Janvier 2009 **Numéro 43**

Cahiers de l'IMA

Comment redresser un test biaisé

Jean-Philippe Antonietti

Institut de Mathématiques Appliquées
Faculté des S.S.P.
Université de Lausanne
Anthropole
1015 Lausanne

Comment redresser un test biaisé

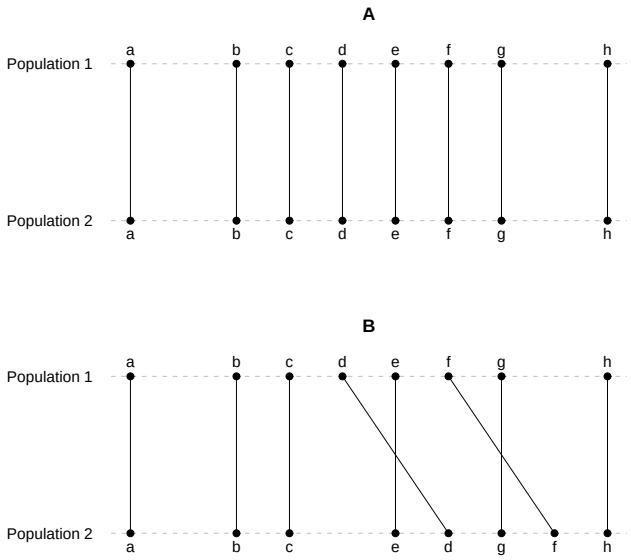
Jean-Philippe Antonietti

Sommaire

1 Introduction

En Suisse, les responsables de l'éducation se demandent si les compétences en mathématiques sont les mêmes dans tous les cantons. Pour pouvoir répondre à une telle question, il faut disposer d'un instrument fiable. Une façon simple de procéder consiste à soumettre aux élèves des différents cantons une épreuve de mathématiques composées de plusieurs items dichotomiques puis à compter le nombre de bonnes réponses fournies par chaque élève. Cette manière classique de faire, qui se fonde sur le calcul d'un score global, souffre de quelques défauts et peut être améliorée en recourant au modèle de Rasch [1, 4]. Dans ce nouveau cadre, pour que la comparaison de la position des populations ait un sens, il ne faut pas que les items fonctionnent différentiellement. Cela veut dire que les difficultés des items doivent être les mêmes dans toutes les populations étudiées. Illustrons notre propos.
La Figure 1A représente une situation dans laquelle tous les items fonctionnent de la même manière dans la Population 1 et dans la Population 2. Dans les deux populations, les items occupent les mêmes positions le long du trait latent : à gauche se trouvent les items les plus faciles (les items a, b et c), à droite les plus difficiles (les items g et h).
La Figure 1B représente, quant à elle, une situation dans laquelle les items d et f fonctionnent différentiellement. Ces deux items sont plus difficiles pour les individus de la Population 2 qu'ils ne le sont pour ceux de la Population 1.

FIGURE 1 – *Difficultés des items estimées dans deux populations différentes. En A, tous les items fonctionnent de la même manière. En B, les items d et f fonctionnent différentiellement.*



Deux individus i_1 et i_2, également compétents, occupant donc la même position θ_i sur le trait latent, mais appartenant l'un à la Population 1 et l'autre à la Population 2 ne réaliseront pas la même performance au test décrit par la Figure 1B. Les items d et f étant plus difficiles pour l'individu issu de la Population 2, il y a moins de chance que ce dernier y réponde correctement. Selon le modèle de Rasch, la probabilité qu'un individu i ayant les compétences θ_i donne une réponse correcte à l'item j de difficulté β_j égale :

$$P(x_{i,j} = 1 \mid \theta_i,\ \beta_j) = \frac{e^{(\theta_i - \beta_j)}}{1 + e^{(\theta_i - \beta_j)}}. \tag{1}$$

Symbolisons la difficulté de l'item d dans les Populations 1 et 2 par $\beta_d(1)$ et $\beta_d(2)$ respectivement. Comme $\beta_d(1) < \beta_d(2)$, il s'ensuit que :

$$P(x_{i_1,d} = 1 \mid \theta_{i_1} = \theta_i,\ \beta_d(1)) > P(x_{i_2,d} = 1 \mid \theta_{i_2} = \theta_i,\ \beta_d(2)). \tag{2}$$

Pour l'item f, le raisonnement est analogue. Comme $\beta_f(1) < \beta_f(2)$, il s'ensuit également que :

$$P(x_{i_1,f} = 1 \mid \theta_{i_1} = \theta_i,\ \beta_f(1)) > P(x_{i_2,f} = 1 \mid \theta_{i_2} = \theta_i,\ \beta_f(2)). \tag{3}$$

Un tel test est une calamité puisqu'en s'y fiant on pourrait conclure que deux populations occupent la même position alors que ce n'est pas le cas ou, à l'inverse, conclure que leur position est différente, alors qu'elles se trouvent au même endroit. À la lumière de ce qui précède, il paraît nécessaire de développer une méthode qui permette de comparer la position de deux populations même lorsque le test utilisé contient un certain nombre d'items fonctionnant différentiellement.
Dans le travail qui suit, nous commencerons par étudier dans une perspective classique l'impact de la composition d'un test sur l'estimation de la position de deux populations. Nous montrerons ensuite comment s'accommoder de la présence dans un test d'items fonctionnant différentiellement.

2 Mauvaise influence des items fonctionnant différentiellement

Nous allons créer artificiellement quelques situations et les étudier en détail à l'aide de méthodes classiques.

2.1 Populations et échantillons

Nous supposerons que les compétences des individus appartenant aux Populations 1 et 2 se distribuent selon une même loi normale centrée et réduite :

$$\text{Population 1 :} \qquad \theta_{i_1} \sim \mathcal{N}(0,\ 1) \tag{4}$$

$$\text{Population 2 :} \qquad \theta_{i_2} \sim \mathcal{N}(0,\ 1) \tag{5}$$

Nous tirerons de chacune des populations un échantillon de taille 1000.

2.2 Tests utilisés

Créons six tests différents. Chaque test est constitué de trois parties A, B et C. La première partie contient 30 items uniformément répartis entre -3 et $+3$ (Figure 2A). Les difficultés de ces items sont listées dans le Tableau 1.

TABLEAU 1 – *Difficulté des items de la partie A de chaque test.*

Item	Difficulté			
1	-3	$+$	0	$\cdot\, 6/29$
2	-3	$+$	1	$\cdot\, 6/29$
3	-3	$+$	2	$\cdot\, 6/29$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
i	-3	$+$	$(i-1)$	$\cdot\, 6/29$
$\vdots$	$\vdots$		$\vdots$	$\vdots$
30	-3	$+$	29	$\cdot\, 6/29$

La deuxième partie est formée de 6 items systématiquement plus faciles pour les individus de la Population 1. Dans la Population 1, les items se distribuent uniformément entre -2.4 et $+1.6$. Dans la Population 2, ils se distribuent uniformément entre -1.6 et $+2.4$ (Tableau 2, Figure 2B). La troisième partie est formée de 6 items, mais cette fois systématiquement plus faciles pour les individus de la Population 2. Dans la Population 1, les items se distribuent uniformément entre $-2+\delta/2$ et $+2+\delta/2$. Dans la Population 2, ils se distribuent uniformément entre $-2-\delta/2$ et $+2-\delta/2$ (Tableau 3, Figure 2C).

TABLEAU 2 – *Difficulté des items de la partie B de chaque test.*

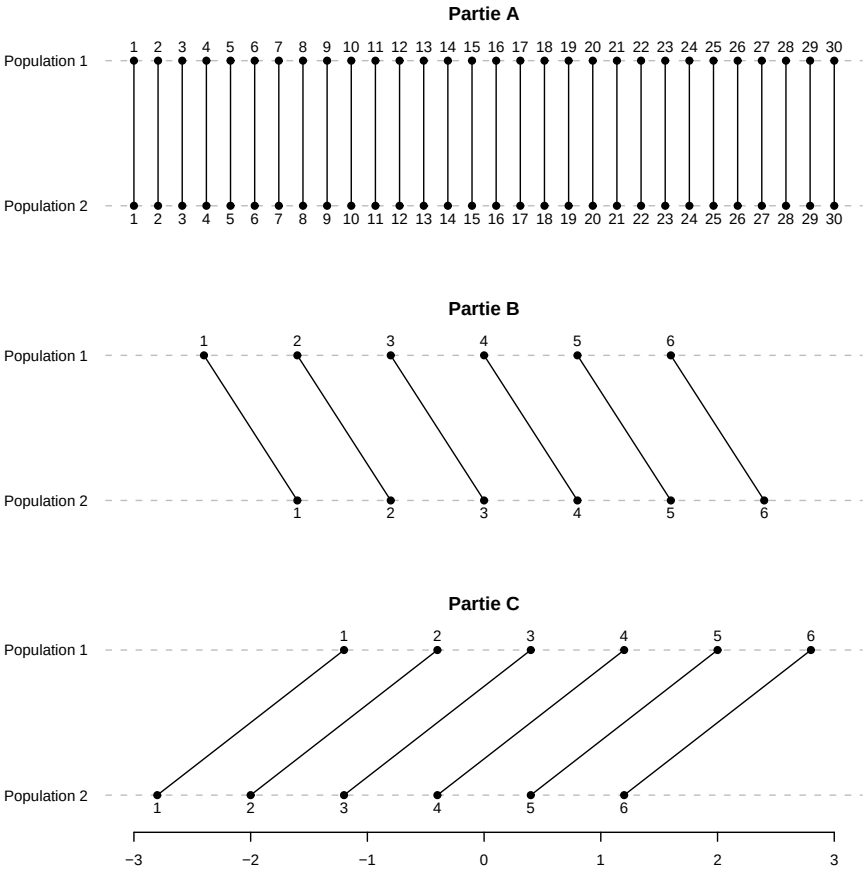
Item	1	2	3	4	5	6
Population 1	-2.4	-1.6	-0.8	0.0	$+0.8$	$+1.6$
Population 2	-1.6	-0.8	0.0	$+0.8$	$+1.6$	$+2.4$

TABLEAU 3 – *Difficulté des items de la partie C de chaque test.*

Item	1	2	3
Population 1	$-2.0+\delta/2$	$-1.2+\delta/2$	$-0.4+\delta/2$
Population 2	$-2.0-\delta/2$	$-1.2-\delta/2$	$-0.4-\delta/2$

Item	4	5	6
Population 1	$+0.4+\delta/2$	$+1.2+\delta/2$	$+2.0+\delta/2$
Population 2	$+0.4-\delta/2$	$+1.2-\delta/2$	$+2.0-\delta/2$

FIGURE 2 – *Composition du Test 6 pour lequel δ est égal à* 1.6.



Les parties A et B sont communes aux six tests. La partie C change d'un test à l'autre en fonction de la valeur attribuée à δ. Ce paramètre δ prend la valeur 0.0, 0.1, 0.2, 0.4, 0.8 ou 1.6 selon le test considéré (Tableau 4).

TABLEAU 4 – *Valeur de δ pour les différents tests.*

Test	1	2	3	4	5	6
δ	0.0	0.1	0.2	0.4	0.8	1.6

2.3 Passation des tests

Les individus sélectionnés aléatoirement pour participer à l'expérience ont été soumis aux six tests décrits ci-dessus (§ 2.2). Leurs réponses sont conformes au modèle de Rasch. La probabilité qu'un individu donne une bonne réponse à un item dépend exclusivement de la différence entre sa compétence et la difficulté de l'item (Tableau 5).

TABLEAU 5 – *Modèle probabiliste utilisé pour simuler les réponses selon l'appartenance des individus et l'origine des items.*

Individu		Item		Probabilité
Appartenance	Compétence	Partie	Difficulté	de réussite
Population 1	θ_{i_1}	A	β_j^A	$\frac{e^{(\theta_{i_1}-\beta_j^A)}}{1+e^{(\theta_{i_1}-\beta_j^A)}}$
	θ_{i_1}	B	$\beta_j^B(1)$	$\frac{e^{(\theta_{i_1}-\beta_j^B(1))}}{1+e^{(\theta_{i_1}-\beta_j^B(1))}}$
	θ_{i_1}	C	$\beta_j^C(1)$	$\frac{e^{(\theta_{i_1}-\beta_j^C(1))}}{1+e^{(\theta_{i_1}-\beta_j^C(1))}}$
Population 2	θ_{i_2}	A	β_j^A	$\frac{e^{(\theta_{i_2}-\beta_j^A)}}{1+e^{(\theta_{i_2}-\beta_j^A)}}$
	θ_{i_2}	B	$\beta_j^B(2)$	$\frac{e^{(\theta_{i_2}-\beta_j^B(2))}}{1+e^{(\theta_{i_2}-\beta_j^B(2))}}$
	θ_{i_2}	C	$\beta_j^C(2)$	$\frac{e^{(\theta_{i_2}-\beta_j^C(2))}}{1+e^{(\theta_{i_2}-\beta_j^C(2))}}$

2.4 Résultats

Calculons classiquement pour chaque individu la somme des points obtenus à nos différents tests et voyons ce qu'il est possible d'inférer concernant la position des populations à partir des caractéristiques des échantillons. Rappelons que si les tests n'étaient pas biaisés, nous devrions

pouvoir conclure que les moyennes des deux populations sont les mêmes. Plus exactement, nous ne devrions pas être en mesure de rejeter l'hypothèse nulle du test de Student à deux groupes indépendants selon laquelle l'espérance de la Population 1 est égale à celle de la Population 2 :

$$H_0 \ : \quad \mu_1 = \mu_2 \tag{6}$$

Les résultats de nos inférences statistiques sont résumés dans le Tableau 6 et représentés graphiquement dans la Figure 3.

TABLEAU 6 – *Performances réalisées aux différents tests par les Groupes 1 et 2 d'individus issus des Populations 1 et 2 respectivement. τ représente la taille de l'effet, t est la valeur empirique de la variable de décision du test de Student à deux groupes indépendants et p est la probabilité critique.*

Test 1

Groupe 1	Groupe 2
$\overline{x}_1 = 21.553$	$\overline{x}_2 = 20.650$
$s_1 = 6.493$	$s_2 = 6.590$
$\tau = -0.138$	
$t[1998] = -3.085$	
$p = 0.002$	

Test 2

Groupe 1	Groupe 2
$\overline{x}_1 = 21.367$	$\overline{x}_2 = 20.663$
$s_1 = 6.496$	$s_2 = 6.659$
$\tau = -0.107$	
$t[1998] = -2.392$	
$p = 0.017$	

Test 3

Groupe 1	Groupe 2
$\overline{x}_1 = 21.503$	$\overline{x}_2 = 20.815$
$s_1 = 6.583$	$s_2 = 6.460$
$\tau = -0.106$	
$t[1998] = -2.371$	
$p = 0.018$	

Test 4

Groupe 1	Groupe 2
$\overline{x}_1 = 21.218$	$\overline{x}_2 = 20.784$
$s_1 = 6.445$	$s_2 = 6.567$
$\tau = -0.067$	
$t[1998] = -1.491$	
$p = 0.136$	

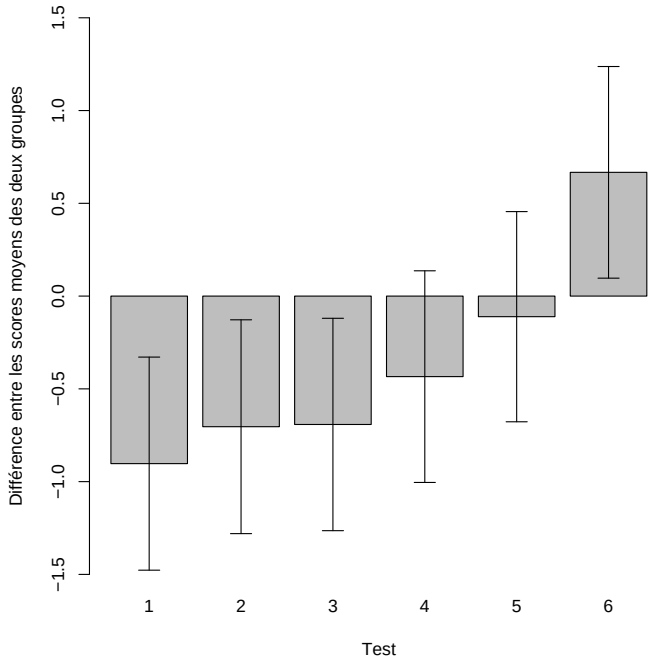
Test 5

Groupe 1	Groupe 2
$\overline{x}_1 = 21.100$	$\overline{x}_2 = 20.989$
$s_1 = 6.399$	$s_2 = 6.539$
$\tau = -0.017$	
$t[1998] = -0.383$	
$p = 0.701$	

Test 6

Groupe 1	Groupe 2
$\overline{x}_1 = 20.688$	$\overline{x}_2 = 21.355$
$s_1 = 6.403$	$s_2 = 6.595$
$\tau = 0.103$	
$t[1998] = 2.293$	
$p = 0.022$	

FIGURE 3 – *Estimation de la différence entre les moyennes des Populations 1 et 2 avec la marge d'erreur à 95%.*



Pour chacun des six tests, nous avons calculé la taille de l'effet τ :

$$\tau = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\hat{\sigma}} \tag{7}$$

avec :

$$\hat{\mu}_1 = \overline{x}_1,\ \hat{\mu}_2 = \overline{x}_2 \text{ et } \hat{\sigma} = \sqrt{\frac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}} \tag{8}$$

la valeur empirique de la variable de décision du test de Student à deux groupes indépendants t :

$$t = \frac{\hat{\mu}_2 - \hat{\mu}_1}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \cdot \hat{\sigma}} \tag{9}$$

et la probabilité critique p :

$$p = 2 \cdot P(|t| < T \mid H_0) \tag{10}$$

avec T une variable aléatoire qui suit une loi de Student à $n_1 + n_2 - 2$ degrés de liberté :

$$T \sim t[n_1 + n_2 - 2]. \tag{11}$$

2.5 Bilan

Nous constatons que dans un test la présence d'items fonctionnant différentiellement compromet fortement la justesse des inférences faites à propos de la position des populations comparées. Les résultats de nos simulations montrent que trois conclusions sont possibles. Première conclusion : la moyenne de la Population 1 est supérieure à celle de la Population 2 ; c'est ce que l'on observe lors de l'utilisation des Tests 1, 2 et 3. Deuxième conclusion : la moyenne de la Population 1 est égale à la moyenne de la Population 2 ; c'est la conclusion tirée suite à l'emploi des Tests 4 et 5. Troisième conclusion : la moyenne de la Population 1 est inférieure à la moyenne de la Population 2 ; c'est ce que l'on conclut en appliquant le Test 6. Comme, par construction, $\mu_1 = \mu_2$, nous devons admettre que, dans la majorité des cas étudiés, les conclusions tirées des observations à l'aide de méthodes classiques sont erronées.
Bien que ces résultats soient déplorables, ils sont compréhensibles. Ils sont, en effet, très liés au biais des tests ; celui-ci est défini comme la somme des différences des difficultés des items appariés :

$$biais = \sum_{j} (\beta_j(2) - \beta_j(1)) \tag{12}$$

ou, plus spécifiquement, dans les situations que nous étudions :

$$biais = \sum_{j=1}^{6} (\beta_j^B(2) - \beta_j^B(1)) + \sum_{j=1}^{6} (\beta_j^C(2) - \beta_j^C(1)) \tag{13}$$

Lorsque le biais est positif, cela signifie qu'en moyenne les items proposés aux individus de la Population 1 sont plus faciles que ceux proposés aux individus de la Population 2 ; il n'est pas surprenant dans ces conditions que la performance des premiers soit supérieure à celle des seconds. Lorsque le biais est négatif, cela veut dire qu'en moyenne les items proposés aux individus de la Population 1 sont plus difficiles que ceux proposés aux individus de la Population 2. Dans ce cas, on s'attend, bien évidemment, à ce que la performance des premiers soit inférieure à celle des seconds (Tableau 7).
Nous pouvons néanmoins tirer un élément positif de ce petit désastre : un test contenant des items fonctionnant différentiellement dans lequel les avantages en faveur d'une population sont compensés par les avantages en faveur de l'autre population – autrement dit, lorsque le biais du test est négligeable (comme dans le Test 5) – permet d'estimer correctement la position des populations. Cette condition n'étant pas toujours remplie, il paraît important de pouvoir disposer d'une méthode plus fiable, permettant d'estimer correctement la position des populations même lorsque le test est biaisé. Une telle méthode serait d'autant plus utile qu'il n'est pas possible de s'assurer avec des outils classiques que le biais d'un test

TABLEAU 7 – *Inférences faites concernant la position de μ_1 et μ_2.*

Test	Biais	Sgn(biais)	Conclusion
1	$(6 \times 0.8) + (6 \times \quad 0.0) = \quad 4.8$	$+1$	$\mu_1 > \mu_2$
2	$(6 \times 0.8) + (6 \times -0.1) = \quad 4.2$	$+1$	$\mu_1 > \mu_2$
3	$(6 \times 0.8) + (6 \times -0.2) = \quad 3.6$	$+1$	$\mu_1 > \mu_2$
4	$(6 \times 0.8) + (6 \times -0.4) = \quad 2.4$	$+1$	$\mu_1 = \mu_2$
5	$(6 \times 0.8) + (6 \times -0.8) = \quad 0.0$	0	$\mu_1 = \mu_2$
6	$(6 \times 0.8) + (6 \times -1.6) = -4.8$	-1	$\mu_1 < \mu_2$

est nul. Dans la section suivante, nous présenterons une procédure qui devrait être plus satisfaisante que celle qui consiste simplement à utiliser les scores globaux d'un test.

3 Détection et correction d'un biais

Nous commencerons par décrire généralement la démarche que nous préconisons puis nous l'appliquerons à trois situations.

3.1 Démarche générale

La méthode que nous proposons se déroule en trois étapes. Lors de la première étape, la difficulté des items est estimée séparément dans chacune des deux populations. Lors de la deuxième, les items fonctionnant différentiellement sont repérés. Lors de la troisième, la difficulté des items est à nouveau estimée mais cette fois les calculs se font sur l'ensemble des individus, issus de la Population 1 ou de la Population 2, et les items qui fonctionnent de manière différentielle sont particularisés.

3.1.1 Première estimation de la difficulté des items

Les compétences des individus et la difficulté des items sont estimées conjointement à l'aide de la méthode du maximum de vraisemblances [3], comme nous l'avons déjà décrit dans un précédent cahier [1]. Ces estimations se font une première fois à partir des performances des individus issus de la Population 1 puis une seconde fois à partir des performances des individus issus de la Population 2. À l'issue de ces calculs, nous obtenons pour chaque individu une estimation de sa compétence et pour chaque item deux estimations de sa difficulté. Nous disposons également des erreurs standards pour chaque estimation.

3.1.2 Identification des items fonctionnant différentiellement

Pour identifier les items qui fonctionnent de manière différentielle nous calculons à partir des deux séries d'estimations des difficultés un *spectre de concordance*. Pour ce faire, nous maintenons fixe la première série des estimations des difficultés et faisons varier la seconde. Pour chaque position de la seconde série, nous calculons, d'une part, le recouvrement de chaque paire d'items puis, d'autre part, la somme des recouvrements. Nous évaluons ainsi dans quelle mesure $\hat{\beta}_1(1)$ concorde avec $\hat{\beta}_1(2)$, $\hat{\beta}_2(1)$ concorde avec $\hat{\beta}_2(2)$, $\hat{\beta}_3(1)$ concorde avec $\hat{\beta}_3(2)$, etc. Pour un décalage d quelconque, le recouvrement $\rho_j(d)$ entre $\hat{\beta}_j(1)$ et $\hat{\beta}_j(2)$ se calcule de la manière suivante :

$$\rho_j(d) = \frac{1}{\sqrt{2\pi}} \sqrt{\frac{1}{\hat{\epsilon}_j^2(1)} + \frac{1}{\hat{\epsilon}_j^2(1)}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\hat{\beta}_j(1)}{\hat{\epsilon}_j(1)}\right)^2} e^{-\frac{1}{2}\left(\frac{x-\hat{\beta}_j(2)-d}{\hat{\epsilon}_j(2)}\right)^2} dx \quad (14)$$

avec :

$\hat{\beta}_j(1)$ l'estimation de la difficulté de l'item j dans la Population 1 ;

$\hat{\epsilon}_j(1)$ l'erreur standard associée à l'estimation de $\beta_j(1)$;

$\hat{\beta}_j(2)$ l'estimation de la difficulté de l'item j dans la Population 2 ;

$\hat{\epsilon}_j(2)$ l'erreur standard associée à l'estimation de $\beta_j(2)$.

Lorsque $\hat{\beta}_j(1) = \hat{\beta}_j(2) + d$, le recouvrement est parfait et vaut 1.
La concordance $\gamma(d)$, quant à elle, est définie comme la somme des recouvrements des items du test :

$$\gamma(d) = \sum_j \rho_j(d). \tag{15}$$

Pour identifier les items qui fonctionnent de la même manière dans les deux populations, il suffit de pointer les items qui se recouvrent lorsque, dans le spectre, la concordance est maximale.
À titre d'exemple, établissons le spectre correspondant à chacun des trois tests suivants :

Test a

Item	1	2	3	4	5	6	7
Population 1	−3	−2	−1	0	1	2	3
Population 2	−3	−2	−1	0	1	2	3

Test b

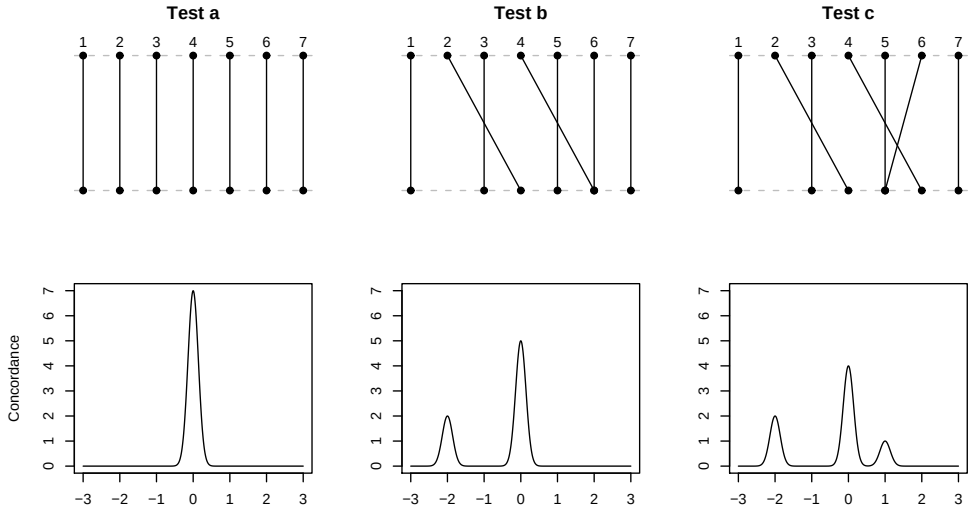
Item	1	2	3	4	5	6	7
Population 1	−3	−2	−1	0	1	2	3
Population 2	−3	0	−1	2	1	2	3

Test c

Item	1	2	3	4	5	6	7
Population 1	−3	−2	−1	0	1	2	3
Population 2	−3	0	−1	2	1	1	3

Nous supposerons que les erreurs standards valent toutes 0.1.
Dans la Figure 4, nous voyons que le spectre de concordance du Test *a* ne possède qu'un pic de hauteur 7. Cela signifie que tous les items se recouvrent, ils se comportent globalement tous de la même manière dans les Populations 1 et 2, ils ne forment qu'une famille. Le spectre du Test *b* possède deux pics, l'un d'une hauteur de 5 et l'autre d'une hauteur de 2. Ainsi cinq items fonctionnent de manière cohérente dans les deux populations et deux items fonctionnent différentiellement. Le décalage entre les deux familles d'items se lit dans le spectre : il est égal à la longueur de l'intervalle qui sépare les deux pics, il vaut en l'occurrence 2. Le spectre du Test *c* contient trois pics, de hauteur 4, 2 et 1 respectivement. Sur les sept items du Test *c*, trois fonctionnent différentiellement : deux fournissent un avantage aux individus de la Population 1 et un fournit un

FIGURE 4 – *Trois tests et leur spectre de concordance.*



avantage aux individus de la Population 2. Ici aussi les distances entre les pics peuvent s'interpréter en termes de biais.
Il est possible de faire une typologie des tests à partir de leur spectre de concordance (Figure 5). Un test qui ne possède aucun item fonctionnant de manière différentielle est caractérisé par un spectre n'ayant qu'un seul pic (Type I). Un test qui contient quelques items fonctionnant différentiellement mais qui n'est pas biaisé a un spectre constitué d'un pic dominant entouré de quelques satellites et si l'on interprète l'ordonnée du spectre de concordance comme une densité associée à chaque point de l'abscisse, alors la position moyenne du spectre coïncide avec la position modale (*i.e.* le pic dominant) (Type II). Un test biaisé, quant à lui, est caractérisé par un spectre dans lequel la moyenne et le mode occupent des positions distinctes (Type III). Finalement, un test totalement disparate se reconnaît à un spectre qui ne possède aucun pic dominant (Type IV).

3.1.3 Nouvelle estimation des paramètres du modèle

L'examen des recouvrements à l'endroit où, dans le spectre, la concordance est maximale permet de séparer les items en deux lots : l'un contenant J_1 items fonctionnant de manière cohérente dans les deux populations et J_2 items fonctionnant différentiellement (Figure 6).
Avant d'estimer les paramètres du modèle, nous devons transformer le tableau de données brutes de telle sorte qu'à l'issue des calculs nous obtenions simultanément :

FIGURE 5 – *Types de spectre de concordance.*

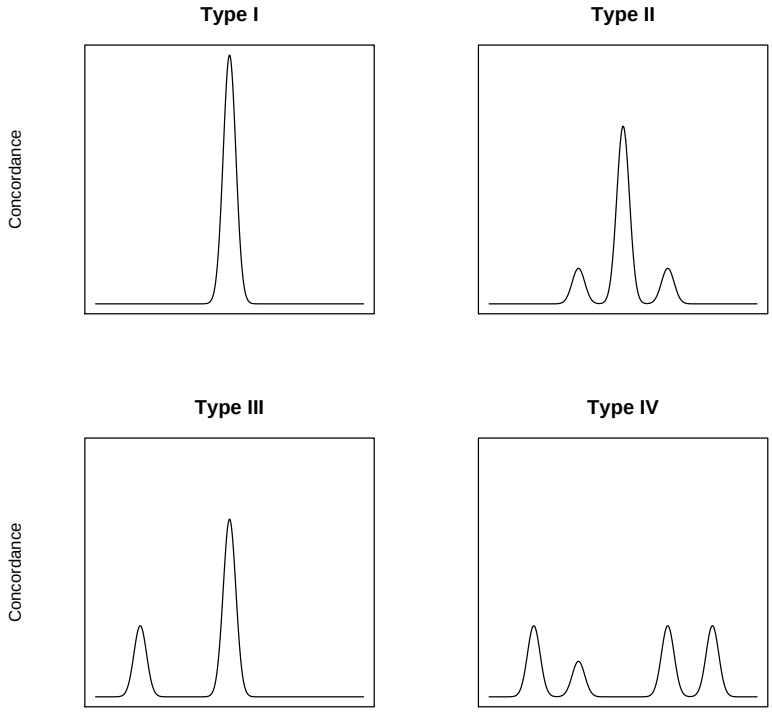


FIGURE 6 – *Structure des données brutes après regroupement des items selon qu'ils fonctionnent ou non différentiellement. x_{ij} représente la réponse de l'individu i à l'item j. Cette réponse est codée par 0 ou 1.*

$$
\begin{array}{cc}
 & \begin{array}{cccccc} 1 & \cdots & j_1 & \cdots & J_1 \end{array} \\
\begin{array}{c} \text{Groupe 1} \\ \\ \\ \\ \\ \end{array} &
\begin{array}{c|ccccc}
1 & x_{11} & \cdots & x_{1j_1} & \cdots & x_{1J_1} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
i_1 & x_{i_1 1} & \cdots & x_{i_1 j_1} & \cdots & x_{i_1 J_1} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
I_1 & x_{I_1 1} & \cdots & x_{I_1 j_1} & \cdots & x_{I_1 J_1}
\end{array}
\end{array}
\qquad
\begin{array}{c}
\begin{array}{cccccc} 1 & \cdots & j_2 & \cdots & J_2 \end{array} \\
\begin{array}{|ccccc|}
x_{11} & \cdots & x_{1j_2} & \cdots & x_{1J_2} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{i_1 1} & \cdots & x_{i_1 j_2} & \cdots & x_{i_1 J_2} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{I_1 1} & \cdots & x_{I_1 j_2} & \cdots & x_{I_1 J_2} \\
\hline
\end{array}
\end{array}
$$

$$
\begin{array}{cc}
\begin{array}{c} \text{Groupe 2} \\ \\ \\ \\ \\ \end{array} &
\begin{array}{c|ccccc}
1 & x_{11} & \cdots & x_{1j_1} & \cdots & x_{1J_1} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
i_2 & x_{i_2 1} & \cdots & x_{i_2 j_1} & \cdots & x_{i_2 J_1} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
I_2 & x_{I_2 1} & \cdots & x_{I_2 j_1} & \cdots & x_{I_2 J_1}
\end{array}
\end{array}
\qquad
\begin{array}{|ccccc|}
x_{11} & \cdots & x_{1j_2} & \cdots & x_{1J_2} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{i_2 1} & \cdots & x_{i_2 j_2} & \cdots & x_{i_2 J_2} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
x_{I_2 1} & \cdots & x_{I_2 j_2} & \cdots & x_{I_2 J_2} \\
\hline
\end{array}
$$

– la compétence des individus :

$$\{\theta_1,\ \theta_2,\ \ldots,\ \theta_{i_1},\ \ldots,\ \theta_{I_1}\} \text{ et } \{\theta_1,\ \theta_2,\ \ldots,\ \theta_{i_2},\ \ldots,\ \theta_{I_2}\}\,;$$

– la difficulté des items se comportant de manière cohérente dans les deux populations :

$$\{\beta_1,\ \beta_2,\ \ldots,\ \beta_{j_1},\ \ldots,\ \beta_{J_1}\}\,;$$

– la difficulté des items fonctionnant différentiellement tels qu'ils apparaissent aux individus de la Population 1 :

$$\{\beta_1(1),\ \beta_2(1),\ \ldots,\ \beta_{j_2}(1),\ \ldots,\ \beta_{J_2}(1)\}\,;$$

– la difficulté des items fonctionnant différentiellement tels qu'ils apparaissent aux individus de la Population 2 :

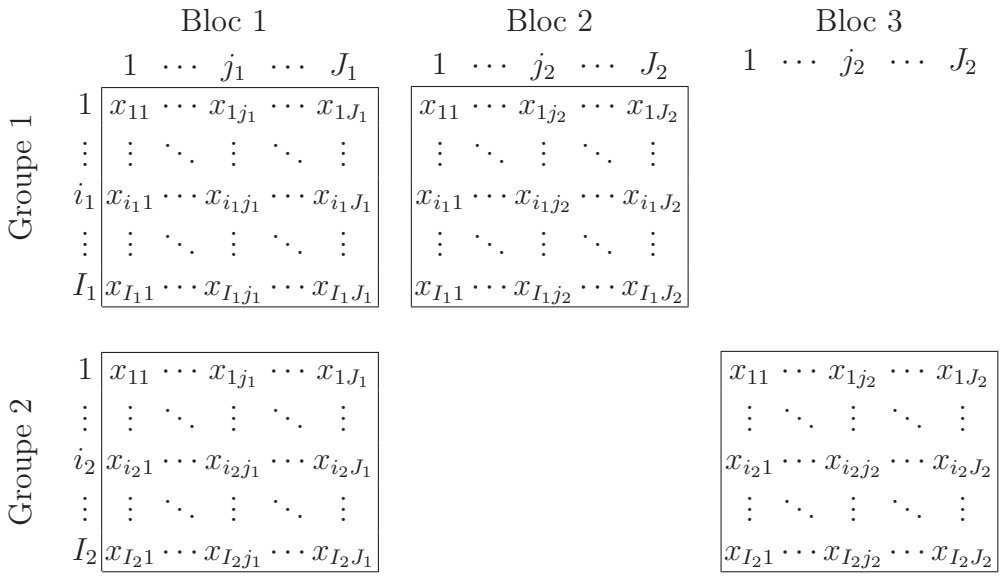
$$\{\beta_1(2),\ \beta_2(2),\ \ldots,\ \beta_{j_2}(2),\ \ldots,\ \beta_{J_2}(2)\}\,;$$

Le nouveau tableau de données est formé de trois blocs verticaux juxtaposés. Chaque ligne représente un individu interrogé. Au sein des trois blocs, les individus sont placés dans le même ordre : en tête se trouvent les individus issus de la Population 1, en queue ceux issus de la Population 2. Le premier bloc contient les réponses fournies aux items fonctionnant de manière cohérente dans les deux populations. Le deuxième bloc est formé des réponses des individus issus de la Population 1 aux items qui fonctionnent différentiellement ; dans les lignes réservées aux individus issus de la Population 2, il y a des données manquantes. Et le troisième bloc est formé des réponses des individus issus de la Population 2 aux items qui fonctionnent différentiellement ; de manière analogue à ce que l'on trouve dans le bloc 2, les lignes réservées aux individus issus de la Population 1 sont vides (Figure 7).
En procédant de cette manière, nous améliorons la précision des estimations des difficultés des items qui fonctionnent de façon cohérente dans les deux populations et pouvons placer simultanément sur un seul et même trait latent tous les paramètres du modèle.

3.2 Illustrations

Nous allons maintenant appliquer la démarche que nous venons de décrire (§ 3.1) à trois situations différentes. Dans la première, le test utilisé sera biaisé positivement et les populations comparées seront les mêmes. Dans la deuxième, le test sera biaisé négativement et les populations seront à nouveau identiques. Dans la troisième, le test sera biaisé positivement et les populations n'occuperont pas la même position.
Dans chaque situation, nous comparerons les résultats obtenus selon

FIGURE 7 – *Structure des données à partir desquelles les paramètres du modèle sont estimés.*



notre méthode à ceux que l'on obtient classiquement sans recourir au modèle de la réponse à l'item.

3.2.1 Exemple 1

Nous utiliserons le Test 1 décrit dans la sous-section 2.2. Nous interrogerons deux échantillons de 1000 individus issus d'une même population normale centrée et réduite. Ainsi, en réalité :

$$\boxed{\mu_1 = \mu_2}$$

Procédure classique Les performances globales des deux groupes sont résumées dans le Tableau 8.
À partir des caractéristiques des échantillons, il est possible d'affirmer, au seuil de 5%, que les populations sont différentes :

$$\boxed{\mu_1 > \mu_2}$$

Cette conclusion, erronée, est la même que celle que nous avions obtenue précédemment en traitant une situation analogue (§ 2.4).

Nouvelle procédure Commençons par calculer la difficulté des items séparément dans chacun des deux groupes (§ 3.1.1) puis, à partir de

TABLEAU 8 – *Performances réalisées par les Groupes 1 et 2 dans l'Exemple 1.*

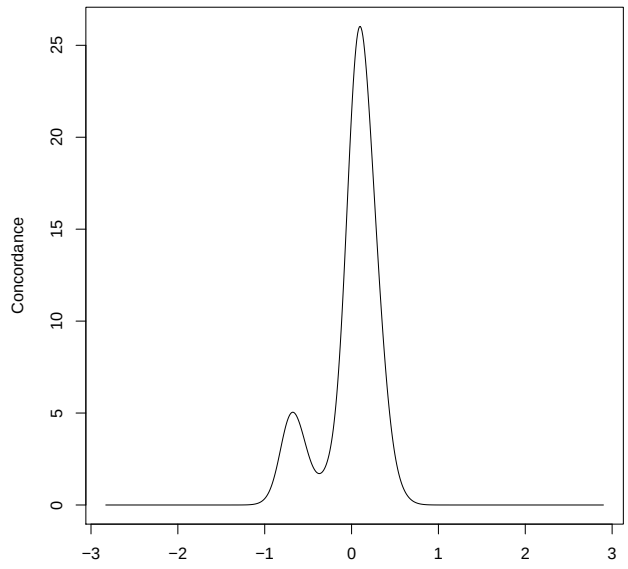
Test 1

Groupe 1	Groupe 2
$\overline{x}_1 = 21.413$	$\overline{x}_2 = 20.782$
$s_1 = 6.698$	$s_2 = 6.613$

τ	$=$	-0.095
$t[1998]$	$=$	-2.119
p	$=$	0.034

ces premiers résultats intermédiaires, établissons le spectre de concordance (§ 3.1.2). Il s'avère que le Test 1 contient quelques items fonctionnant différentiellement et est biaisé. En effet le spectre possède un pic dominant et le mode du spectre ne coïncide pas avec sa moyenne (Figure 8).

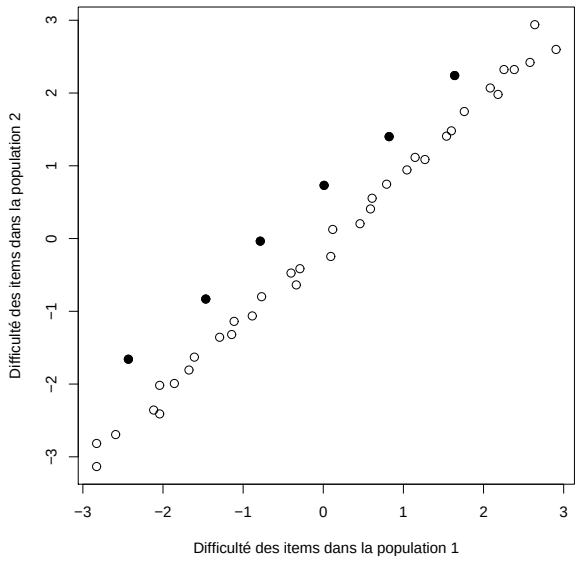
FIGURE 8 – *Spectre de concordance établi à partir des données de l'Exemple 1.*



En appliquant la méthode d'agrégation autour de deux centre mobiles [2] aux logarithmes des recouvrements déterminés à l'endroit où, dans le spectre, la concordance est maximale, nous identifions deux groupes d'items : le premier formé de 36 items qui fonctionnent de manière cohérente dans les deux populations et le second formé de 6 items qui fonctionnent différentiellement.

Dans la Figure 9, nous avons représenté la difficulté des items estimée dans la Population 2 en fonction de la difficulté des items estimée dans la Population 1. Les items fonctionnant différentiellement sont représentés par une pastille noire, les autres par une pastille blanche. Ce graphique fournit, à l'imprécision des estimations près, une image fidèle de la structure du Test 1.

FIGURE 9 – *Lien entre la difficulté des items perçue par les individus de la Population 1 et la difficulté des mêmes items perçue par les individus de la Population 2 dans l'Exemple 1. Les items fonctionnant différentiellement sont en noir.*



Estimons les paramètres du modèle en tenant compte cette fois de la spécificité des items comme nous l'avons décrit précédemment (§ 3.1.3) et comparons la position des moyennes des Populations 1 et 2 en nous appuyant sur les estimations des compétences des individus des Groupes 1 et 2 (Tableau 9).

TABLEAU 9 – *Compétences des Groupes 1 et 2 dans l'Exemple 1.*

Test 1

Groupe 1	Groupe 2
$\overline{x}_1 = 0.010$	$\overline{x}_2 = 0.036$
$s_1 = 1.125$	$s_2 = 1.104$

τ	=	0.024
$t[1998]$	=	0.529
p	=	0.597

À l'issue de cette procédure, nous ne sommes pas en mesure de rejeter, au seuil de 5%, l'hypothèse H_0 selon laquelle les moyennes des Populations 1 et 2 sont égales. Nous conclurons donc que :

$$\boxed{\mu_1 = \mu_2}$$

3.2.2 Exemple 2

Nous utiliserons pour cet exemple le Test 6 décrit dans la sous-section 2.2. Nous interrogerons deux échantillons de 1000 individus chacun issus d'une même population normale centrée et réduite. Ainsi, par construction :

$$\boxed{\mu_1 = \mu_2}$$

Pour ne pas trop nous répéter, ni abuser de la patience de notre lecteur, nous ne fournirons dans cet exemple et le suivant que les figures et tableaux qui parlent presque d'eux-mêmes !

Procédure classique À partir des caractéristiques des échantillons (Tableau 10), il est possible d'affirmer, au seuil de 5%, que les populations sont différentes :

$$\boxed{\mu_1 < \mu_2}$$

TABLEAU 10 – *Performances réalisées par les Groupes 1 et 2 dans l'Exemple 2.*

Test 6

Groupe 1	Groupe 2
$\overline{x}_1 = 20.351$	$\overline{x}_2 = 21.495$
$s_1 = 6.595$	$s_2 = 6.450$
$\tau = 0.175$	
$t[1998] = 3.919$	
$p = 0.000$	

Nouvelle procédure À l'issue de cette procédure (Figure 10 et Tableau 11), nous ne sommes pas en mesure de rejeter, au seuil de 5%, l'hypothèse H_0 selon laquelle les moyennes des Populations 1 et 2 sont égales. Sommairement, nous conclurons que :

$$\boxed{\mu_1 = \mu_2}$$

FIGURE 10 – À gauche : *Spectre de concordance établi à partir des données de l'Exemple 2.* À droite : *Lien entre la difficulté des items perçue par les individus de la Population 1 et la difficulté des mêmes items perçue par les individus de la Population 2 dans l'Exemple 2. Les items fonctionnant différentiellement sont en noir.*

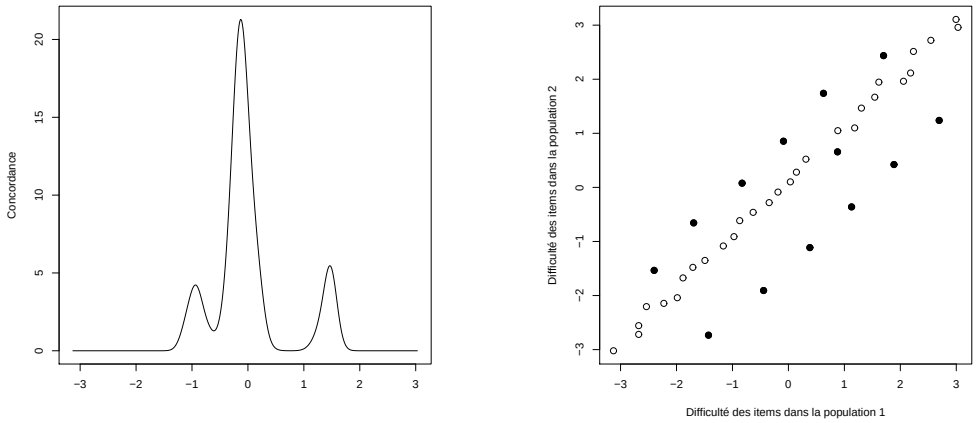


TABLEAU 11 – *Compétences des Groupes 1 et 2 dans l'Exemple 2.*

Test 6

Groupe 1	Groupe 2
$\overline{x}_1 = -0.068$	$\overline{x}_2 = -0.002$
$s_1 = 1.120$	$s_2 = 1.091$

τ	=	0.059
$t[1998]$	=	1.321
p	=	0.187

3.2.3 Exemple 3

Nous allons reprendre le test utilisé dans l'Exemple 1 (Test 1 du § 2.2) mais cette fois les Populations 1 et 2 seront différentes, elles ne posséderont pas la même moyenne :

$$\text{Population 1 :} \qquad \theta_i(1) \sim \mathcal{N}(\mu_1 = -0.1,\ \sigma^2 = 1) \tag{16}$$

$$\text{Population 2 :} \qquad \theta_i(2) \sim \mathcal{N}(\mu_2 = +0.1,\ \sigma^2 = 1) \tag{17}$$

Par construction donc :

$$\boxed{\mu_1 < \mu_2}$$

Les échantillons interrogés sont tous les deux de taille 1000.

Procédure classique La différence entre les moyennes des deux populations n'est pas statistiquement significative au seuil de 5% (Tableau 12) :

$$\boxed{\mu_1 = \mu_2}$$

TABLEAU 12 – *Performances réalisées par les Groupes 1 et 2 dans l'Exemple 3.*

Test 1

Groupe 1	Groupe 2
$\overline{x}_1 = 20.729$	$\overline{x}_2 = 21.080$
$s_1 = 6.766$	$s_2 = 6.774$

τ	=	0.052
$t[1998]$	=	1.159
p	=	0.247

Nouvelle procédure À partir des caractéristiques des échantillons (Figure 11 et Tableau 13), il est possible d'inférer, au seuil de 5%, la conclusion suivante :

$$\boxed{\mu_1 < \mu_2}$$

Par ailleurs, l'estimation des positions des moyennes est excellente.

FIGURE 11 – À gauche : *Spectre de concordance établi à partir des données de l'Exemple 3.* À droite : *Lien entre la difficulté des items perçue par les individus de la Population 1 et la difficulté des mêmes items perçue par les individus de la Population 2 dans l'Exemple 3. Les items fonctionnant différentiellement sont en noir.*

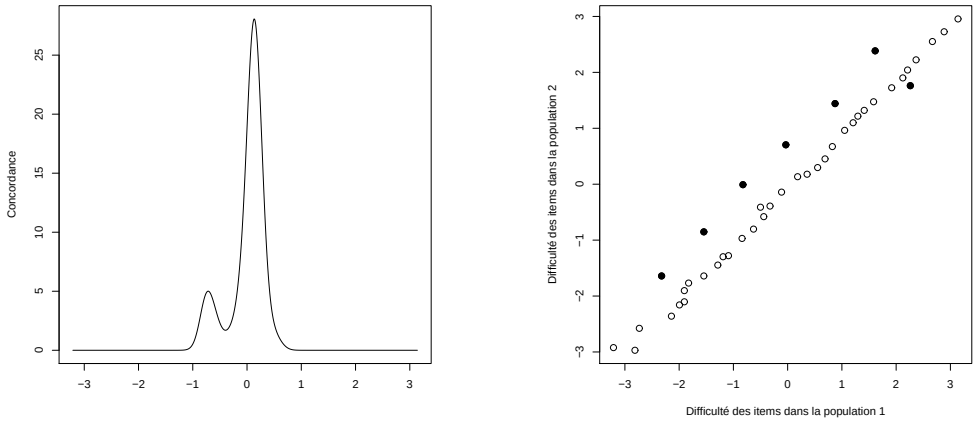


TABLEAU 13 – *Compétences des Groupes 1 et 2 dans l'Exemple 3.*

Test 1

Groupe 1	Groupe 2
$\overline{x}_1 = -0.150$	$\overline{x}_2 = 0.034$
$s_1 = 1.146$	$s_2 = 1.228$
$\tau = 0.162$	
$t[1998] = 3.617$	
$p = 0.000$	

4 Conclusion

Dans le Tableau 14 nous avons rassemblé toutes les conclusions tirées de l'analyse des Exemples 1, 2 et 3 (§ 3.2).

TABLEAU 14 – *Comparaison des conclusions tirées lors de l'application de deux procédures : l'une classique et l'autre nouvelle.*

	Position	Procédure	
Exemple	des moyennes	classique	nouvelle
1	$\mu_1 = \mu_2$	$\mu_1 > \mu_2$	$\mu_1 = \mu_2$
2	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$	$\mu_1 = \mu_2$
3	$\mu_1 < \mu_2$	$\mu_1 = \mu_2$	$\mu_1 < \mu_2$

Nous constatons de manière indéniable, la supériorité de notre méthode. Par curiosité et pour compléter cette étude, nous avons aussi utilisé ConQuest [5] pour analyser les données des Exemples 1, 2 et 3. Ce logiciel, très largement utilisé dans la communauté scientifique[1], capable – selon les concepteurs – de détecter les items fonctionnant différentiellement et pallier leur effet conduit à deux conclusions erronées sur trois (la première et la deuxième).
Refrénons tout de même notre enthousiasme en rappelant qu'actuellement notre méthode ne permet de comparer que deux populations à la fois et que le test utilisé ne doit contenir que des items dichotomiques qui fonctionnent dans chaque population conformément aux hypothèses du modèle de Rasch.

[1] ConQuest a été employé, entre autres, pour analyser les données collectées dans le cadre des enquêtes PISA et HarmoS.

Bibliographie

[1] J.-Ph. Antonietti, *Mesures objectives de traits latents*, Cahiers de l'IMA, 39 (2006).

[2] L. Lebart, A. Morineau, & M. Piron, *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 1997.

[3] I. W. Molenaar, *Estimation of item parameters*, in Rasch models : Foundations, recent developments, and applications, G. H. Fischer & I. W. Molenaar, eds., Springer, New York, 1995, pp. 39–51.

[4] G. Rasch, *Probabilistic models for some intelligence and attainment tests*, The University of Chicago Press, Chicago, 1980.

[5] M. L. Wu, R. J. Adams, M. R. Wilson, & S. A. Haldane, *ConQuest Version 2.0 : Generalised Item Response Modelling Software*, ACER, Camberwell, 2007.