Revue des sciences de l'éducation

Évaluation d'un test de lecture en anglais par deux méthodes de détection du fonctionnement différentiel d'items
Evaluation of an English reading test based on two methods for detecting differential item functioning
Evaluación de una prueba de lectura en inglés mediante dos métodos de detección del funcionamiento diferencial de los ítems

François Pichette, Gilles Raîche, Sébastien Béland and David Magis

Volume 37, Number 3, 2011

URI: https://id.erudit.org/iderudit/1014757ar
DOI: https://doi.org/10.7202/1014757ar

See table of contents

Publisher(s)

Revue des sciences de l'éducation

ISSN

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

Explore this journal

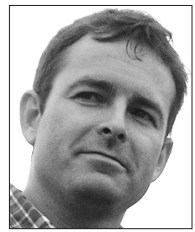
Cite this article

Pichette, F., Raîche, G., Béland, S. & Magis, D. (2011). Évaluation d'un test de lecture en anglais par deux méthodes de détection du fonctionnement différentiel d'items. *Revue des sciences de l'éducation*, *37*(3), 543–568. https://doi.org/10.7202/1014757ar

Article abstract

This study examines the presence of gender-based differential item functioning on a reading comprehension test in English administered to 171 French-speaking university students. Two non-parametric methods are used : the Mantel-Haenszel test and the logistic regression model. On a total of 64 items, two items are identified by the Mantel-Haenszel test, whereas five additional items are identified by logistic regression. This low number of items suggests reasonable fairness for the test, but the discrepancies observed stress the need for further analyses to clarify the status of those items.

Évaluation d'un test de lecture en anglais par deux méthodes de détection du fonctionnement différentiel d'items *



François Pichette
professeur
Télé-Université /
Université du Québec à Montréal

Gilles Raîche
professeur
Université du Québec à Montréal

Sébastien Béland
doctorant
Université du Québec à Montréal

David Magis
chargé de recherches FNRS
Université de Liège

RÉSUMÉ • Cette étude vise à examiner la présence de fonctionnement différentiel d'items selon le sexe des répondants dans un test de compréhension en lecture en anglais administré à 171 universitaires francophones. Deux méthodes non paramétriques sont utilisées : le test Mantel-Haenszel et le modèle de régression logistique. Sur un total de 64 items, deux présentent un fonctionnement différentiel selon le test Mantel-Haenszel, alors que cinq items supplémentaires ressortent par la régression logistique. Ce faible nombre d'items suggère une bonne équité du test, mais les différences observées soulignent la nécessité d'analyses additionnelles pour clarifier le statut de ces items.

MOTS CLÉS • fonctionnement différentiel d'items, test de vérification de phrases, biais lié au sexe, test Mantel-Haenszel, modèle de régression logistique.

* La présente recherche a été financée par le Fonds institutionnel de la Télé-Université (FIR, n°P710571) ainsi que par le programme de subvention ordinaire du Conseil de recherches en sciences humaines du Canada (CRSH, n° 410-2008-1856).

1. Introduction

1.1 Différences hommes-femmes liées au texte lu

De nombreuses études ont porté, dans les années 1980, sur l'impact de la connaissance du contenu, sujet ou thème d'un texte sur la compréhension de celui-ci (Bramski et Williams, 1984 ; Carrell 1984 ; Koh, 1985 ; Stevens, 1980). Ces études ont été menées surtout autour de la théorie du schéma, hypothèse constructiviste selon laquelle le savoir est organisé en réseaux (ou schémas) de connaissances et que de nouvelles connaissances ne peuvent s'acquérir qu'en s'arrimant à celles déjà possédées par l'apprenant, qui servent à remplir les espaces vides et à faire des inférences (Anderson, 1984 ; Carrell, 1987). Ces études ont conduit à l'idée que, puisque les femmes et les hommes afficheraient des domaines d'intérêt différents, ces différences se répercuteraient dans leur compréhension en lecture. Ainsi, on a fait ressortir des différences de performance en lecture entre hommes et femmes, imputables au sujet des textes lus, où les hommes surclassent les femmes sur des sujets tels les sports ou les sciences, alors que les femmes les surpassent quant aux textes sur la littérature, les langues ou les sciences humaines (Bügel et Buunk, 1996 ; Chavez, 2001 ; Doolittle et Welch, 1989 ; Hyde et Linn, 1988). Ce phénomène est évidemment lié aux connaissances préalables du contenu du texte qui agit comme variable médiatrice, puisqu'il existe évidemment un fort lien entre ce que nous aimons et ce que nous connaissons, ce qui nous porte davantage à nous documenter et à nous informer sur des sujets qui nous intéressent (Carrell et Wise, 1998). Ainsi, quand on prend soin de choisir des sujets neutres, qu'on ne peut associer au masculin ou au féminin, ou alors quand on mesure le degré de familiarité avec le contenu au lieu de le supposer à partir du sexe du participant, on remarque le même effet facilitateur (Brantmeier, 2003 ; Al-Shumaimeri, 2005).

1.2 Différences hommes-femmes liées à la motivation

Ainsi, si l'on contrôle le facteur de la familiarité avec le contenu, par exemple en choisissant des textes neutres, les comparaisons entre les hommes et les femmes donnent un portrait différent. En effet, une étude de grande ampleur menée par Chiu et McBride-Chang (2006) auprès de quelque 200 000 adolescents de 15 ans dans 43 pays différents montre une supériorité significative des filles en lecture dans tous les pays, et cette différence serait imputable à la motivation et au plaisir que les filles éprouvent quand elles lisent, qui ressort comme la seule variable d'impact et qui explique 42 % de l'effet lié au sexe. Déjà en 1996, Bügel et Buunk évoquaient les habitudes de lecture comme variable médiatrice probable. Pour leur part, Lynn et Mikk (2009), analysent les scores obtenus à plusieurs tests internationaux récents, et arrivent à une conclusion similaire : une performance supérieure des filles qui s'expliquerait par une plus grande inclination pour la lecture. Malgré quelques indications contraires quant à l'idée d'une plus grande compétence en lecture chez les filles (par exemple, White, 2007), la plupart des études confirment ce résultat, et suggèrent que cette supériorité apparaît tôt dans

l'apprentissage de la lecture chez les enfants (Lynn et Graham, 1993 ; Wolf et Gow, 1986). Hormis ce facteur de la motivation envers la lecture, peu de différences sont observées : aux tout débuts de l'acquisition de la lecture, certains chercheurs notent des différences d'habileté en lecture entre jeunes garçons et jeunes filles (Wolf et Gow, 1986), d'autres non (Harper et Pelletier, 2008). Quoi qu'il en soit, d'éventuelles différences liées à l'émergence de la littératie semblent rapidement disparaître, dès l'âge de 9 ans (Rosen, 2001). On relève par la suite peu ou pas de différences entre les hommes et les femmes (adolescents et adultes) pour ce qui est des stratégies en lecture (Aek, 2003 ; Poole, 2005 ; Young et Oxford, 1997).

1.3 Différences hommes-femmes liées aux items des tests

Il existe toutefois des chercheurs qui ont abordé le phénomène de la performance et de la familiarité du contenu sous l'angle non pas du sujet abordé dans le texte lu, mais sous l'angle de la question posée pour en mesurer la compréhension. Ainsi, grâce à une classification des items selon leur nature et leurs caractéristiques, on remarque alors des différences chez les hommes et chez les femmes, où les femmes seraient meilleures, par exemple, lorsqu'il faut identifier les idées principales d'un texte, alors que les hommes seraient meilleurs pour y retrouver des détails (Im et Huh, 2007 ; Yazdanpanah, 2007). Les femmes domineraient pour les questions ou items portant sur les sentiments et les impressions, alors que les hommes l'emporteraient sur celles qui demandent de faire des inférences logiques, peu importe le sujet du texte (Im et Huh, 2007 ; Pae, 2004).

Les études de jadis sur le biais des tests lié aux thèmes abordés dans les tests (Chen et Henning, 1985) qui favorisaient un sexe au détriment d'un autre ont suggéré et permis l'élimination de ce biais par une sélection prudente de thèmes neutres. Il est aussi possible, en créant un test de lecture, de prendre en considération les données de recherche susmentionnées afin d'offrir un éventail de questions ou d'items qui, par leur nature, ne favorisent pas plus un sexe que l'autre. Il demeure toutefois impossible comme créateur de tests d'éliminer tout avantage, puisqu'on ne peut changer le fait que les femmes qui passeront le test sont naturellement plus enclines à la lecture, et que ce facteur motivationnel les favorisera dans leur performance.

Néanmoins, une fois un test mis sur pied, il peut toujours subsister des items qui, contre toute attente, favoriseront un sexe plutôt qu'un autre, sans même parfois que l'on sache pourquoi. Identifier la raison n'est pas essentiel ; ce qui compte, c'est de pouvoir identifier et éliminer de tels items afin de présenter un instrument qui soit le plus juste possible.

1.4 Le fonctionnement différentiel d'items et les tests de langues

Le fonctionnement différentiel d'items peut se définir comme une situation où des groupes d'élèves (par exemple, les garçons et les filles) ayant la même habileté estimée ne répondent pas de la même façon à un item. Ainsi, une fois les niveaux d'habiletés de ces deux groupes ramenés sur une même échelle de mesure, il

existerait un fonctionnement différentiel lorsqu'un item serait plus facile pour les filles que pour les garçons.

Plusieurs chercheurs se sont intéressés à étudier le fonctionnement différentiel d'items (aussi noté FDI) dans le cadre de tests de langues. Par exemple, Kunnan (1990) a analysé les résultats obtenus à un test de classement en anglais, langue seconde. Cette étude a permis de mettre en exergue trois principales sources de fonctionnement différentiel d'items susceptibles d'être liées au sexe et à l'origine du répondant : le contexte éducatif d'origine, le fait que certaines langues sont plus familières que d'autres (par exemple, le français et l'espagnol ont la même racine linguistique), et la capacité qu'a un groupe à se reconnaître dans la formulation d'un item. Parmi ces sources, les deux premières relèvent du fonctionnement différentiel d'items lié à une origine culturelle différente et la dernière, pertinente pour cette étude, est celle qui est liée au sexe.

Abbott (2007), de son côté, a étudié le fonctionnement différentiel d'items en utilisant la méthode non paramétrique SIBTEST (Shealy et Stout, 1993). Son étude montre que l'on peut trouver des différences statistiquement significatives entre certains groupes linguistiques qui ont répondu à un test de lecture en anglais langue seconde. Pae (2004), enfin, a étudié les résultats d'étudiants coréens à un test de compréhension en lecture en langue anglaise à l'aide du test de rapport de vraisemblance. Ses résultats ont montré que les items traitant des émotions tendent à être plus facilement réussis par les filles, alors que les garçons réussissent mieux les items à contenu logique.

1.5 Objectif spécifique

Quelques recherches antérieures ont déjà montré une utilisation efficace des méthodes de détection du fonctionnement différentiel d'items dans le contexte des tests de compréhension en lecture en langue anglaise. Avec l'apparition récente d'un nouveau type de test de lecture reposant sur une technique de création d'items réservée auparavant à d'autres fins, la présente recherche vise à examiner la qualité de ce test en vérifiant la présence de fonctionnement différentiel d'items à l'aide de deux méthodes de détection non paramétriques.

2. Contexte théorique

2.1 Le fonctionnement différentiel d'items considéré

Selon Magis, Béland, Tuerlinckx et DeBoeck (2010), les premières méthodes de détection ont été développées durant les années 1960-1970. À ce moment, les auteurs parlaient surtout de détection de l'incidence d'item (situation où un item serait plus difficile pour un groupe d'élèves que pour un autre) ou de biais d'item (certains facteurs tels qu'une mauvaise formulation de la question provoquent une différence de difficulté entre les groupes d'élèves). Dans le cadre de cette recherche, nous nous intéresserons plutôt aux approches qui ont été développées au cours des 35 dernières années.

Plusieurs enjeux importants entourent l'examen du fonctionnement différentiel d'items ainsi que la façon dont on peut aborder et interpréter ce phénomène. Par exemple, on retrouve, comme facteurs importants, le nombre de groupes de participants à l'étude (il y a traditionnellement un groupe focal et un groupe de référence), la nature des données à analyser (à deux choix ou à plus de deux choix de réponse), de même que la nature de la méthode – de type paramétrique ou non paramétrique – à utiliser pour identifier le fonctionnement différentiel d'items. Le lecteur intéressé à en savoir plus sur ce sujet peut consulter l'article de Magis, Béland, Tuerlinckx et De Boeck (2010).

Dans le cadre de cet article, nous nous concentrerons uniquement sur le type de fonctionnement différentiel d'items qui est détecté. Ainsi, il existe deux grandes catégories de fonctionnement différentiel d'items. La plus utilisée est le fonctionnement différentiel d'items uniforme, qui survient lorsqu'un item favorise constamment un groupe de participants au détriment d'un autre pour tous les niveaux de score obtenus à un test. L'autre catégorie de fonctionnement différentiel d'items qui peut être détecté est le fonctionnement différentiel d'items non uniforme, où un item favorise un groupe de participants pour un certain éventail de scores, et un autre groupe pour d'autres niveaux d'habileté. La figure 1 ci-dessous présente un exemple de fonctionnement différentiel d'items pour chacune

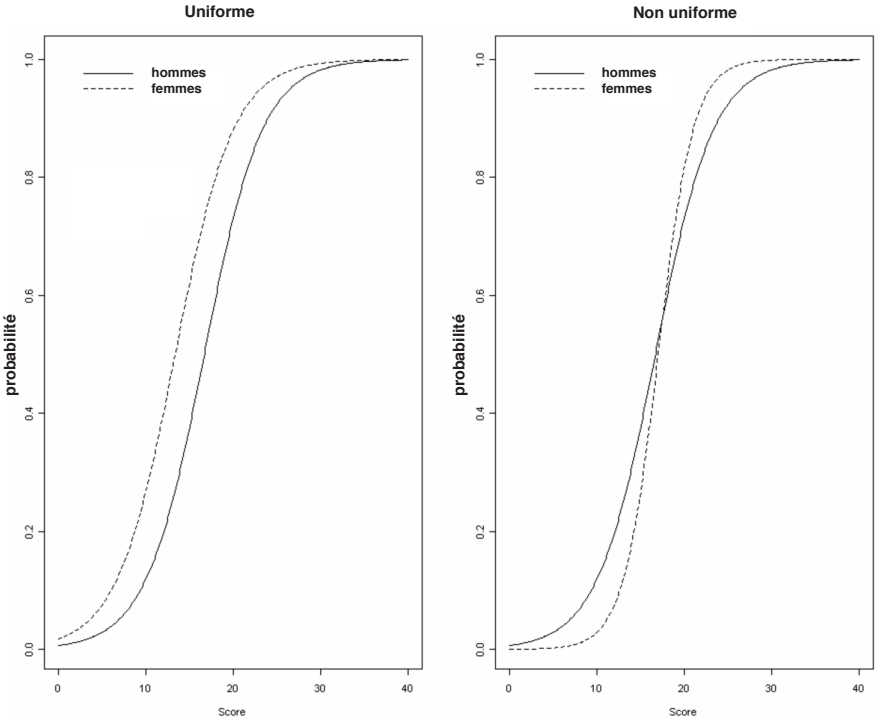


Figure 1. Exemples de fonctionnement différentiel d'items uniforme et non uniforme

des deux catégories. Alors que, dans ces exemples, le fonctionnement différentiel d'items uniforme (image de gauche) favorise constamment les femmes, un autre item du même test présente un fonctionnement différentiel non uniforme, favorisant les hommes pour les niveaux inférieurs à un score total approximatif de 18 au test, et favorisent les femmes pour les scores supérieurs à ce seuil.

Les deux méthodes choisies pour détecter la présence de fonctionnement différentiel lié au sexe sont le test Mantel-Haenszel et la régression logistique.

2.2 Le test Mantel-Haenszel

Le test Mantel-Haenszel (Holland et Thayer, 1988) se fonde sur la comparaison des réponses à un item *j* par l'intermédiaire d'une table de contingence. Ainsi, tel qu'il apparaît dans le tableau 1 ci-dessous, pour chaque item, on analyse une table à deux entrées : appartenance à un groupe focal (F) ou de référence (R) et réponse à l'item (0 ou 1) :

Tableau 1
Tableau de contingence de Mantel-Haenszel

	Score à l'item		
Groupe	1	0	Total
R	A_j	B_j	n_{Rj}
F	C_j	D_j	n_{Fj}
Total	m_{1j}	m_{0j}	T_j

Adapté de Holland et Thayer (1988, p. 130)

Ici, les cellules A_j et B_j sont les bonnes (1) et mauvaises réponses (0) pour le groupe de référence (par exemple, les garçons). Les cellules C_j et D_j correspondent aux bonnes et aux mauvaises réponses pour le groupe focal (par exemple, les filles). Les scores totaux à chacun des groupes sont notés n_{Rj} et n_{Fj} (où $n_{Rj} = A_j + B_j$ and $n_{Fj} = C_j + D_j$) et le nombre total de bonnes et de mauvaises réponses, respectivement, m_{1j} et m_{0j} (où $m_{1j} = A_j + C_j$ et $m_{0j} = B_j + D_j$). Le test Mantel-Haenszel est représenté par l'équation 1.

$$MH = \frac{\left(\left|\sum_j A_j - \sum_j E(A_j)\right| - 0.5\right)^2}{\sum_j Var(A_j)} \qquad (1)$$

où $E(A_j) = \dfrac{n_{Rj}\, m_{1j}}{T_j}$ et $Var(A_j) = \dfrac{n_{Rj}\, n_{Fj}\, m_{1j}\, m_{0j}}{T_j^{\,2}\,(T_j - 1)}$.

Le test Mantel-Haenszel est basé sur un test du χ^2 où le rejet de l'hypothèse nulle consiste à conclure à l'existence d'une relation entre l'appartenance à un groupe et la réponse à l'item.

Ce test est l'une des méthodes de détection du fonctionnement différentiel d'items les plus populaires. L'*Educational testing service* (Holland et Thayer, 1988 ; Zwick et Ercikan, 1989) a aussi développé une norme, *ΔMH*, visant à calculer la taille de l'effet du coefficient obtenu à ce test. Cette norme fournit une information supplémentaire intéressante sur l'importance du fonctionnement différentiel d'items potentiellement détecté. Cette statistique est obtenue en calculant en premier lieu un rapport de cote (*odds ratio*) tel que représenté par l'équation 2.

$$\alpha_{MH} = \frac{\sum A_j D_j / T_j}{\sum B_j C_j / T_j} \tag{2}$$

Ensuite, basée sur la métrique delta de l'*Educational testing service*, la statistique est obtenue à partir de l'équation 3.

$$\Delta_{MH} = -2{,}35 \ln(\alpha_{MH}) \tag{3}$$

Un item présente un fonctionnement différentiel négligeable lorsque *ΔMH* est inférieur à 1,00 en valeur absolue. Un item présente un fonctionnement différentiel intermédiaire lorsque cette statistique se situe entre 1,00 et 1,50 en valeur absolue. Le fonctionnement différentiel d'un item est important lorsque *ΔMH* est supérieur à 1,50 en valeur absolue (Holland et Thayer, 1985).

2.3 Le modèle de régression logistique

Le test Mantel-Haenszel ne permet cependant que de détecter les items qui affichent un fonctionnement différentiel uniforme. Une autre solution est donc nécessaire pour identifier les items qui présentent un fonctionnement différentiel non uniforme. L'approche développée par Swaminathan et Rogers (1990) consiste à utiliser un modèle de régression logistique afin d'estimer les coefficients de l'équation suivante :

$$\text{logit}(\pi) = \beta_0 + \beta_1 S + \beta_2 G + \beta_3 SG \tag{4}$$

Dans l'équation 4, π représente la probabilité de répondre correctement à un item, *S* représente le score total au test (nombre de bonnes réponses) et *G* représente l'appartenance (G = 1) ou non (G = 0) au groupe focal d'intérêt (par exemple, les garçons ou les filles). *SG* représente l'interaction entre l'appartenance ou non au groupe focal et le score total au test.

Ce sont les coefficients estimés qui vont permettre de se prononcer sur le fonctionnement différentiel d'un item. Ici, l'ordonnée à l'origine $\beta 0$ et la pente $\beta 1$

sont communs à tous les répondants au test. S'il n'y a pas de fonctionnement différentiel de l'item, la probabilité d'obtention d'une bonne réponse à l'item est expliquée uniquement par le niveau de difficulté de l'item ($\beta 0$) et par le score du répondant au test (S). Il y a fonctionnement différentiel de l'item lorsque l'un des coefficients $\beta 2$ ou $\beta 3$ est non nul ou lorsque les coefficients sont tous deux non nuls. Le coefficient $\beta 2$ représente la différence existant entre l'ordonnée à l'origine de chacun des deux groupes. Le coefficient $\beta 3$ modélise la différence de pentes selon l'appartenance au groupe focal ou au groupe de référence.

Il existe un fonctionnement différentiel d'items uniforme lorsqu'on a $\beta 3 = 0$ et $\beta 2 \neq 0$. Le fonctionnement différentiel d'items est non uniforme lorsqu'on a $\beta 3 = 0$ indépendamment du fait que $\beta 2$ soit nul ou non. À l'opposé, selon cette logique, il n'existe pas de fonctionnement différentiel d'items si $\beta 2 = 0$ et $\beta 3 = 0$. En effet, la probabilité de répondre correctement à un item particulier ne dépendra pas de l'appartenance au groupe focal ou au groupe de référence.

La détection du fonctionnement différentiel d'items à l'aide de ce modèle est basé sur les quantiles de la distribution de probabilité du χ^2. De plus, des auteurs (Jodoin et Gierl, 2001 ; Zumbo et Thomas, 1997) ont proposé de mesurer de l'importance du fonctionnement différentiel d'items à partir d'un pseudo coefficient de détermination, soit le R^2 de Nagelkerke (1991). Plus spécifiquement, ils utilisent comme mesure de la taille de l'effet la différence des pseudo coefficients entre les groupes focal et de référence, soit ΔR^2. À partir de cette mesure, ces auteurs ont classé l'effet du fonctionnement différentiel d'items en trois grandes catégories : effet négligeable, modéré et important. Selon Jodoin et Gierl (2001), l'effet du fonctionnement différentiel d'items est négligeable si le ΔR^2 est plus petit que 0,035, modéré s'il se situe entre 0,035 et 0,070, et important s'il est supérieur à 0,070. De leur côté, Zumbo et Thomas (1997) ont aussi proposé une classification de l'effet du fonctionnement différentiel d'items, mais les résultats sont alors beaucoup plus conservateurs et moins d'items sont alors identifiés. Nous croyons, à l'instar de Béland, Magis et Raîche (2011), qu'utiliser cette dernière approche pourrait augmenter le risque de ne pas détecter certains items présentant un fonctionnement différentiel d'items (erreur de type II) : ce qui explique pourquoi nous la rejetons dans le cadre de cet article.

En conformité avec la pratique usuelle, les étapes pour identifier un fonctionnement différentiel d'items à l'aide de la méthode de la régression logistique sont les suivantes. En premier lieu, il s'agit de vérifier l'hypothèse selon laquelle les coefficients $\beta 2$ et $\beta 3$ sont tous deux égaux à 0. Dans l'affirmative, on considère la présence d'un fonctionnement différentiel de l'item. Par la suite, on peut consulter directement des experts de contenu pour expliquer ce résultat ou tout simplement rejeter l'item. De plus, si l'on désire expliquer le type de fonctionnement différentiel, il est préférable de vérifier la catégorie dans laquelle il s'inscrit : uniforme ou non uniforme. À cette fin, à la condition que l'hypothèse précédente ait été rejetée, on vérifie une seconde hypothèse selon laquelle le coefficient $\beta 3$ est égal à 0. Si l'hypothèse est rejetée, on considère que cet item affiche un fonctionnement

différentiel non uniforme. Si l'hypothèse ne peut pas être rejetée, et seulement à cette condition, on vérifie une troisième hypothèse selon laquelle le coefficient $\beta 2$ est égal à 0. Le rejet de cette hypothèse permet d'identifier un fonctionnement différentiel uniforme. De plus, à la fin du processus, on peut estimer l'importance du fonctionnement différentiel à l'aide de la taille de l'effet.

2.4 Justification des méthodes de détection sélectionnées

La première raison pour laquelle ces méthodes non paramétriques ont été sélectionnées pour cette étude est qu'elles sont plus simples et plus faciles à utiliser que les méthodes paramétriques, basées sur la théorie de la réponse à l'item. De telles méthodes basées sur la théorie de la réponse à l'item sont contraignantes, car cette théorie repose sur des modèles probabilistes complexes. Les modèles probabilistes qui les sous-tendent exigent le respect de certains postulats : principalement, les tests utilisés pour obtenir les données doivent être unidimensionnels (notons que l'unidimensionnalité s'applique aussi à la théorie classique des tests, bien que les conditions d'application en soient plus souples) et l'indépendance locale entre les items doit être démontrée (voir Bertrand et Blais, 2004, pour plus d'information sur ces postulats). Enfin, ces méthodes requièrent généralement des corpus de données assez importants (Lai, Teresi et Gershon, 2005).

Parmi les modèles non paramétriques, ceux choisis pour cette étude figurent parmi les plus couramment utilisés par les agences de premier plan en administration de tests (par exemple, *Educational Testing Service*) et ils sont cités dans de nombreuses recherches (Ferreres-Traver, Fidalgo-Aliste et Muniz, 2000 ; Zheng, Gierl et Cui, 2007 ; Zwick, Thayer et Lewis, 1997). De plus, ces modèles offrent la possibilité de calculer un indice de la taille de l'effet, c'est-à-dire qu'ils ne font pas qu'identifier quels items présentent un fonctionnement différentiels, mais ils donnent une idée de l'ampleur de ce phénomène. Enfin, ils fournissent des données fiables pour des matrices de données plus réduites, de l'ordre de grandeur habituel que l'on retrouve pour les données réelles issues de plusieurs recherches à petite ou moyenne échelle. Le lecteur intéressé pourra d'ailleurs consulter les recherches d'Awuor (2008) et de Lai, Teresi et Gershon (2005) pour en savoir davantage sur le sujet.

Il faut aussi souligner que Lord (1980) a proposé de réaliser de manière itérative l'analyse du fonctionnement différentiel d'items. C'est ce qu'il a nommé la purification des items. Ainsi, à la première itération, tous les items sont analysés. À la seconde itération, l'analyse est effectuée de nouveau en retirant du calcul du score les items qui ont été identifiés comme présentant un fonctionnement différentiel à l'étape précédente. Les itérations se répètent jusqu'au moment où deux itérations mènent à la même classification des items. On obtient alors les scores de chacun des sujets en-dehors de la présence d'items qui affichent un fonctionnement différentiel, et la dernière itération du processus fournit les items qui affichent réellement un fonctionnement différentiel. Malheureusement, l'approche suggérée par Lord n'est pas toujours appliquée.

L'utilisation conjointe de ces deux méthodes non paramétriques est faite dans une perspective de confirmation mutuelle pour valider la détection de fonctionnement différentiel d'items. Comme ces deux méthodes reposent sur des logiques différentes (régression plutôt que tableau de contingence), une détection des items par les deux méthodes augmente la certitude qu'on est effectivement en présence d'un fonctionnement différentiel pour ces mêmes items. Le but de la présente étude est d'étudier la qualité d'un test d'habileté de compréhension en lecture en anglais par l'utilisation conjointe du test Mantel-Haenszel et du modèle de régression logistique pour détecter des items porteurs de fonctionnement différentiel lié au sexe.

3. Méthodologie

3.1 Sujets

Les participants à l'étude sont 171 étudiants universitaires bilingues (anglais-français), soit 119 femmes et 52 hommes. Ces participants proviennent d'un éventail de disciplines dans le domaine des sciences humaines, recrutés dans deux universités québécoises. Ce rapport hommes/femmes est assez représentatif des facultés de sciences humaines en milieu universitaire québécois, où l'on observe une présence plus élevée de femmes que d'hommes. Une période de 30 minutes durant leur temps de classe habituel a été réservée aux étudiants intéressés à participer à la recherche, les autres pouvant simplement s'absenter, pendant ce temps, sans subir de préjudice. L'âge moyen des participants était de 24 ans, avec un minimum de 18 et un maximum de 53, sans différence notable d'âge entre participants masculins et féminins. Le groupe de participants était très homogène, puisque 162 des 171 participants étaient des Québécois francophones de souche. Le niveau de compétence en anglais indiqué par les participants variait d'intermédiaire à avancé, comme ils l'ont indiqué eux-mêmes sur la page liminaire du test. Aucune personne de niveau débutant ne figure parmi les participants, puisqu'un tel niveau de compétence la rendrait incapable d'effectuer les tâches de lecture requises.

3.2 Instrumentation

Le test qui a servi à recueillir les données à analyser a été créé en 2008 pour mesurer l'habileté de compréhension en lecture en anglais (Pichette, Lafontaine, et de Serres, 2009). Il est basé sur la technique par vérification de phrases (*Sentence verification technique* ou SVT) (Royer, Hastings et Hook, 1979) qui consiste à lire quatre textes de 12 phrases chacun et à indiquer de mémoire, après chacun de ces passages, si oui ou non chacun des 16 items sous forme de phrases correspond au contenu du passage lu. L'historique de cette technique et les détails de nature psychométrique qui y sont associés sont présentés dans Royer (2004). Selon les études qui ont porté sur les mesures de validité à partir de plusieurs applications de la technique SVT à des textes en plusieurs langues, cette technique permet de mesurer ce à quoi on la destine, c'est-à-dire l'habileté en lecture. Par exemple,

avec 54 participants, Royer note une corrélation de 0,73 entre les scores à un test SVT et ceux à la partie lecture du test standardisé *Iowa Test of Basic Skills*. L'auteur note en outre que le coefficient de fidélité d'un test SVT consistant en quatre passages (64 questions) se situe habituellement entre 0,70 et 0,80.

Ainsi, la lecture de quatre textes en anglais d'intérêt général de 12 phrases est suivie de 16 items créés pour chaque texte. Les items se répartissent en quatre catégories distinctes : paraphrases (P), changements de sens (CS), phrases intactes (PI) et distracteurs (D). Ainsi, cinq des 12 phrases du texte sont paraphrasées, en changeant le plus de mots possible tout en maintenant le même sens. Cinq autres phrases subissent un changement de sens : seuls un ou deux mots sont changés, tout en affectant le sens entier de la phrase. Les deux phrases restantes sont laissées intactes, puis quatre phrases plausibles sont ajoutées comme distracteurs. Un exemple de texte et des items qui lui sont associés est présenté en annexe.

Le niveau de facilité des items mesurés par la proportion de bonnes réponses varie entre 0,28 et 0,99. Leur moyenne est égale à 0,82 et leur écart type à 0,15. La figure 2 illustre la distribution de la fréquence relative (pourcentage) du niveau de facilité des items. Somme toute, le niveau de facilité des items du test semble couvrir une étendue raisonnable, quoiqu'on remarque une importante concentration d'items très faciles (niveau de facilité autour de 0,80).

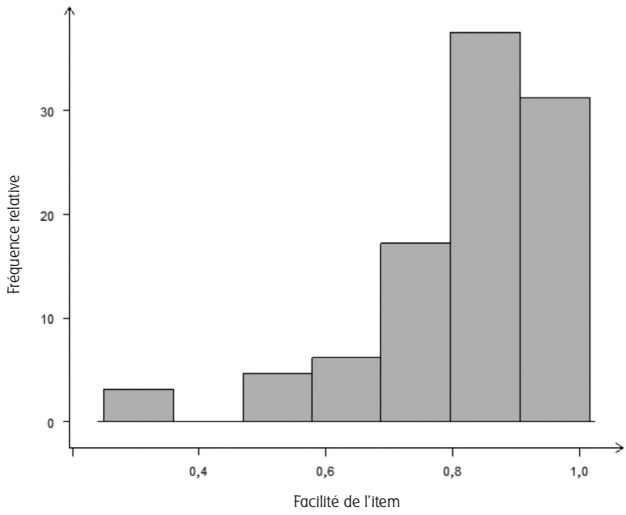


Figure 2. Histogramme représentant la fréquence relative des catégories de niveau de facilité des items au test

3.3 Déroulement

Ce nouvel instrument a été administré en copies papier dans une salle de classe. La durée moyenne de passation du test a été de 21 minutes, et un éventail de durée de 16 à 24 minutes. Les participants se sont faits montrer un exemple de test SVT couplé d'une description du test et de son utilité : nombre de textes et d'items, longueur des textes, durée prévue, etc. Plusieurs de ces informations étaient déjà en possession des participants sur la copie du formulaire de consentement signé qui leur avait été remise. Des consignes leur ont aussi été données oralement par l'expérimentateur, même si certaines se trouvaient également sur le test, comme en témoigne l'extrait fourni en annexe. Les principales consignes consistaient à inviter les participants à lire les textes attentivement, au rythme normal d'une lecture pour fins d'apprentissage, à ne lire chaque texte qu'une seule fois, à ne pas revenir au texte au moment de répondre aux items, à laisser le test face contre le bureau et à quitter la salle lorsqu'ils auraient répondu à tous les items.

3.4 Méthode d'analyse des données

Une fois les tests complétés par les participants, les données manquantes avec lesquelles il a fallu composer étaient de l'ordre de 0,38 % des données recueillies (42 sur 10944) et ne concernaient ni un item ni une section du test en particulier. Ces données manquantes ont été traitées comme de mauvaises réponses et se sont alors vu assigner un score de zéro. Il est à noter que ce type de traitement est fréquent et qu'il a aussi été, par exemple, adopté par Raîche (2002). Les sujets masculins ont été choisis pour former le groupe de référence, les sujets féminins formant le groupe focal.

Les coefficients du test Mantel-Haenszel et du modèle de régression logistique ont été tous deux calculés à l'aide de la librairie difR 4.0 (Magis, Béland et Raîche, 2011), disponible via le logiciel statistique R (R development core team, 2011). Le seuil de signification statistique (erreur de type I) a été fixé à 0,05, aussi bien pour le test Mantel-Haenszel que pour le modèle de régression logistique, et cela, pour chacun des items. La technique de purification des items a été utilisée. La pratique habituelle des études sur le fonctionnement différentiel d'items qui consiste à ne pas corriger l'erreur par test de signification de manière à limiter l'importance de l'erreur d'ensemble a été retenue : le logiciel utilisé n'effectue d'ailleurs pas cette correction. Pour comparer les items identifiés par chacune des méthodes, une représentation graphique du numéro des items pour lesquels on a identifié un fonctionnement différentiel est effectuée. Une indication quant à l'importance de la taille de l'effet est aussi ajoutée. Ensuite, les caractéristiques des items identifiés sont analysées. Enfin, les courbes caractéristiques de fonctionnement différentiel des items identifiés sont affichées. Ces courbes sont produites à partir des paramètres de la fonction associée à la méthode de régression logistique selon le type de fonctionnement différentiel (uniforme ou non uniforme) qui a été identifié à chacun des items retenus par les méthodes de détection.

3.5 Considérations éthiques

Le projet de recherche a été approuvé par les comités d'éthique de la recherche avec les êtres humains des deux universités où les participants ont été testés. Les participants ont été informés au préalable des détails de la recherche : objectifs, déroulement, durée prévue, participation volontaire et droit de retrait en tout temps, etc. Ils ont tous signé un formulaire de consentement à cet effet avant la collecte des données. Ils se sont vu assurés par écrit de la confidentialité dans la gestion des données, qui implique, entre autres, que les données sont anonymisées et qu'aucun nom ne paraîtra dans aucun rapport. Le fait pour les participants de fournir leur nom était d'ailleurs optionnel.

4. Résultats : comparaison de la détection par les deux méthodes

Les résultats au test Mantel-Haenszel appliqué à nos données pour détecter un fonctionnement différentiel d'items uniforme sont présentés à la figure 3, et se trouvent sous la forme d'items individuels répartis autour d'un niveau seuil (ligne horizontale), au-dessus duquel la différence entre les probabilités des hommes et des femmes d'obtenir la bonne réponse est considérée statistiquement significative.

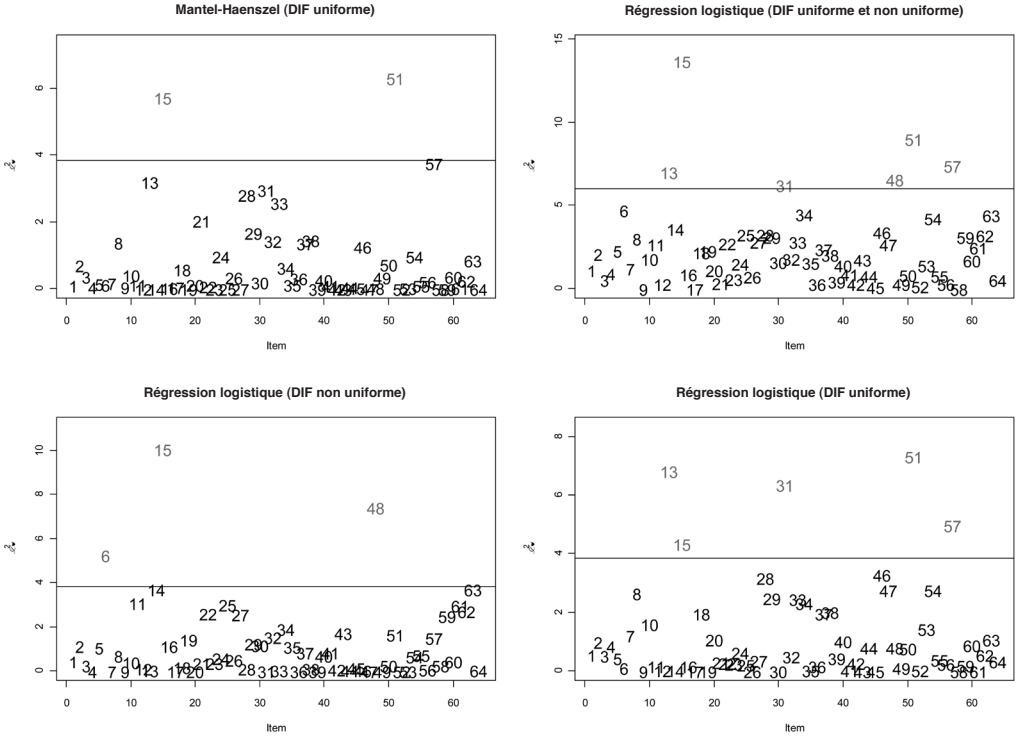


Figure 3. Détection des items par le test Mantel-Haenszel et la régression logistique après purification des items

Dans cette figure, les données présentées ont été obtenues après la purification des items. Plus l'item se trouve éloigné au-dessus du seuil, plus cet item est considéré comme présentant un fonctionnement différentiel. Ce test identifie donc deux items comme présentant un fonctionnement différentiel uniforme, soit les items 15 et 51.

Quant aux résultats qui concernent le modèle de régression logistique, ils sont aussi présentés de façon similaire, à la même figure 3. Ces résultats permettent cependant d'identifier maintenant aussi bien le fonctionnement différentiel d'items uniforme que non uniforme. Dans ce cas-ci, ce sont plutôt sept items qui sont suggérés comme étant sources de fonctionnement différentiel, soit les mêmes deux items identifiés préalablement par le test Mantel-Haenszel, plus cinq items supplémentaires, soit les items 6, 13, 31, 48 et 57. L'item 31 est de toute évidence un cas limite, comme en témoigne sa proximité avec le niveau seuil. Selon les critères présentés préalablement dans le cadre théorique, le modèle de régression logistique indique que quatre items affichent un fonctionnement différentiel uniforme (items 13, 31, 51 et 57), tandis que trois affichent un fonctionnement différentiel non uniforme (items 6, 15 et 48). À l'item 6, on notera une dérogation à la stratégie de détection d'un fonctionnement différentiel. Plus précisément, l'item 6 est le seul pour lequel l'hypothèse nulle selon laquelle le coefficient $\beta 3$ est égal à 0 est rejetée sans avoir préalablement rejeté l'hypothèse nulle selon laquelle les coefficient $\beta 2$ et $\beta 3$ sont tous deux égaux à 0. Il nous a semblé important de noter cette situation, qui sera d'ailleurs commentée plus loin à l'intérieur de la discussion.

À des fins d'analyse, le tableau 2 présente en détail les sept items identifiés par les méthodes utilisées. Dans ce tableau, on indique aussi par des étoiles la taille de l'effet. Celle-ci est mesurée par les statistiques ΔMH et ΔR^2 respectivement pour le test de Mantel-Haenszel et la méthode de la régression logistique. En examinant la nature du fonctionnement différentiel des items en question, on constate de nouveau que, selon la méthode de la régression logistique, quatre présentent un fonctionnement différentiel d'items uniforme et trois un fonctionnement différentiel d'items non uniforme. Fait à souligner, alors que selon le test Mantel-Haenszel, les items 15 et 51 affichent un fonctionnement différentiel d'items uniforme, ce sont les items 13, 31, 51 et 57 qui affichent un tel type de fonctionnement différentiel selon la méthode de la régression logistique. Ainsi, selon la méthode de la régression logistique, l'item 15 affiche plutôt un fonctionnement différentiel non uniforme plutôt qu'uniforme, comme le suggérait, à l'opposé, la méthode de Mantel-Haenszel : l'analyse de la figure 4 montre clairement que cet item affiche bien un fonctionnement différentiel non uniforme. De plus, selon cette même méthode, deux autres items affichent un fonctionnement différentiel uniforme. Cette méthode, avec ces données, permet donc d'identifier deux fois plus d'items dont le fonctionnement différentiel est uniforme. Dans tous les cas, qu'il s'agisse du test Mantel-Haenszel ou de la régression logistique, la taille de l'effet est toujours jugée comme importante. On remarquera qu'aucun item de type phrase intacte (PI) n'a été identifié comme affichant un fonctionnement différentiel.

Tableau 2
Sept items affichant un fonctionnement différentiel et taille de l'effet associée selon les catégories de l'*Educational testing service* (MH) et celles de Jodoin et Gierl (régression logisique)

				Régression logistique	
#	Description de l'item	Type	MH	U	NU
6	*David Blair had left the ship without warning*	D			***
13	*The binoculars were used for seeing nearby ships*	CS		***	
15	*Blair forgot to put the binoculars back one morning.*	CS	***		***
31	*The average rainfall in this rainforest is over 80 inches per year.*	D		***	
48	*Kay played with the chain puzzle with great interest, and managed to solve it in an hour.*	CS			***
51	*The dog would go near the children and wait to be petted.*	P	***	***	
57	*He never learned to cope with his sick eyes.*	CS		***	

CS = changement de sens ; P = paraphrase ; D = distracteur
*** = Effet important
U = uniforme ; NU = non uniforme

La figure 4 ci-dessous montre les courbes de fonctionnement différentiel des sept items identifiés par les deux méthodes. L'inspection de ces courbes montre clairement le fonctionnement différentiel des items correspondants. Plus spécifiquement, les items 6 et 15 indiquent que pour le groupe de référence, celui des hommes, la discrimination entre les sujets est extrêmement forte autour d'un score de 40. L'item 48 affiche une courbe similaire, mais l'étendue des valeurs du score où il permet une bonne discrimination est plus large : de 35 à 45 environ. Les items 13, 51 et 57, pour leur part, sont des cas très clairs de fonctionnement différentiel uniforme, les courbes caractéristiques d'items étant presque parallèles. Dans le cas de l'item 31, la situation est quelque peu particulière. Le niveau de difficulté de l'item ne semble pas du tout varier avec le score des sujets du groupe de référence. Le pouvoir de discrimination de cet item est donc nul sur tout le continuum du score pour ce groupe. Selon la méthode de régression logistique, cet item serait associé à un fonctionnement différentiel uniforme, mais l'inspection de la figure 4 permet toutefois d'en douter.

5. Discussion des résultats

Il arrive que des items qui montrent un fonctionnement différentiel puissent être expliqués à la lumière des résultats de recherche des dernières décennies, évoqués en introduction. Comme nous l'avons vu, le fonctionnement différentiel d'items lié au sexe de l'apprenant peut découler de la capacité qu'a un groupe à se reconnaître dans la formulation d'un item (Kunnan, 1990). Nos résultats semblent en être un cas probant par l'examen de nos items qui présentent un fonctionnement différentiel d'items uniforme, soit les items 13, 31, 51 et 57 (voir figure 4). Deux

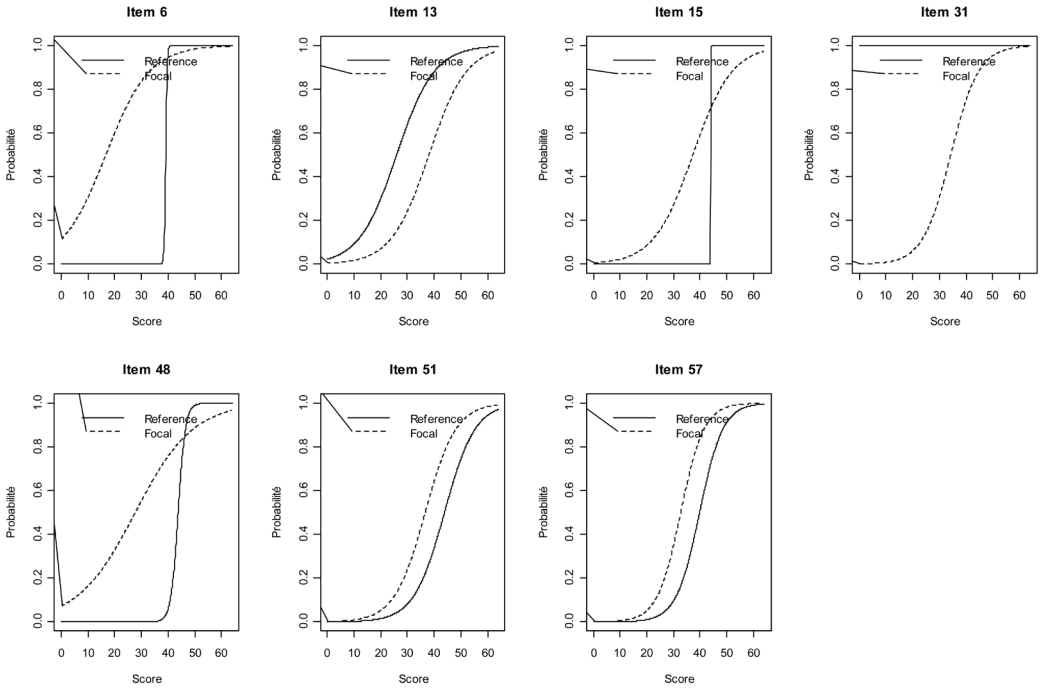


Figure 4. Courbes de fonctionnement différentiel des items identifiés par les deux méthodes pour les hommes (groupe de référence) et les femmes (groupe focal).

items semblent assurément favoriser les femmes pour tous les niveaux de compétence ; il s'agit des items 51 (*The dog would go near the children and wait to be petted*) et 57 (*He never learned to cope with his sick eyes*). Or, ces items sont justement les deux seuls parmi les sept items de la figure 4 qui possèdent clairement une charge affective ou émotionnelle, ce qui, rappelons-le, est justement censé favoriser les femmes dans de tels tests et va dans le même sens que les données issues des études menées par Im et Huh (2007) et par Pae (2004). De façon intéressante, comme ces chercheurs avaient évalué des apprenants coréens, nos résultats viennent s'ajouter aux résultats des recherches antérieures en suggérant que cette tendance, chez les femmes, à prêter davantage attention aux éléments émotionnels en cours de lecture se manifeste également chez les lectrices québécoises.

En outre, toujours parmi les sept items présentant un fonctionnement différentiel selon la figure 4, les deux seuls items qui semblent posséder un contenu clairement scientifique/logique sont les items 13 (*Binoculars are used to see nearby ships*) et 31 (*The average rainfall in this rainforest is over 80 inches per year*). L'item 13 semble reposer sur une inférence logique et devrait donc, sur la base des recherches antérieures, favoriser les hommes (Bügel et Buunk, 1996 ; Chavez,

2001 ; Doolittle et Welch, 1989 ; Hyde et Linn, 1988), ce qui semble être le cas ici, et ce, pour tous les niveaux de compétence observés. Pour ce qui est de l'item 31, le fonctionnement différentiel d'items ressort aussi comme uniforme et il favorise les participants masculins, ce qui va dans le sens des études mentionnées plus haut, puisqu'en plus de sa teneur scientifique, il s'agit d'un item qui porte sur un élément de détail secondaire par rapport à l'essentiel du texte et que de tels items sont censés favoriser les hommes (Im et Huh, 2007 ; Yazdanpanah, 2007). Toutefois, un examen de notre matrice de données laisse voir que seules neuf personnes ont raté cet item, et qu'elles sont toutes de sexe féminin. Les participants masculins ont donc tous réussi l'item 31, une situation qui survient rarement et qui ne permet pas de statuer sur la présence de fonctionnement différentiel pour l'item 31, puisque la courbe de probabilité qui en résulte est égale à un, partout pour ce groupe (voir la figure 4). Il n'est pas impossible qu'il y ait fonctionnement différentiel pour cet item, notamment en raison de sa nature scientifique/logique qui est effectivement censée avantager les hommes, mais le profil de résultats laisse trop peu d'information pour analyser et discuter l'item à cette fin. Il est possible que le score parfait pour les participants masculins soit une aberration ou un artefact statistique qui disparaîtrait avec un nombre plus élevé de participants ou avec un groupe de participants masculins différents. Les cas de fonctionnement différentiel d'items uniforme de notre étude semblent appuyer clairement les hypothèses basées sur les recherches antérieures. Malgré le caractère convaincant de ces observations, une certaine réserve est de mise, car il reste à savoir si la raison formulée sur la base des résultats de recherche est effectivement la seule et la bonne.

En somme, la difficulté d'expliquer les cas de fonctionnement différentiel d'items observés ici est probablement imputable au fait qu'il s'agit d'une habileté de compréhension en lecture en langue seconde, ainsi qu'à la présence de participants de compétence langagière limitée en anglais. En effet, nombre de chercheurs depuis plusieurs décennies soutiennent qu'une compétence limitée dans la langue lue amène de plus grands efforts de décodage et de traitement de la langue en cours de lecture, ce qui saturerait la mémoire de travail et nuirait à la compréhension en laissant moins de resources cognitives disponibles pour le traitement du contenu (Block, 1992 ; Miyake, Carpenter et Just, 1994, Pichette, 2005 ; Pulido, 2009). On a suggéré que la présence de mots inconnus provoque une espèce de vision tunnel (Smith, 1994) et qu'une connaissance insuffisante de la langue lue a tendance à court-circuiter le système de lecture du lecteur (Clarke, 1980). Le point de vue adopté par la plupart des chercheurs en lecture est que la demande trop forte exigée par le décodage d'un texte trop difficile taxe la mémoire de travail et empêche l'utilisation efficace de stratégies de lecture. Ce handicap causé par une connaissance lexicale et syntaxique limitée fait en sorte que c'est chez les personnes plus compétentes en langue seconde que l'on trouvera un profil semblable à ceux observés en lecture en langue maternelle.

Par conséquent, une analyse efficace de nos items qui présentent un fonctionnement différentiel, à la lumière des recherches passées, consiste à porter notre regard sur les scores de niveaux supérieurs pour identifier le sexe qui est favorisé par les fonctionnement différentiel d'items uniformes et non uniformes. Les considérations théoriques mentionnées ci-dessus fournissent une base solide pour décider comment aborder maintenant, parmi les sept items qui présentent un fonctionnement différentiel, les trois items dont le fonctionnement différentiel est non uniforme, soit les items 6, 15 et 48. Ce faisant, on observe pour les trois items un même profil de courbe, où l'item favorise les femmes pour les niveaux de compétence plus bas et, ce qui est le point de mire de nos observations, favorise les hommes pour les niveaux de compétences supérieurs (de l'ordre de 40 à 45 et plus sur un score maximal de 64, selon l'item). Ce que ces items ont en commun et partagent avec l'item 31, c'est qu'ils portent sur un élément d'information qui n'est pas central au scénario du texte :

- dans le cas de l'item 15 (*Blair forgot to put the binoculars back one morning*), le fait que ce soit le soir que le marin ait oublié de remettre les jumelles plutôt que le matin est secondaire : c'est le fait de l'avoir oublié qui est l'élément central ;
- dans le cas de l'item 48 (*Kay played with the chain puzzle with great interest, and managed to solve it in an hour*), le fait que le perroquet ait résolu le problème en moins d'une minute plutôt qu'en moins d'une heure est secondaire, ce qui importe étant l'exploit de l'avoir réussi ;
- enfin, l'item 6 (*David Blair had left the ship without warning*) est un distracteur ajouté, mais on peut supposer qu'avoir eu l'autorisation ou non de s'absenter importe beaucoup moins que le fait en soi que le marin a quitté le navire.

Il est également à noter que les items 6, 15 et 48 semblent aussi présenter une légère charge émotive qui serait plus faible que pour les items 51 et 57, ainsi qu'un contenu légèrement technique. On pourrait alors penser que la charge émotive faible associée à ces items favorise les femmes qui ont un faible niveau de compréhension en lecture, mais que leur contenu plus scientifique ou logique favorise les hommes qui ont un niveau de compréhension élevé en lecture. Si cette interprétation est plausible, pour expliquer, sinon prévoir le fonctionnement différentiel uniforme ou non uniforme d'un item, il ne suffirait pas de se limiter à considérer uniquement le fait qu'un item est associé à une charge émotive ou non, ou encore à un contenu scientifique/logique ou non. Il serait alors préférable d'apprécier le niveau de charge émotive et l'importance du contenu scientifique ou logique au lieu de se limiter simplement à considérer leur présence ou leur absence.

Ainsi, la façon dont nos items favorisent un sexe ou l'autre de façon uniforme ou non uniforme (dans ce dernier cas, en passant par l'observation, chez les apprenants de compétence plus élevée) semble fournir un appui convaincant aux hypothèses émises au cours des années par les chercheurs en langue seconde.

Il importe toutefois de souligner que transformer les items qui affichent un fonctionnement différentiel s'avère presque impossible pour ce test. Comme nous

l'avons indiqué plus tôt, exception faite des distracteurs, chaque item est basé sur une phrase de texte. Par conséquent, si une phrase du texte original concerne les émotions, l'item qui consistera en cette phrase transformée concernera forcément lui aussi les émotions.

Cela dit, avant même d'envisager la modification du test, il importe de nous assurer que nous sommes effectivement en présence de fonctionnement différentiel d'items. Qu'adviendrait-il des six items identifiés si beaucoup plus de participants étaient testés : continueraient-ils de montrer un fonctionnement différentiel ? Il est probable que seuls les items 15 et 51 seraient susceptibles de le faire, mais il s'agit de conjectures que seules des études supplémentaires permettront de vérifier.

5. Conclusion

La présente étude visait à comparer l'identification de fonctionnement différentiel d'items lié au sexe dans un nouveau test d'habileté de compréhension en lecture de l'anglais. Pour ce faire, deux méthodes non paramétriques ont été comparées : le test Mantel-Haenszel et le modèle de régression logistique. Alors que le premier a permis d'identifier deux des 64 items du test comme présentant un fonctionnement différentiel, le second a permis de relever ces items en plus de cinq items supplémentaires.

Cependant, la prudence est de mise pour ces détections qui reposent sur des tests d'hypothèses autour d'un seuil α fixé par convention à 0,05 dans le domaine des sciences humaines et sociales, en-dessous duquel on juge le test d'hypothèse statistiquement significatif. Si on se place hors de l'approche fréquentiste des tests d'hypothèse, ce dernier concept est une question de degrés et n'est pas un phénomène binaire selon lequel un fait observé est soit significatif soit non significatif. Tout ce qui est modérément proche d'un niveau seuil aux fondements subjectifs, (par exemple, l'item 31 selon la régression logistique), s'avère hasardeux à qualifier. De plus, comme dans la plupart des études de fonctionnement différentiel d'items, les tests d'hypothèse utilisés ne tenaient pas compte de l'erreur d'ensemble. Il serait d'intérêt de reproduire cette recherche en en tenant compte.

Pour les méthodes d'analyses choisies, Zieky (1993) recommande des échantillons selon les standards de l'Éducational testing service (ETS) avec au moins 200 sujets dans le plus petit groupe ainsi qu'un total de plus de 600 participants pour les deux groupes combinés. Bien que cela aurait été préférable, le défi aurait été de taille, puisque qu'à un certain moment plus de 500 personnes ont été sollicitées en classe, ce qui n'a permis de rejoindre qu'une dizaine de participants seulement. Comme le reconnaît Zieky, face au risque de ne tout simplement pas atteindre le nombre désiré, de deux maux nous avons choisi le moindre.

Non seulement il serait intéressant de voir si un nombre supérieur de participants confirmerait ces items comme porteurs de fonctionnement différentiel, en particulier pour le cas intéressant de l'item 31, mais il serait intéressant de mener les mêmes analyses avec des méthodes paramétriques ; par exemple, celle du χ^2

de Lord. Une telle entreprise permettrait non seulement d'obtenir une image plus complète du phénomène, mais également d'examiner si les patrons de détection correspondent ou diffèrent entre les méthodes paramétriques et non paramétriques, en particulier en fonction de groupes de tailles différentes. Une telle approche s'avère malheureusement inapplicable dans notre étude, vu que le faible nombre de participants dans chaque groupe engendrerait des estimations des paramètres d'items peu fiables.

Enfin, il n'est pas toujours, voire rarement, possible d'expliquer avec certitude pour quelle(s) raison(s) un item montrerait un fonctionnement différentiel, mais veiller à en prévenir la présence est une étape importante dans la mise sur pied de tests de langue équitables pour tous les participants.

Quand aux pistes de recherches futures, il serait d'un grand intérêt de poursuivre nos travaux par une étude de l'impact du niveau de la charge émotionnelle et de l'importance du contenu scientifique ou logique des items sur l'ampleur du fonctionnement différentiel uniforme et non uniforme des items. Cette approche permettrait éventuellement d'estimer à l'avance l'ampleur du fonctionnement différentiel d'un item sans toujours avoir à mettre en place un plan d'analyse lourd à chaque fois que de nouveaux items sont ajoutés à un test. Dans ce contexte, on pourrait imaginer appliquer une modélisation à niveaux multiples où les paramètres du modèle de la régression logistique seraient eux-mêmes expliqués par les caractéristiques des items que sont la charge émotionnelle et le contenu scientifique ou logique. Cette approche cadre d'ailleurs très bien avec les propositions bayésiennes d'analyse exploratoire de réponses aux items d'un test (De Boeck et Wilson, 2004 ; Fox, 2010).

ENGLISH TITLE • Evaluation of an English reading test based on two methods for detecting differential item functioning

SUMMARY • This study examines the presence of gender-based differential item functioning on a reading comprehension test in English administered to 171 French-speaking university students. Two non-parametric methods are used : the Mantel-Haenszel test and the logistic regression model. On a total of 64 items, two items are identified by the Mantel-Haenszel test, whereas five additional items are identified by logistic regression. This low number of items suggests reasonable fairness for the test, but the discrepancies observed stress the need for further analyses to clarify the status of those items.

KEYWORDS • Differential item functioning, sentence verification technique, gender bias, Mantel-Haenszel test, logistic regression model.

TITULO • Evaluación de una prueba de lectura en inglés mediante dos métodos de detección del funcionamiento diferencial de los ítems

RESUMEN • El presente estudio se propone examinar la presencia de funcionamiento diferencial de los ítems según el género de los informantes en una prueba de comprensión de lectura en inglés administrada a 171 universitarios francófonos. Se utilizan dos métodos no paramétricos:

la prueba de Mantel-Haenszel y el modelo de regresión logística. En un total de 64 ítems, dos presentan un funcionamiento diferencial según la prueba de Mantel-Haenszel, mientras cinco ítems más se destacan por la regresión logística. Este bajo número de ítems sugiere una buena equidad de la prueba, pero las diferencias observadas muestran la necesidad de análisis adicionales para aclarar el estatuto de estos ítems.

PALABRAS CLAVES • funcionamiento diferencial de los ítems, prueba de verificación de frases, sesgo de género, prueba de Mantel-Haenszel, modelo de regresión logística.

Références

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading. *Language testing, 24*, 8-36.

Aek, P. (2003). A closer look at gender and strategy use in L2 reading. *Language learning*, *53*(4), 649-702.

Al-Shumaimeri, Y. A. N. (2005). *Gender differences in reading comprehension performance in relation to content familiarity of gender-neutral texts*. Présentation faite au Second international conference : Language, culture and literature, Minia University, Égypte.

Anderson, R. C. (1984). Role of the reader's schema in comprehension, learning, and memory. Dans R. C. Anderson, J. Osborn et R. J. Tierney (Dir.) : *Learning to read in American schools : basal readers and content texts*. Hillsdale, New Jersey : Lawrence Erlbaum.

Awuor, R. A. (2008). *Effect of unequal sample sizes on the power of DIF detection : an IRT-based monte carlo study with SIBTEST and Mantel-Haenszel procedures*. Thèse de doctorat inédite, Virginia Polytechnic Institute and State University, Blacksburg : Virginie.

Béland, S., Magis, D. et Raîche, G. (2011). Étude de l'habileté des sujets et de la détection du fonctionnement différentiel d'items dans le cadre du test de classement en anglais-langue seconde (TCALS-II). Dans J.-G. Blais et J.-L. Gilles (Dir.) : *Évaluation des apprentissages et TIC*. Québec, Québec : Presses de l'Université Laval.

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

Block, E. (1992). See how they read : comprehension monitoring of L1 and L2 readers. *TESOL quarterly, 26*, 319-344.

Bramski, D. et Williams, R. (1984). Lexical familiarization in economics text, and its pedagogic implications in reading comprehension. *Reading in a foreign language, 2*, 169-181.

Brantmeier, C. (2003). Does gender make a difference ? Passage content and comprehension in second language reading. *Reading in a foreign language*, *15*(1), 1-27.

Bügel, K. et Buunk, B. P. (1996). Sex differences in foreign language text comprehension : the role of interests and prior knowledge. *The modern language journal, 80*, 15-31.

Carrell, P. L. (1984). Schema theory and ESL reading : classroom implications and applications. *The modern language journal, 68*, 332-343.

Carrell, P. L. (1987). Content and formal schemata in ESL reading. *TESOL quarterly, 21*, 461-481.

Carrell, P. L. et Wise, T. E. (1998). The relationship between prior knowledge and topic interest in second language reading. *Studies in second language acquisition*, *20*, 285-309.

Chavez, M. (2001). *Gender in the language classroom*. Boston, Massachussetts : Heinle et Heinle

Chen, Z. et Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language testing, 2*(2), 155-163.

Chiu, M. M. et McBride-Chang, C. (2006). Gender, context, and reading : a comparison of students in 43 countries. *Scientific studies of reading*, *10*(4), 331–362.

Clarke, M. A. (1980). The short-circuit hypothesis of ESL reading or when language competence interferes with reading performance. *The modern language journal, 64*, 203-209.

De Boeck, P. et Wilson, M. (2004). *Explanatory item response models. A generalized linear and nonlinear approach.* New York, New Jersey : Springer.

Doolittle, A. et Welch, C. (1989). *Gender differences in performance on a college-level achievement test (ACT Research Report Series 89-9)*. Iowa City, Iowa : American college testing program.

Ferreres-Traver, D. F., Fidalgo-Aliste, A. M. et Muniz, J. (2000). Detection of non-uniform DIF : Mantel-Haenszel and logistic regression methods. *Psicothema*, *12*(Supplément 2), 220-225.

Fox, J.-P. (2010). *Bayesian item response modeling. Theory and applications*. New York, New Jersey : Springer.

Harper, S. N. et Pelletier, J. P. (2008). Gender and language issues in assessing early literacy group differences in children's performance on the Test of Early Reading Ability (SAGE). *Journal of psychoeducational assessment, 26*(2), 185-194.

Holland, P. W. et Thayer, D. T. (1985). *An alternate definition of the ETS delta scale of item difficulty* (rapport de recherche RR-85-43). Princeton, New Jersey : Educational testing service.

Holland, P. W. et Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. Dans H. Wainer et H. I. Braun (Dir.) : *Test validity*. Hillsdale, New Jersey : Lawrence Erlbaum Associates.

Hyde, J. et Linn, M. (1988). Gender differences in verbal activity : a meta-analysis. *Psychological bulletin, 104*, 53–69.

Im, B.-B. et Huh, J.-H. (2007). Gender Differences in L2 proficiency test by test material. *Journal of Pan-Pacific association of applied linguistics*, *11*(1), 1-15.

Jodoin, M. G. et Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with logistic regression procedure for DIF detection. *Applied measurement in education*, *14*, 329-349.

Koh, M. Y. (1985). The role of prior knowledge in reading comprehension. *Reading in a foreign language, 3*, 375-380.

Kunnan, A. J. (1990). DIF in native language and gender groups in an ESL placement test. *TESOL quarterly, 24*, 741-746.

Lai, J-S., Teresi, J. et Gershon, R. C. (2005). Procedures for the analysis of differential item functioning (DIF) for small sample sizes. *Evaluation and the health profession, 28*, 283-294.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, New jersey : Lawrence Erlbaum Associates.

Lynn R. et Graham, W. R. (1993). Sex differences in second-language ability : an Irish study. *School psychology international*, *14*(3), 275-279.

Lynn, R. et Mikk, J. (2009). Sex differences in reading achievement. *Trames*, *13*(1), 3-13.

Magis, D., Béland, S. et Raîche, G. (2011). *difR 4.0 : Collection of methods to detect dichotomous differential item functioning (DIF) in psychometrics*. Vienne, Autriche : R foundation for statistical computing.

Magis, D., Béland, S., Tuerlinckx, F. et DeBoeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior research methods*, *42*, 847-862.

Miyake, A., Carpenter, P. A. et Just, M. A. (1994). A capacity approach to syntactic comprehension disorders : making normal adults perform like aphasic patients. *Cognitive neuropsychology, 11*(6), 671-717.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*, 691-692.

Pae, T.-I. (2004). Gender effect on reading comprehension with Korean EFL learners. *System*, *32*, 265–281.

Pichette, F. (2005). Time spent on reading and reading comprehension in second language learning. *Canadian modern language review, 62*(2), 243-262.

Pichette, F., Lafontaine, M. et de Serres, L. (2009). *A new tool for measuring L2 reading comprehension ability*. Présentation faite au colloque EUROSLA. Cork, Irlande.

Poole, A. (2005). Gender differences in reading strategy use among ESL college students. *Journal of college reading and learning*, *36*(1), 7-20.

Pulido, D. (2009). Vocabulary processing and acquisition through reading : evidence for the rich getting richer. Dans Z. Han et N. J. Anderson (Dir.) : *Second language reading research and instruction : crossing the boundaries*. Ann Arbor, Michigan : University of Michigan Press.

R development core team (2011). *R : a language and environment for statistical computing. Reference index*. Vienne, Autriche : R foundation for statistical computing.

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

Rosen, M. (2001). Gender differences in reading performance on documents across countries. *Reading and writing, 14*(1-2), 1-38.

Royer, J. M. (2004). *Uses for the sentence verification technique for measuring language comprehension*. Amherst, Massachussets : Reading success lab.

Royer, J. M., Hastings, C. N. et Hook, C. (1979). A sentence verification technique for measuring reading comprehension. *Journal of reading behavior, 11*, 355-363.

Shealy, R. T. et Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability diffrences and detects tes bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Smith, F. (1994). *Understanding reading : a psycholinguistic analysis of reading and learning to read* (5e édition). New York, New Jersey : Holt, Rinehart et Winston.

Stevens, K. (1980). The effect of background knowledge on the reading comprehension of ninth graders. *Journal of reading behavior, 12*, 151-154.

Swaminathan, H. et Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of educational measurement, 27*, 361-370.

White, B. (2007). Are girls better readers than boys ? Which boys ? Which girls ? *Canadian journal of education*, *30*(2), 554-581.

Wolf, M. et Gow, D. (1986). A longitudinal investigation of gender differences in language and reading development. *First language, 6*(17), 81-110.

Yazdanpanah, K. (2007). The effect of background knowledge and reading comprehension test items on male and female performance. *The reading matrix*, *7*(2), 64-80.

Young, D. J. et Oxford, R. (1997). Gender-related analysis of strategies used to process written input in the native language and a foreign language. *Applied language learning*, *8*(1), 43-73.

Zheng, Y., Gierl, M. J. et Cui, Y. (2007). *Using real data to compare DIF detection and effect size measures among Mantel-Haenszel, SIBTEST, and logistic regression procedures*. Edmonton, Alberta : University of Alberta, Centre for research in applied measurement and evaluation.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. Dans P. W. Holland et H. Wainer (Dir.) : *Differential item functioning*. Hillsdale, New Jersey : Lawrence Erlbaum and associates.

Zumbo, B. D. et Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF*. Prince George, Colombie-Britanique : University of Northern British Columbia, Edgeworth Laboratory for Quantitative Behavioral Science.

Zwick, R. et Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of educational measurement*, *26*, 55-66.

Zwick, R., Thayer, D. T. et Lewis, C. (1997) An investigation of the validity of an empirical Bayes approach to Mantel-Haenszel DIF analysis. *ETS research report No. 97-21*. Princeton, New Jersey : Educational testing service.

Correspondance
Pichette.francois@teluq.ca
raiche.gilles@uqam.ca
beland.sebastien@uqam.ca
david.magis@ulg.ac.be

Contribution des auteurs
François Pichette : 45 %
Gilles Raîche : 25 %
Sébastien Béland : 20 %
David Magis : 10 %

Ce texte a été révisé par Sandra Najac.

Texte reçu le : 7 septembre 2011
Version finale reçue le : 21 mars 2012
Accepté le : 18 juin 2012

Annexe : exemple de texte et des items accompagnateurs

Lisez l'histoire suivante lentement et attentivement, une seule fois, en vous concentrant.

A special volunteer

Barkley, our dog, came to me when he was three years old after living with a family that could no longer take care of him.

I took him to visit the school for the blind where I worked as a teacher. He would walk over to the children and wait for a child to pet him.

One day, he started bumping into the walls of our house. When we played ball in the yard, I noticed that he could not catch it. I took him to the veterinarian, who found that he had an eye illness. Barkley had to have several operations. He soon learned to function with his weak eyes. When he got better, he stood at the door, blocking my way, trying to tell me that he wanted to go to school with me and visit his friends. I started taking him to school again. Everyone was happy. Barkley was the happiest of all.

UNE FOIS TERMINÉ, TOURNEZ LA PAGE ET RÉPONDEZ AUX QUESTIONS.
NE REVENEZ PAS À L'HISTOIRE

Lisez attentivement chacune des phrases suivantes, dont l'ordre peut être différent de celui du texte.
- Écrivez « YES » si la phrase lue signifie la même chose que dans le texte.
- Écrivez « NO » la phrase a un sens différent du texte, ou si cela n'a pas été dit explicitement dans le texte.
Les mots n'ont pas à être les mêmes.

1. I took him to visit the old-age home where I worked as a nurse.
2. I received Barkley, our dog, when he was three years old after he lived with people who could not take care of him anymore.
3. The dog would go near the children and wait to be petted.
4. Barkley was so happy that he pulled the leash on the way to school.
5. I took him to the veterinarian, who found that he had an eye illness.
6. One day, he started falling.
7. Barkley was not happy after the operation because he could not see his friends.
8. When we played outside, I realized that he could not catch the ball.
9. He never learned to cope with his sick eyes.
10. The vet had to operate on Barkley several times.
11. That dog was intelligent and eager to please.
12. When he got better, he stood at the door, blocking my way, trying to tell me that he wanted to go to school with me and visit his friends.
13. I started playing with him at school again.
14. Barkley was always very affectionate to the family with whom he lived.
15. The blind children were happiest of all.
16. All the kids were joyful.