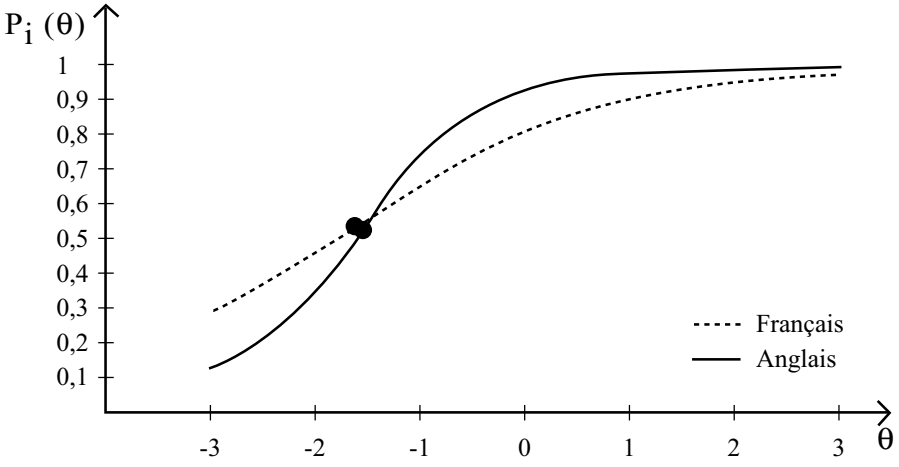

Figure 2. ***Chacune des deux courbes représente le même item. La courbe en pointillé concerne les francophones alors que la courbe en trait plein se rapporte aux anglophones. Le fait que les courbes se croisent est un signe de FDI non uniforme.***

Un *item* i sera dit *biaisé* envers un groupe de sujets (que ce soit le groupe de référence ou le groupe focal) si les deux critères suivants sont respectés :

- deux sujets d'habileté égale mais appartenant à des groupes distincts ont une probabilité différente de réussir l'item i (c'est-à-dire qu'il y a présence d'un FDI) *et*
- la raison de cette différence de probabilité de réussite n'a rien à voir avec l'interprétation usuelle que l'on fait des scores au test (c'est-à-dire qu'il y a absence de validité).

Le deuxième critère mérite une explication supplémentaire. Tout d'abord, notons que seul un comité d'experts pourra décréter qu'un item présentant un FDI est biaisé ou non car il s'agit d'un jugement largement basé sur l'analyse du contenu de l'item. Imaginons que plusieurs des items d'un test de résolution de problèmes comportent des FDI (le premier critère est donc satisfait !), disons en faveur des élèves du groupe de référence, et que l'on se rende compte, après analyse par un comité d'experts, que les items ont été suffisamment bien traduits. Sommes-nous en présence de biais d'item ? En d'autres termes, est-ce que le deuxième critère est satisfait ? Peut-être que oui, peut-être que non. Cela dépend vraiment de la façon dont les scores seront interprétés. Supposons que les items de résolution de problèmes comportent beaucoup plus de mots en français qu'en anglais (situation fréquente lorsqu'on traduit de l'anglais, langue plus synthétique, au français) et que les élèves de 13 ans n'aient pas encore tous acquis une compétence élevée en compréhension en lecture. Si, d'aventure, les scores sont interprétés comme signifiant une habileté générale en mathématique, sans plus, il pourrait être légitime alors de considérer ces items comme biaisés car alors le deuxième critère est bien satisfait : la raison du FDI, de cette différence de probabilité de réussite de cet item, est liée à la compréhension en lecture alors que l'interprétation des scores ne concernent pas cette habileté. Si, par contre, le comité d'experts juge que la compréhension en lecture fait partie des habiletés (secondaires) légitimes visées par (quelques items de) ce test (arguant que quelqu'un qui veut réussir en mathématique, ce qui inclut la résolution de problèmes, doit aussi savoir bien poser des problèmes et donc bien comprendre le texte, etc., etc.), ces items ne seraient pas considérés comme biaisés même si les valeurs observées des FDI sont très élevées : en d'autres termes, le deuxième critère ne serait pas satisfait puisque la compréhension en lecture, habileté secondaire qui explique en partie cette différence de probabilité de réussite, fait partie de l'interprétation usuelle des scores.

Il existe un nombre impressionnant de procédures visant à détecter les FDI, certaines étant fondées sur les modèles de la théorie des réponses aux items (TRI) alors que d'autres ne s'appuient sur aucun modèle. Parmi les procédures ne s'appuyant pas sur les modèles de la TRI, on retrouve la méthode de Mantel-Haenszel et la régression logistique (Clauser & Mazor, 1998). Parmi les procédures s'appuyant sur les modèles de la TRI, on compte la méthode du calcul de l'aire non signée (UPD[2]) de Shepard, Camilli et Williams (1984), la méthode non compensatoire du calcul de l'aire entre les deux CCI (NCDIF[3]) de Raju, van der Linden et Fleer (1995) et la méthode de Thissen, Steinberg et Wainer (1988) basée sur la différence de modèles. Cette liste ne constitue bien sûr qu'un échantillon seulement des procédures permettant de détecter un FDI. Il faut aussi remarquer que, même si elles ne prennent pas spécifiquement appui sur un modèle de la TRI, les méthodes comme celle de Mantel-Haenszel ou encore celle basée sur la régression logistique peuvent aussi utiliser le score TRI comme variable de contrôle plutôt que le score classique.

Bien que toutes ces procédures ne s'accordent pas parfaitement, certaines par contre, comme on le verra bientôt, convergent tout de même vers une décision semblable. Pour le montrer, les tableaux 1 et 2 présentent les résultats que nous avons eus en étudiant le FDI des 110 items du test d'évaluation du Programme des indicateurs du rendement scolaire (PIRS)[4] administré en 1997. Comme le révèle le tableau 1, il existe une grande disparité entre les méthodes. En tout, 20 des 110 items ont été reconnus comme présentant un FDI modéré ou sévère par au moins une des cinq méthodes. Sur ces 20 items, trois seulement ont été reconnus FDI modérés ou sévères par toutes les méthodes : ce sont les items I_102 et I_25 et I_56. L'item I_75 a été relevé comme présentant un FDI modéré par quatre des cinq méthodes. Seule la méthode de l'indice non compensatoire de Raju n'a pu le détecter. Fait à noter : tous les items détectés FDI par la méthode de différence de modèles de Thissen l'ont aussi été par au moins une autre et parfois par plusieurs autres méthodes. Par contre, parmi les items détectés FDI par une seule méthode, c'est la procédure de Mantel-Haenszel (M-H) qui l'emporte avec trois items. Cette observation est d'autant plus percutante que la procédure de Mantel-Haenszel est considérée comme le standard de l'industrie (Roussos, Schnipke & Pashley, 1999) !

Tableau 1

Items reconnus FDI sévères (S) ou modérés (M) par l'une ou l'autre des cinq méthodes, Mantel-Haenszel (M-H), régression logistique (Reg.log), méthode non compensatoire de Raju (NCDIF), méthode de différence de modèles de Thissen (Diff.mod) et méthode de l'aire (UPD) de Shepard, Camilli et Williams

Item	*M-H*	*Reg.log*	*NCDIF*	*Diff.mod*	*UPD*
I_102	S	S	S	S	M
I_25	S	S	S	S	M
I_121	M	-	S	-	-
I_56	M	S	S	S	M
I_100	M	-	-	-	-
I_75	M	M	-	M	M
I_66	M	-	-	-	-
I_113	M	-	-	-	-
I_72	-	-	-	-	M
I_59	-	S	S	M	-
I_76	-	M	M	M	-
I_46	-	M	-	M	-
I_92	-	M	-	-	-
I_60	-	M	-	M	M
I_95	-	M	-	-	-
I_49	-	M	-	-	M
I_91	-	-	S	-	-
I_62	-	M	-	-	M
I_35	-	-	-	-	M
I_63	-	-	-	-	M

Le tableau 2 présente les taux d'entente entre les cinq méthodes prises deux à deux. Chaque valeur du tableau est constituée du rapport entre, d'une part, le nombre d'items reconnus FDI conjointement par les deux méthodes comparées et, d'autre part, le nombre d'items reconnus FDI par l'une *ou* l'autre de ces deux méthodes. Si on se fie au tableau 1, il appert, par exemple, que quatre items ont été relevés comme présentant un FDI conjointement par la méthode de Mantel-Haenszel et la régression logistique. Or, puisque seize items ont été identifiés FDI par l'une ou l'autre de ces deux méthodes, la valeur du taux d'entente entre ces deux méthodes est de 4/16 ou 0,25. On retrouve, au bas du tableau, la moyenne des taux d'entente associés à l'une ou l'autre des méthodes. Ce taux moyen associé à une méthode donnée est une indication globale de sa capacité à détecter les FDI que les autres méthodes ont

aussi détectés. À la lumière des valeurs observées au tableau 2, il appert que c'est la méthode de différence de modèles de Thissen (Diff.mod) qui a montré la plus grande moyenne avec une valeur de 0,4712. C'est donc cette méthode qui s'accorderait le mieux avec les autres méthodes pour découvrir les FDI, du moins si on se limite à ces données. Et c'est la méthode de Mantel-Haenszel (le standard de l'industrie !) qui a présenté la moyenne la plus faible. Autre fait marquant, les deux méthodes qui se sont le moins bien entendues sont celles qui sont basées sur le même principe, à savoir le calcul de l'aire entre les deux courbes : la méthode de Shepard (UPD) et la méthode non compensatoire de Raju (NCDIF). Les deux méthodes qui ont généré le plus haut taux d'entente sont la régression logistique (Reg.log) et la différence de modèles (Diff.mod), deux méthodes dont les fondements ne sont pourtant pas du tout semblables.

Tableau 2
Taux d'entente entre chaque paire de méthodes de détection de biais

	M-H	*Reg.log*	*NCDIF*	*Diff.mod*	*UPD*
M-H	-	0,2500	0,3636	0,3333	0,2857
Reg.log	0,2500	-	0,3571	0,6667	0,4667
NCDIF	0,3636	0,3571	-	0,5	0,2142
Diff.mod	0,3333	0,6667	0,5	-	0,3846
UPD	0,2857	0,4667	0,2142	0,3846	-
Moyenne	0,3082	0,4351	0,3587	0,4712	0,3378

Tel que nous venons de le constater, il existe une quantité impressionnante de méthodes de détection de biais d'item. En outre, ces méthodes ne mènent pas toutes au même ensemble d'items considérés comme FDI. Il est donc pertinent de se questionner sur le choix d'une méthode et encore plus sur le choix d'une méthode dans le contexte d'un test adaptatif. Steinberg, Thissen et Wainer (2000, p. 219) suggèrent de ne pas utiliser la méthode de Mantel-Haenszel : basé sur nos observations, il est inutile de dire que nous endossons cette suggestion. Ces auteurs suggèrent, en lieu et place, d'utiliser la méthode de différence de modèles qui serait plus adaptée au contexte du TAO. Nous sommes pourtant d'avis d'employer une autre méthode beaucoup moins élaborée et qui, comme on vient de le voir, donne des résultats similaires : la méthode basée sur la régression logistique. La méthode de différence de modèles, en effet, est extrêmement laborieuse puisqu'il faut effectuer une analyse TRI pour chacun des items analysés. La méthode basée sur la régression logistique est beaucoup plus simple à utiliser : elle est donc à la portée de plus de chercheurs.

Comme il existe plusieurs façons d'interpréter les résultats d'une étude de FDI fondée sur la régression logistique, nous proposons une interprétation qui est d'ailleurs celle qui, mieux que l'interprétation de Gierl, Rogers et Klinger (1999), donne les résultats les plus conformes aux autres méthodes comparées ici. Rappelons qu'il serait bien téméraire d'interpréter un item comme présentant un FDI seulement si le test du khi-deux émanant de la régression logistique est statistiquement significatif. Jodoin et Gierl (2001) ont montré que la puissance du test du khi-deux est très sensible à la taille de l'échantillon ; or, comme la taille de l'échantillon utilisé pour calibrer une banque d'items utilisée en TAO est généralement très grande, il convient d'être très prudent avant d'interpréter qu'un item comporte un FDI en ne se basant que sur ce test statistique. C'est pourquoi Gierl, Rogers et Klinger (1999) proposent l'interprétation suivante :

- Un item comportera un FDI *sévère* si le test du khi-deux est statistiquement significatif ET SI la différence du R^2 (entre l'étape 1 et l'étape 2 de la régression logistique[5]) est supérieure à ,07.
- Le FDI sera dit *modéré* si le test du khi-deux est statistiquement significatif ET SI la différence de R^2 se situe entre ,035 et ,07.
- Dans tous les autres cas, le FDI sera considéré *négligeable*.

Nous n'avons trouvé qu'une faible ressemblance entre les FDI provenant de cette interprétation et les FDI provenant des interprétations obtenues des autres méthodes. Nous proposons donc l'interprétation suivante.

Nous dirons qu'un item présente un FDI si

- le khi-deux donne un verdict statistiquement significatif ET SI
- OU BIEN la valeur du khi-deux est considérée comme une valeur extrême (une valeur est dite extrême si elle est située à plus de trois fois l'étendue interquartile du troisième quartile) en regard du diagramme en boîte[6] (il s'agit alors d'un FDI *sévère*) ;
- OU BIEN elle est considérée comme une valeur aberrante sans être extrême (une valeur est dite aberrante si elle est située à plus de 1,5 fois l'étendue interquartile du troisième quartile) en regard du diagramme en boîte (il s'agit alors d'un FDI *modéré*).

C'est cette interprétation qui a prévalu pour classer les FDI selon la méthode de la régression logistique présentée au tableau 1. La figure 3 qui suit montre que, conformément au tableau 1, ce sont les items I_56, I_25, I_102 et I_59 qui comportent un FDI sévère suivant notre interprétation. De la même façon, la figure 4 permet de découvrir les items du tableau 1 qui présentent un

FDI modéré suivant la méthode de la régression logistique. Cette classification en FDI sévère et FDI modéré permettra au comité d'experts chargé d'étudier le contenu des items de mettre un ordre de priorité dans l'examen des items.

Il existe bien sûr d'autres façons de détecter des biais, des façons supposément plus spécifiques au TAO (voir les travaux de Zwick, Thayer et Wingersky, 1994, 1995), mais nous proposons d'employer cette méthode basée sur la régression logistique car elle est peu onéreuse, accessible à l'aide d'un logiciel très connu (SPSS : qui permet non seulement d'obtenir les valeurs de chi-deux associées à la régression logistique mais aussi d'obtenir le diagramme en boîte) et permet de détecter autant les FDI uniformes que les FDI non uniformes, ce que ne permettent pas d'autres méthodes, celle de Mantel-Haenszel par exemple.

Notons également que la méthode fondée sur la régression logistique que nous proposons pour découvrir les FDI doit être utilisée sur tous les items de la banque à l'aide des échantillons qui ont servi à la calibration, donc avant que ne commence le TAO en tant que tel.

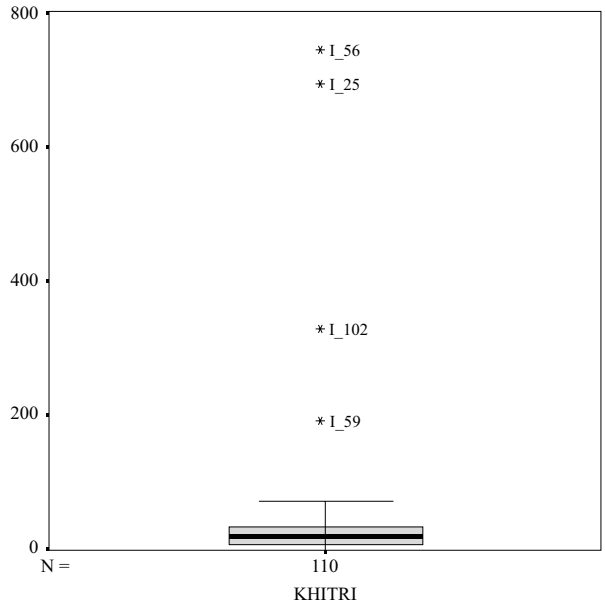

Figure 3 ***Diagramme en boîte indiquant que quatre des 110 items de l'enquête PIRS de 1997, soit les items I_56, I_25, I_102 et I_59, représentés par le symbole (&), sont des valeurs extrêmes et sont donc considérés comme présentant un FDI sévère.***

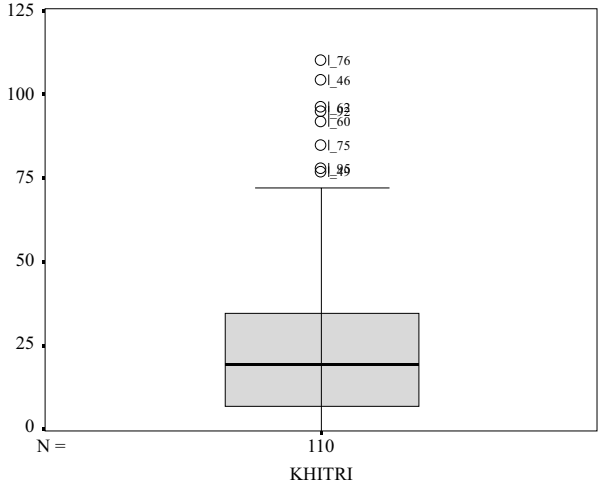

Figure 4 ***Diagramme en boîte indiquant que huit des 110 items de l'enquête PIRS de 1997, soit les items I_76, I_46, I_62, I_92, I_60, I_75, I_95 et I_49 repérés par le symbole (;), sont des valeurs aberrantes (non extrêmes) et sont donc considérés comme présentant comme un FDI modéré.***

Biais associés aux caractéristiques des répondants : patrons de réponses atypiques

Il arrive que des sujets adoptent un comportement atypique, voire bizarre dans leur façon de répondre aux items d'un test. Par exemple, un sujet pourrait réussir la plupart des items difficiles tout en échouant la plupart des items faciles. Il est encore possible de voir un sujet réussir tous les items pairs et ratés les items impairs du test. Or, il est tout à fait possible que le score au test ne révèle pas ces anomalies (van Krimpen-Stoop & Meijer, 2001), causées par des phénomènes comme la tricherie ou une tendance systématique à répondre volontairement au hasard. Puisque c'est le score au test qui servira, en fin de compte, pour classer un individu ou lui donner un diplôme, il ne serait pas approprié de l'utiliser dans les circonstances où la façon de répondre est jugée atypique. La situation est encore plus cruciale en TAO puisque le test est, en général, plutôt court, puisqu'il n'est pas possible de réviser les items et puisque la réponse à un item donné guide le choix du futur item : ainsi, toute séquence de réponses atypiques peut avoir des conséquences déterminantes.

Comment faire pour évaluer le caractère atypique des réponses données par un sujet à un test ? Deux types d'indices du caractère atypique des réponses au test peuvent être distingués : les indices externes et les indices internes. Les indices externes au test sont ceux basés sur des échelles de désirabilité sociale, de « faking bad » ou de « faking good ». Ces indices peuvent parfois avoir un certain intérêt mais ils sont souvent limités, onéreux, puisqu'ils exigent l'administration d'un autre test, et pas toujours valides (Zickar & Drasgow, 1996). C'est pourquoi plusieurs ont proposé des indices internes, c'est-à-dire des indices n'exigeant pas d'autres données que celles apportées par les réponses des sujets au test. La plupart des auteurs, en effet, ont proposé d'utiliser des indices associés directement au patron de réponses au test pour mesurer le caractère atypique du patron (Drasgow, Levine & Williams, 1985 ; Harnisch & Linn, 1982 ; Hulin, Drasgow & Parsons, 1983 ; Levine & Drasgow, 1988 ; Levine & Rubin, 1979 ; McArthur, 1987 ; van Krimpen-Stoop & Meijer, 2001 ; Zickar & Drasgow, 1996). Il s'agit d'examiner les séquences de 1 (bonne réponse) et de 0 (mauvaise réponse) dans le patron de réponses. Une séquence de 1 aux items difficiles et de 0 aux items faciles d'un test devrait normalement attirer l'attention par son caractère atypique !

Le tableau 3 montre que l'on peut caractériser certains patrons de réponses à partir d'une séquence observée de bonnes (1) et de mauvaises (0) réponses. Notons que les items sont classés de gauche à droite du plus facile (I_1) au plus difficile (I_8). Le patron du « parfait-fort » est aussi qualifié de « too good to be true » (Wright & Stone, 1979). Il devrait attirer l'attention tout autant que le « parfait-faible ». Le patron « normal » présente une séquence auquelle on serait en droit de s'attendre, alors que le patron du « chanceux » pourrait bien aussi être celui d'un tricheur. Un bon indice interne devrait permettre de distinguer, par exemple, le patron du répondant chanceux ou aléatoire de celui du répondant normal. Dit autrement, un bon indice interne devrait accorder une valeur différente aux patrons vraisemblables (comme le patron normal ou le patron consciencieux-lent) et aux patrons carrément atypiques, celui du chanceux par exemple. Ceci dit, toutes les formes de patrons atypiques ne sont pas aussi facilement détectables. Par exemple, Meijer, Molenaar et Sijtsma (1994) ont montré que les sujets qui copiaient les réponses aux items difficiles sur leurs voisins plus habiles étaient beaucoup plus faciles à repérer que des sujets répondant de façon nonchalante ou complètement au hasard. Il va sans dire que les qualificatifs associés aux patrons de ce tableau, à la limite du caricatural, ont été donnés plus pour frapper l'esprit que pour réellement caractériser les sujets.

Tableau 3
Exemples de qualificatifs des patrons de réponses : les items sont classés, de gauche à droite, du plus facile au plus difficile

Sujet	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8
parfait-fort	1	1	1	1	1	1	1	1
normal	1	1	1	1	1	0	1	0
endormi	1	1	1	0	0	1	0	1
aléatoire	1	0	1	0	1	0	1	0
consciencieux-lent	1	1	1	0	0	0	0	0
chanceux	1	0	0	0	0	1	1	1
parfait-faible	0	0	0	0	0	0	0	0

Deux types d'indices internes ont été proposés pour quantifier le caractère atypique d'un patron de réponses : ceux fondés sur les modèles de la TRI et les indices dits heuristiques ou empiriques qui ne font appel à aucun modèle spécifique.

Parmi les indices heuristiques, l'indice de précaution (« caution index ») présentée sous une forme modifiée par Harnisch et Linn (1982) mais attribuée d'abord à Sato et Kurata (1977) est sûrement l'un des plus utiles (Thibault, 1992). Ce dernier indice est basé sur le principe suivant : si tous les items du test sont classés du plus facile au plus difficile, un patron de Guttman est celui qui, comme 1111100000, implique que la réussite à un item donné engendre la réussite à tout item plus facile et l'échec à un item donné engendre l'échec à tout item plus difficile. L'indice de Harnisch et Linn donne une mesure standardisée, entre 0 et 1, de l'écart entre un patron donné et un patron de Guttman.

Toute une série d'indices faisant appel de près ou de loin aux modèles de la TRI ont été proposés parmi lesquels l'indice L_0, le logarithme du maximum de la fonction de vraisemblance et sa version standardisée, l'indice L_z, attribués à Levine et Rubin (1979), les indices polytomiques P_0 et P_z, de Drasgow, Levine et Williams (1985), ou enfin l'indice multitest L_{zm} de Drasgow, Levine et McLaughlin (1991).

L'indice L_0, le logarithme du maximum de la fonction de vraisemblance est particulièrement instructif puisqu'il est basé sur le principe suivant : l'estimé d'habileté en TRI, l'estimé du thêta, est la valeur sur l'échelle des thêtas qui correspond au maximum de la fonction de vraisemblance, c'est la valeur qui maximise les chances d'observer le patron de réponses. Or, il est possible (Hulin, Drasgow & Parsons, 1983) que le maximum de la fonction de

vraisemblance soit tellement faible qu'il rende très peu vraisemblable l'observation d'un patron de réponses : la valeur de thêta, l'estimé, est donc peu représentative de l'habileté du sujet. C'est bien ce qui arrive lorsqu'un patron de réponses est atypique, un peu comme le patron du chanceux (tableau 3).

Puisque les indices heuristiques sont moins puissants (Levine & Drasgow, 1988) et qu'ils exigent, en général, de connaître les réponses au test de tous les sujets d'un groupe, ils sont donc beaucoup moins appropriés au *testing* adaptatif. Nous nous devons donc de concenter notre attention plus spécifiquement sur les indices dont les fondements originent des modèles de la théorie des réponses aux items et plus particulièrement sur l'indice L_z, dont les valeurs sont facilement accessibles lors d'un TAO.

Les paragraphes qui suivent donnent un exemple simple du calcul de l'indice L_z. Afin de faciliter l'interprétation de cet indice, on peut supposer, à l'instar de van Krimpen-Stoop et Meijer (1999), que ses valeurs sont distribuées suivant la loi normale centrée (à 0) et réduite (à 1). L'indice L_0, tel que précisé avant, est constitué du logarithme de la fonction de vraisemblance associée à son patron de réponses où $\hat{\theta}$ est l'estimé de l'habileté thêta. Donc,

$$L_0=\sum_{i=1}^{n}\left\{u_i Ln[P_i(\hat{\theta})]+(1-u_i)Ln[1-P_i(\hat{\theta})]\right\}$$

L'indice L_z, par ailleurs est obtenu en standardisant les valeurs de l'indice L_0. Ainsi,

$$L_z=\frac{L_0-E(L_0)}{\sqrt{Var(L_0)}}$$

où

$$E(L_0)=\sum_{i=1}^{n}\left\{P_i(\hat{\theta})Ln[P_i(\hat{\theta})]+(1-P_i(\hat{\theta}))Ln[1-P_i(\hat{\theta})]\right\}$$

et

$$Var(L_0)=\sum_{i=1}^{n}\left\{P_i(\hat{\theta})[1-P_i(\hat{\theta})]Ln[\frac{P_i(\hat{\theta})}{1-P_i(\hat{\theta})}]^2\right\}$$

Précisons que u_i est la fonction bien connue qui vaut 1 si l'item i est réussi et 0 si l'item i est manqué. De plus, $P_i(\hat{\theta})$ est la probabilité de réussite de l'item i pour un sujet d'habileté (estimée) $\hat{\theta}$.

$E(L_0)$ et $Var(L_0)$ sont appelées respectivement valeur attendue et variance de la fonction L_0.

Insistons encore sur le fait que, lors d'un TAO, toutes ces valeurs sont déjà disponibles : il ne faut donc pas de calculs supplémentaires pour les produire.

Comme nous le verrons, la valeur de l'indice L_0 sera d'autant plus faible que les patrons seront peu vraisemblables : par exemple, elle sera plus particulièrement faible pour un sujet qui a réussi des items difficiles et échoué des items faciles. C'est sur la base des faibles valeurs de cet indice que les patrons seront classés atypiques puisque invraisemblables. Or comme les valeurs de L_0 dépendent de la valeur de thêta et donc du niveau d'habileté du sujet, il est plus approprié d'interpréter les valeurs de L_z. Ainsi, il sera possible de déterminer un seuil (ex. $| L_z | > 2$) à partir duquel le patron est considéré atypique, peu importe le niveau d'habileté du sujet.

Voici maintenant, présenté au tableau 4, un exemple de quatre sujets de *même habileté* à qui on a administré un test de cinq items mais dont le patron de réponses diffère d'un sujet à l'autre. Cet exemple est tiré de l'idée de Hulin, Drasgow et Parsons (1983, p. 123) que nous avons enrichie en ajoutant d'autres patrons de réponses et en calculant les valeurs de l'indice de Sato. Remarquons également que les items sont classés de gauche à droite par ordre de difficulté ($P_i(\hat{\theta})$) du plus facile, líitem 1, au plus difficile, líitem 5. Comme on peut le voir, chacun des sujets a rÈussi trois items ($X_j = 3$) mais il ne semble pas, d'après les valeurs des indices de Sato, de L_0 et de L_z, que le degré de vraisemblance des patrons de réponses soit le même pour les quatre sujets. En effet, le sujet 3 a échoué l'item le plus facile, l'item 1, mais réussi l'item le plus difficile, l'item 5. Puisque son patron de réponses s'éloigne d'un patron de Guttman, la valeur de l'indice de Sato est élevé, soit ,67. Par contre, le sujet 1 a réussi les items les plus faciles et échoué les items les plus difficiles : son patron de réponses est donc identique à un patron de Guttman, d'où une valeur de 0 à l'indice de Sato.

En remarquant que $u_i = 1$ si l'item i est réussi et $u_i = 0$ si l'item i est échoué et en utilisant les données du tableau 4, il est facile, à l'aide des formules données précédemment, de calculer les valeurs de l'indice L_0 puis de l'indice L_z. Comme attendu, c'est le patron du sujet 3 qui génère la valeur de

L_z la plus élevée, soit –3,18 : c'est donc ce patron qui est considéré comme le plus atypique, si, bien sûr, on se fie à l'indice L_z. De plus, si on suppose que les valeurs de l'indice L_z sont distribuées selon la loi normale standardisée, une valeur telle que –3,18 sera considérée significative puisqu'en valeur absolue elle est plus élevée que 2. Lors d'un TAO, un sujet qui obtiendrait une telle valeur pourrait se faire demander de reprendre le test, son patron de réponses étant trop peu vraisemblable : il faut expliquer pourquoi il a réussi l'item le plus difficile et échoué l'item le plus facile. Convenons, bien sûr, qu'un test de cinq items n'est pas suffisamment long pour prendre une décision sur la capacité d'un individu. Il faudrait probablement plusieurs dizaines d'items pour qu'une telle décision soit reconnue comme valable. Un patron comme 00000011111101011111, par exemple, serait sûrement de nature à susciter certaines interrogations, si, comme toujours, les items faciles se situent à gauche et les items difficiles à droite du patron.

Tableau 4

Patrons de réponses de quatre sujets à un test de cinq items auxquels nous avons associées les valeurs de l'indice de Sato et des indices L_0 et Lz

	Item 1	*Item 2*	*Item 3*	*Item 4*	*Item 5*	X_j	*Sato*[8]	L_0	L_z
Sujet 1	1	1	1	0	0	3	0	-1,62	0,88
Sujet 2	1	1	0	1	0	3	,33	-2,46	0,10
Sujet 3	0	1	1	0	1	3	,67	-6,01	-3,18
Sujet 4	1	0	1	1	0	3	,33	-3,31	-0,68
n_i	3	3	3	2	1				
$P_i(\hat{\theta})$	,9	,7	,5	,3	,1			$E(L_0)$	=-2,57
								Var (L_0)	= 1,17

Puisque le TAO doit nécessairement se faire à l'aide d'un logiciel programmé pour estimer les valeurs des thêtas, ce logiciel peut aussi calculer assez facilement les valeurs d'un indice comme L_z : de sorte que, au terme du TAO, en présentant son estimé d'habileté thêta au sujet à qui est administré un TAO, on pourrait également lui donner la valeur de l'indice L_z qui quantifie le côté atypique du patron.

Avant d'interpréter l'indice L_z, il faut cependant tenir compte de la remarque de Schmitt, Cortina et Whitney (1993), qui spécifie que les indices basés sur un modèle de la TRI sont d'autant plus efficaces que le test est long : or, le *testing* adaptatif vise justement à réduire la longueur du test, à précision

égale. En outre, avant de prendre une décision (absolue) sur le caractère atypique d'un patron, il faut, comme on l'a vu, supposer que les valeurs de L_z sont distribuées selon la loi normale, une condition qui ne semble pas toujours remplie (van Krimpen-Stoop & Meijer, 1999).

Conclusion

Après avoir revu les principales méthodes visant à détecter les biais d'item et celles permettant de relever les patrons de réponses atypiques, nous avons présenté une procédure visant à améliorer la validité d'interprétation des scores émanant d'un TAO en suggérant une méthode de détection de biais d'item, la régression logistique et un indice d'identification de patron atypique, le L_z.

Nous suggérons d'utiliser la régression logistique *avant* le TAO, en examinant tous les items de la banque d'où seront puisés les items qui constitueront le TAO. En effet, il semble peu souhaitable d'éliminer *a posteriori* un item comportant un FDI, comme cela peut se faire lors d'une situation de *testing* habituelle, car en TAO la réussite ou l'échec à un item oriente le choix des items subséquents. En outre, il serait plus logique que la variable de contrôle soit le thêta des sujets qui ont servi à la calibration des items. Les sujets à qui sera administré le TAO, en effet, devraient faire partie des mêmes sous-groupes de la population qui ont servi à la calibration. Chaque sous-groupe minoritaire visé par le test sera considéré comme le groupe focal, le groupe de référence étant constitué en général du sous-groupe majoritaire. La méthode de détection de FDI fondée sur la régression logistique comporte plusieurs avantages : elle est facile à utiliser (à l'aide de SPSS !), elle permet de détecter les biais uniformes et les biais non uniformes, elle permet de trouver le sens du FDI (qui détermine qui, du groupe de référence ou du groupe focal, est favorisé), elle fournit une classification (sévère, modéré, négligeable) qui permet au comité d'experts d'établir un ordre de priorité dans l'examen des FDI et elle possède un des meilleurs taux d'entente moyens avec les autres méthodes (comme illustré au tableau 2).

L'indice d'identification du patron de réponses atypique, le L_z, sera disponible *après* le TAO, soit au moment où le sujet recevra son score d'habileté thêta. Cet indice est facile à obtenir en TAO puisque toutes les valeurs qui le composent doivent de toutes façons être calculées dans la procédure du TAO. De plus, c'est un indice particulièrement intéressant du fait qu'il permet de prendre une décision (absolue !) sur l'invraisemblance du patron (ex. $| L_z | > 2$).

Notons cependant un certain nombre de limites inhérentes à l'utilisation de la régression logistique. Par exemple, plusieurs critères ont été proposés pour interpréter le degré du FDI (différence de R^2, diagramme en boîte des valeurs du khi-deux, niveau de signification du khi-deux). De plus, plusieurs variables de contrôle (score classique, score thêta) peuvent être envisagées à la première étape de la régression logistique. Or le choix du ou des critères d'interprétation et de la variable de contrôle est crucial : en général, ce choix déterminera une série unique d'items comportant un FDI. Un autre choix mène généralement à une autre série d'items comportant un FDI.

Un certain nombre de limites associées à l'utilisation de l'indice L_z doivent aussi être indiquées. Mentionnons tout d'abord le fait que cet indice est d'autant plus efficace que le nombre d'items est élevé : ce qui semble ne pas concorder avec l'idée d'un TAO qui tend à minimiser le nombre d'items (Schmitt, Cortina & Whitney, 1993). En outre, comme nous l'avons déjà souligné, les valeurs de l'indice L_z peuvent s'éloigner de la loi normale (van Krimpen-Stoop & Meijer, 1999).

Bien qu'il existe des limites à l'utilisation de la régression logistique et de l'indice L_z, la mise sur pied d'un TAO est si dispendieuse qu'il serait négligeant de ne pas utiliser des indices de biais associés aux items ou aux personnes, de façon à obtenir une interprétation du score TAO qui soit non seulement fidèle (ça le TAO le contrôle bien !) mais aussi valide.

NOTES

1. Traduction libre de « It seems likely that in the not too distinct future mental tests will be administered and scored by computer. »
2. UPD pour « unsigned probability difference ».
3. NCDIF pour « non compensatory differential item functionning ».
4. Il s'agit d'un programme d'enquêtes pancanadiennes géré par le Conseil des ministres de l'Éducation du Canada.
5. L'étape 1 de la régression logistique ne considère que la variable de contrôle (score total ou thêta) comme variable indépendante ; c'est le groupe (focal ou de référence) qui est inséré dans l'équation de régression à l'étape 2 comme variable indépendante.
6. Au sens de Bertrand et Valiquette (1986, p. 44).
7. Même si ce n'est pas possible de le voir clairement à partir de la figure elle-même, ces huit valeurs aberrantes ont été relevées en utilisant la commande « Explore » de SPSS associée à la production du diagramme en boîte et moustaches.
8. Tel qu'indiqué dans Bertrand et Blais (à paraître), l'indice de Sato est donné par

$$S_j = \frac{\sum_{i=1}^{X_j}(1-u_{ij})n_i - \sum_{i=X_j+1}^{I} u_{ij}n_i}{\sum_{i=1}^{X_j} n_i - \sum_{i=I+1-X_j}^{I} n_i}$$

RÉFÉRENCES

Bertrand, R., & Blais, J.-G. (à paraître). *Modèles de mesure : l'apport de la théorie des réponses aux items*. Québec : Presses de l'Université du Québec.

Bertrand, R., & Valiquette, C. (1986). *Pratique de l'analyse statistique des données*. Québec : Presses de l'Université du Québec.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA : Sage Publications.

Camilli, G., & Congdom, P. (1999). Application of a method of estimating DIF for polytomous test items. *Journal of Educational and Behavioral Statistics, 24*(4)*,* 323-341.

Clauser, B.E., & Mazor, K.M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement : Issues and Practices*, Spring, 31-44.

Drasgow F., Levine, M.V., & Williams, E.A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. British Journal of Mathematical and Statistical Psychology, 38(1), 67-86.

Drasgow F., Levine, M.V., & McLaughlin, M.E. (1991). Appropriateness measurement for some multidimensional test batteries. Applied Psychological Measurement, 15(2), 171-191.

Gierl, M.J., Rogers, W.T., & Klinger, D.A. (1999). Using statistical judgement reviews to identify and interpret translation DIF. Paper presented at the Annual Meeting of the National Council for Measurement in Education, Montréal, Canada.

Gitomer, D.H. (2000). Foreword to the second edition. In H. Wainer (éd.), *Computerized adaptive testing : A primer*. Hillsdale, NJ : Lawrence Erlbaum.

Harnisch, D.L., & Linn, R.L. (1982). *Identification of aberrant response patterns*. Research Report, Champaign, IL : University of Illinois.

Holland, P. W., & Thayer, D.T. (1986). *Differential item functionning and the Mantel-Haenszel procedure*. Technical Report, Princeton, NJ : Educational *Testing* Service.

Holland, P.W., & Wainer, H. (éds) (1993). *Differential item functionning*. Hillsdale, NJ : Lawrence Erlbaum.

Hulin, C.L., Drasgow, F., & Parsons, C.K. (1983). *Item response theory : applications to psychological measurement*. Homewood, IL : Dow-Jones Irwin.

Jodoin, M.G., & Gierl, M.J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education* (à paraître).

Levine, M. V., & Drasgow, F. (1988). Optimal appropriateness measurement. *Psychometrika*, *53*(2), 161-176.

Levine, M. V., & Rubin, D.B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*(3)*,* 269-290.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ : Lawrence Erlbaum.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719-748.

McArthur, D.L. (1987). Analysis of patterns : the S-P technique. In D.L. McArthur (éd.), *Alternative approaches to the assessment of achievement*. Boston, Mass : Kluwer Academic.

Meijer, R.R, Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement*, *18*(2), 111-120.

Meijer, R.R., & Nering, M.L. (1999). Computerized adaptive testing : overview and introduction. *Applied Psychological Measurement, 23*, 187-194.

Oshima, T.C., Raju, N.S., & Flowers, C.P. (1997). Development and demonstration of multidimensional IRT-based internal measures of differential functionning of items and tests. *Journal of Educational Measurement, 34*(3), 253-272.

Raju, N.S. (1990). Determining the significance of the estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197-207.

Raju, N.S., van der Linden, W.J., & Fleer, P.F. (1995). IRT-Based internal measures of differential functionning of items and tests. *Applied Psychological Measurement, 19*(4), 353-368.

Roussos, L.A., Schnipke, D.L., & Pashley, P.J. (1999). A generalized formula for the Mantel-Haenszel differential item functionning parameter. *Journal of Educational and Behavioral Statistics, 24*(3), 292-322.

Sands, W.A., Waters, B.K., & McBride, J.R. (éds) (1997). *Computerized adaptive testing : from inquiry to operation*. Washington, DC : APA.

Sato, T., & Kurata, M. (1977). Basic S-P score table characteristics. *NEC Research and Development*, *47*, 64-71.

Schmitt, N., Cortina, J.M., & Whitney, D.J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement*, *17*(2), 143-150.

Shepard. L. A. (1982). Definitions of bias. In R.A. Berk (éd.), *Handbook of methods for detecting test bias*. Baltimore : The Johns Hopkins University Press.

Shepard. L. A., Camilli, G. & Williams, D.M. (1984). Accouting for statistical artifacts in item bias research. *Journal of Educational Statistics, 9*(1), 93-128.

Steinberg, L., Thissen, D., & Wainer, H. (2000). Validity. In H. Wainer (éd.), *Computerized adaptive testing : a primer*. Hillsdale, NJ : Lawrence Erlbaum.

Thibault, J. (1992). *L'apport de fidélité intra-individuelle de trois modes de conception distincts estimés selon le modèle logistique à trois paramètres et selon le modèle polytomique de Bock-Samejima utilisés en TRI*. Thèse de doctorat, Université Laval, Sainte-Foy.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (éds), *Test validity*. Hillsdale, NJ : Lawrence Erlbaum.

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*, 121-140.

van Krimpen-Stoop, E.M.L.A., & Meijer, R.R. (2001). CUSUM-Based person-fit statistics for adaptive *testing*. *Journal of Educational and Behavioral Statistics, 26*(2), 199-218.

Wainer, H. (2000). *Computerized adaptive testing : a primer*. Hillsdale, NJ : Lawrence Erlbaum.

Wright, B.D., & Stone, M.H. (1979). *Best test design*. Chicago : MESA.

Zickar, M.J. & Drasgow F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement*, *20*(1), 71-87.

Zwick, R., Thayer, D.T., & Wingersky, M. (1994). A simulation study of methods for assessing differential item functionning in computerized adaptive tests. *Applied Psychological Measurement, 18*, 121-140.

Zwick, R., Thayer, D.T., & Wingersky, M. (1995). Effect of Rasch calibration on ability and DIF estimation in computer-adaptive tests. *Journal of Educational Measurement, 32*, 341-363.