Novembre 2006 **Numéro 39**

Cahiers de l'IMA

Mesures objectives de traits latents

Jean-Philippe Antonietti

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

Mesures objectives de traits latents

Jean-Philippe Antonietti

Résumé

Nous commençons par montrer, dans cet opuscule, que les mesures faites classiquement en psychologie sont relatives, qu'elles dépendent parfois fortement de l'instrument utilisé. Puis nous présentons sommairement, en nous appuyant sur un exemple, le moyen d'éliminer ce défaut. La méthode que nous proposons à cette fin se fonde sur l'utilisation d'un modèle de la réponse à l'item. Finalement nous décrivons méticuleusement comment construire pratiquement un instrument capable de fournir des mesures objectives.

1 Relativité des mesures classiques

Comment un psychologue mesure-t-il, par exemple, les compétences spatiales d'un sujet ? Généralement, il recourra à un test, ou autrement dit à une épreuve standardisée dans son administration et sa cotation. Il pourra ainsi situer la personne évaluée parmi les membres de la population de référence du test. Il saura précisément quelle proportion de la population de référence possède de moins bonnes compétences spatiales que le sujet testé. Une telle mesure est relative, elle dépend entièrement de la population de référence. Un même sujet pourrait paraître excellent dans une population possédant de piètres compétences spatiales et ne paraître que moyen dans une population beaucoup plus forte.

Ce genre de mesure souffre d'un autre défaut majeure : la longueur de l'intervalle qui sépare la position de deux sujets est aussi relative. La différence de compétences entre deux sujets change selon la population de référence.

Afin d'être plus explicite, illustrons notre propos par quelques simulations. Nous allons construire trois tests. Nous supposerons que chacun de ces tests mesure le même attribut, le même trait psychologique ou la même compétence. La compétence d'un sujet quelconque, θ_i, prendra sa valeur dans l'ensemble des nombres réels $\mathbb{R}$. Les items, que nous supposerons tous dichotomiques, seront caractérisés par un unique paramètre : leur difficulté. La difficulté d'un item j quelconque prendra également sa valeur dans l'ensemble des nombres réels.

La réponse du sujet i à l'item j dépendra de la différence entre la compétence du sujet θ_i et la difficulté de l'item β_j. Si l'item j est difficile comparativement à la compétence du sujet i, il y a peu de chance pour que le sujet i fournisse une réponse correcte à cet item :

$$\text{Si } \theta_i - \beta_j \ll 0, \text{ alors } P(X_{ij} = 1) \simeq 0. \tag{1}$$

Dans le cas contraire, si l'item j est facile pour le sujet i, il y a toutes les chances que ce dernier fournisse la bonne réponse :

$$\text{Si } \theta_i - \beta_j \gg 0, \text{ alors } P(X_{ij} = 1) \simeq 1. \tag{2}$$

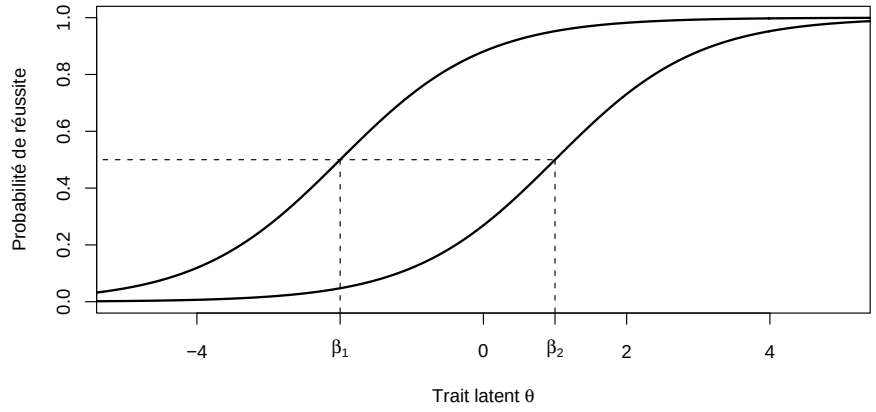
Pour traduire plus formellement ces liens entre compétence, difficulté et probabilité de bien répondre, nous utiliserons une fonction logistique :

$$P(X_{ij} = 1|\theta_i,\ \beta_j) = \frac{e^{(\theta_i - \beta_j)}}{e^{(\theta_i - \beta_j)} + 1}. \tag{3}$$

Dans la figure 1, nous avons représenté la fonction caractéristique de deux items. Nous constatons que lorsque la compétence d'un sujet θ_i est égale à la difficulté de l'item β_j, la probabilité de réussite est égale à 0.5, en effet :

$$\text{Si } \theta_i = \beta_j, \text{ alors } P(X_{ij} = 1|\theta_i,\ \beta_j) = \frac{e^0}{e^0 + 1} = \frac{1}{1+1} = \frac{1}{2}. \tag{4}$$

FIGURE 1 – *Fonction caractéristique de deux items. L'item 1 est moins difficile que l'item 2* ($\beta_1 < \beta_2$).



1.1 Tests utilisés

La population de référence visée par le premier test est une population relativement faible. Les compétences des individus de cette population se distribuent selon une loi normale de moyenne -2 et de variance 1. Le test est composé de 200 items tirés aléatoirement d'une population d'items ayant les mêmes caractéristiques que la population de référence (difficulté moyenne : -2 ; variance des difficultés : 1).
La population de référence visée par le deuxième test est une population moyenne. Les compétences des individus de cette population se distribuent selon une loi normale de moyenne égale à 0 et de variance égale à 1. Ce deuxième test est composé, comme

le premier, de 200 items de difficulté similaire aux compétences de la population de référence.
Le troisième test, construit selon le même canevas que les deux premiers, permet d'évaluer au mieux une population forte. Dans cette population, la moyenne des compétences vaut 2 et la variance 1 (tableau 1).

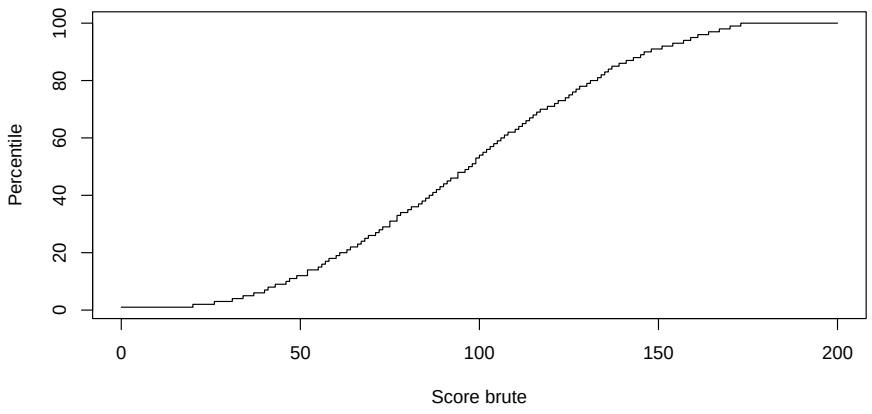
TABLEAU 1 – *Caractéristiques des trois tests utilisés dans la simulation.*

	Population de référence	Difficulté des items	Nombre d'items
Test 1	$\theta_i \sim \mathcal{N}(-2,\ 1)$	$\beta_j \sim \mathcal{N}(-2,\ 1)$	200
Test 2	$\theta_i \sim \mathcal{N}(\ \ 0,\ 1)$	$\beta_j \sim \mathcal{N}(\ \ 0,\ 1)$	200
Test 3	$\theta_i \sim \mathcal{N}(\ \ 2,\ 1)$	$\beta_j \sim \mathcal{N}(\ \ 2,\ 1)$	200

1.2 Établissement des normes

Pour chaque test, nous établissons des normes. Ces dernières sont élaborées à partir d'un échantillon de 2000 sujets tirés aléatoirement de chacune des populations de référence. Les normes expriment le lien entre le score brut au test – ce score est égale au nombre d'items résolus correctement – et le rang au sein de la population. Pour exprimer les positions, nous avons opté pour un centilage. La distribution des scores bruts est donc ramenée à 100 échelons contenant chacun 1% des sujets de l'échantillon. La figure 2 représente le barème du premier test.

FIGURE 2 – *Représentation graphique de la relation entre scores bruts au premier test et rangs, exprimés en percentiles, occupés dans la première population de référence.*



1.3 Groupes comparés

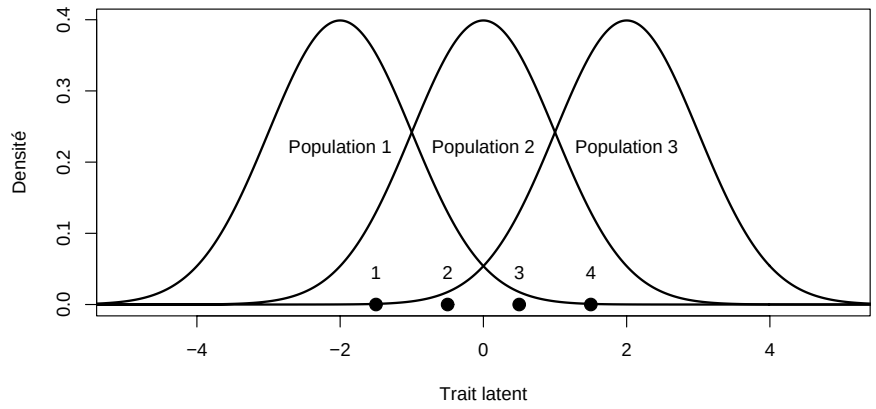
Nous venons de créer trois tests. Nous allons évaluer et comparer avec chacun de ces trois instruments quatre groupes homogènes. Le premier groupe est constitué de 200 individus ayant tous une compétence de -1.5. Le deuxième groupe est composé de 200 individus ayant chacun une compétence égale à -0.5. Les troisième et quatrième groupes ont aussi un effectif de 200 mais sont formés d'individus ayant une compétence égale à $+0.5$ et $+1.5$ respectivement (tableau 2).

TABLEAU 2 – *Caractéristiques des groupes comparés.*

	Effectif	Compétence
Groupe 1	200	-1.5
Groupe 2	200	-0.5
Groupe 3	200	$+0.5$
Groupe 4	200	$+1.5$

La figure 3 rassemble dans un graphique tous les ingrédients que nous allons utiliser pour montrer la relativité des mesures faites avec des instruments construits selon une procédure classique fondée, rappelons-le, sur l'établissement d'un score total obtenu simplement par dénombrement des réponses correctes.

FIGURE 3 – *Distribution des populations de référence évaluées par chacun des trois tests et position des quatre groupes homogènes* (•).



1.4 Résultats de l'évaluation

1.4.1 Premier test

Soumettons les individus des groupes 1, 2, 3 et 4 au premier test. Transformons leurs scores bruts en percentiles à l'aide des normes que nous avons précédemment établies (§ 1.2). Leurs scores sont représentés dans la figure 4 (à gauche). Globalement les individus testés obtiennent de hauts scores. Quasiment tous décrochent un score supérieur à 50 ; ce qui signifie que leurs compétences sont supérieures à celles de plus de la moitié de la population de référence. Notons que théoriquement la proportion des individus de la population 1 moins compétents que les individus du groupe 1 est égale à la probabilité qu'une distribution normale de moyenne -2 et de variance 1 prenne une valeur inférieure à -1.5. Cette probabilité vaut 0.691 :

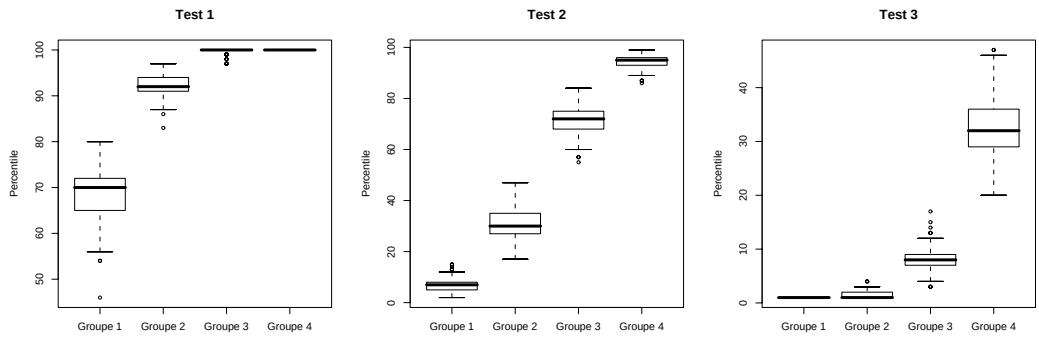
$$\text{Si } X_1 \sim \mathcal{N}(-2,\ 1), \text{ alors } P(X_1 \leq -1.5) = P(X^* \leq 0.5) = \Phi(0.5) = 0.691 \qquad (5)$$

avec X^* une distribution normale centrée et réduite et $\Phi(x)$ la valeur de la fonction de répartition de cette distribution en x.
De manière analogue, nous pouvons calculer la proportion des individus de la première population de référence inférieurs aux individus des groupes 2, 3 et 4 :

$$P(X_1 \leq -0.5) = \Phi(1.5) = 0.933 \qquad (6)$$
$$P(X_1 \leq 0.5) = \Phi(2.5) = 0.994 \qquad (7)$$
$$P(X_1 \leq 1.5) = \Phi(3.5) = 1.000. \qquad (8)$$

Avec le test 1, nous constatons qu'il est possible de distinguer le groupe 1 du groupe 2 et, dans une moindre mesure, le groupe 2 du groupe 3. Par contre les groupes 3 et 4 sont indiscernables. Les individus de ces deux groupes résolvent correctement tous les items du test 1 qui sont pour eux manifestement trop faciles.

FIGURE 4 – *Distribution des scores obtenus par les individus des groupes 1, 2, 3 et 4 soumis aux tests 1, 2 et 3 respectivement.*



1.4.2 Deuxième test

Soumettons maintenant les individus des groupes 1, 2, 3 et 4 au test 2. Leurs résultats, exprimés en percentiles, sont représentés dans la figure 4 (au centre). Les performances des quatre groupes se répartissent régulièrement le long de l'échelle qui va de 1 à 100. Théoriquement la proportion des individus de la deuxième population de référence ayant des compétences inférieures à celles du groupe 1 devrait valoir $\Phi(-1.5) = 0.067$; cette proportion devrait être égale à $\Phi(-0.5) = 0.309$ pour le groupe 2 et s'élever à $\Phi(0.5) = 0.691$ et $\Phi(1.5) = 0.933$ pour les groupes 3 et 4 respectivement. Les distributions empiriques se positionnent exactement selon nos attentes (tableau 3).

TABLEAU 3 – *Position moyenne des groupes évalués à l'aide de trois tests différents.*

	Test 1		Test 2		Test 3	
	Score moyen		Score moyen		Score moyen	
	empirique	théorique	empirique	théorique	empirique	théorique
Groupe 1	68.5	70	7.1	7	1.0	1
Groupe 2	92.4	94	30.8	31	1.6	2
Groupe 3	99.8	100	71.4	70	8.2	7
Groupe 4	100.0	100	94.6	94	32.4	31

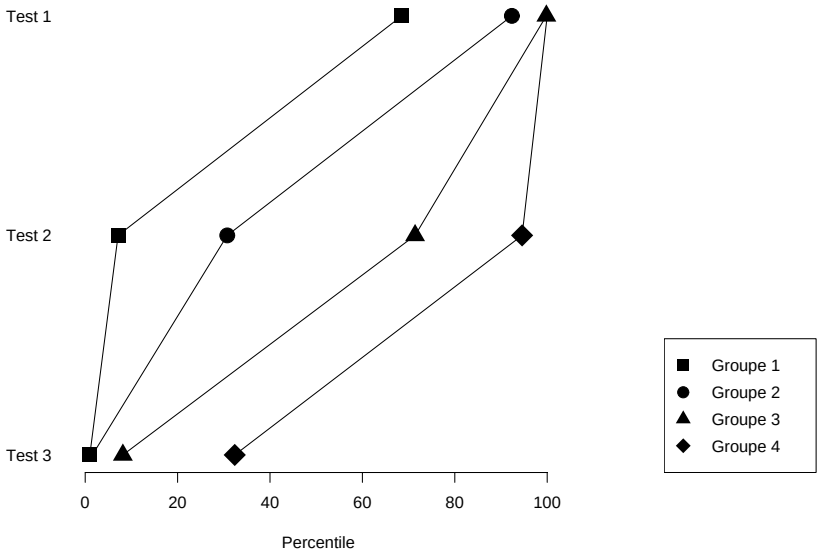
1.4.3 Troisième test

Alors que, dans l'ensemble, le test 1 était trop facile pour les groupes que nous comparons, le test 3 est trop difficile. Avec cet instrument, il n'est plus possible de distinguer les deux groupes les plus faibles car pour les individus de ces deux groupes tous les items sont trop difficiles et ils n'en résolvent correctement aucun (figure 4, à droite).

1.5 Bilan

La figure 5 illustre très clairement les limites de l'approche classique de la mesure en psychologie. Nous voyons à quel point les mesures sont relatives. En changeant d'instrument non seulement on modifie la position des groupes mais aussi les écarts qui les séparent. Il nous faut néanmoins reconnaître que les failles de l'approche classique que nous révélons ici auraient pu être estompées. Les compétences se distribuant normalement au sein des populations de référence, il aurait été préférable d'exprimer les normes à l'aide d'une échelle en scores standard normalisés. Avouons que notre but n'était pas de rendre l'approche classique acceptable mais, bien au contraire, d'en marquer le contour en rendant ses limites plus criardes.

FIGURE 5 – *Évaluation classique de quatre groupes à l'aide de trois tests différents.*

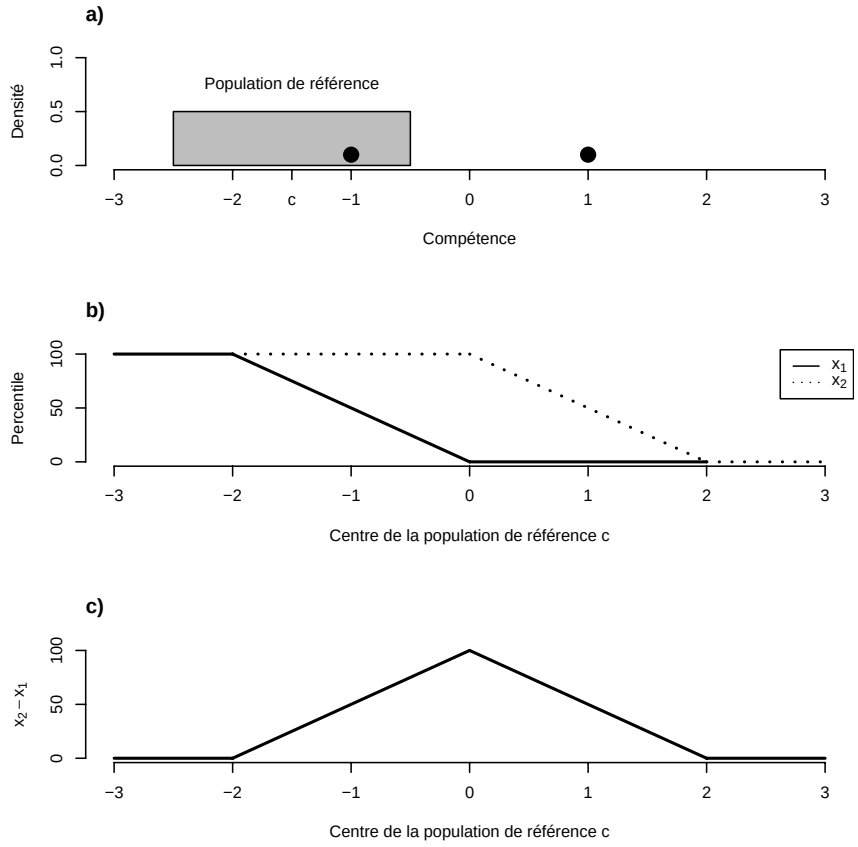


L'exemple qui suit, bien que totalement irréaliste, permet de mieux comprendre pourquoi les mesures psychologiques classiques sont relatives. Supposons que les compétences des individus de la population de référence se distribuent uniformément sur l'intervalle $[c-1;\ c+1]$. Voyons comment évoluent les performances de deux individus témoins – l'un ayant des compétences de -1 et l'autre des compétences de 1 – lorsque le centre c de la population de référence utilisée pour étalonner le test varie de -3 à $+3$. Nommons x_1 le score, exprimé en percentile, de l'individu ayant les compétences de -1 et x_2 le score de l'individu ayant les compétences de 1 :

$$x_1 = \begin{cases} 100 & \text{si} \quad -3 < c \leq -2 \\ -50c & \text{si} \quad -2 < c \leq 0 \\ 0 & \text{si} \quad 0 < c \leq 3 \end{cases} \qquad \text{et} \qquad x_2 = \begin{cases} 100 & \text{si} \quad -3 < c \leq 0 \\ -50c & \text{si} \quad 0 < c \leq 2 \\ 0 & \text{si} \quad 2 < c \leq 3. \end{cases} \tag{9}$$

Lorsque la population de référence change, nous constatons que non seulement x_1 et x_2 varient mais que leur différence $x_2 - x_1$ aussi (figure 6).

FIGURE 6 – *Situation extravagante permettant de mettre en évidence la relativité des mesures classiques. a) Position de la population de référence et des deux individus témoins. b) Scores des deux individus témoins selon la position de la population de référence. c) Écart entre les scores des deux individus témoins lorsque la position de la population de référence varie.*



2 Mesures objectives

Nous venons de montrer que les mesures faites classiquement en psychologie sont relatives. L'évaluation des capacités d'une personne dépend toujours de l'instrument utilisé. Il serait souhaitable de pouvoir disposer d'une gamme d'instruments fournissant tous les mêmes mesures, afin que l'évaluation des compétences d'un individu ne dépende plus du test employé. De tels instruments existent. Ils ont été conçu par Rasch [12] aux environs des années 1960.
Dans certaines conditions bien précises, il est possible de construire des instruments de mesure qui fournissent des résultats objectifs. Ces instruments permettent donc d'évaluer les compétences des individus indépendamment du test utilisé ; ils permettent également de mesurer la difficulté des items indépendamment des personnes à qui l'on fait passer le test.

2.1 Postulats

Pour que l'on puisse véritablement réaliser une mesure objective, deux postulats doivent être satisfaits : l'unidimensionnalité, d'une part, et l'indépendance locale, d'autre part.
L'exigence d'*unidimensionnalité* signifie que tous les items d'un test doivent mesurer un seul et même trait.
L'exigence d'*indépendance locale* signifie, quant à elle, que le trait qui fait l'objet de l'évaluation doit être le seul facteur qui détermine la variabilité des réponses aux items d'un test. Une fois que le trait mesuré a été pris en compte, aucune relation ne doit exister entre les réponses d'un individu aux différents items.
Soit $\mathbf{X} = (X_1,\ X_2,\ \ldots,\ X_k)$ le vecteur des variables des scores des items et $\mathbf{x} = (x_1,\ x_2,\ \ldots,\ x_j,\ \ldots,\ x_k)$ l'une des réalisations de $\mathbf{X}$. Les items j étant dichotomiques, les x_j prennent comme valeur 0 ou 1. La probabilité qu'un individu ayant un niveau de compétences θ_i obtienne le score x_j à l'item j vaut $P(X_j = x_j|\theta_i,\ \beta_j)$. L'indépendance locale se traduit formellement par l'égalité suivante :

$$P(\mathbf{X} = \mathbf{x}|\theta_i,\ \beta_j) = \prod_{j=1}^{k} P(X_j = x_j|\theta_i,\ \beta_j). \tag{10}$$

Par ailleurs, la probabilité qu'un individu ayant la capacité θ_i réponde correctement à l'item de difficulté β_j doit satisfaire la relation suivante :

$$P(X_j = x_j|\theta_i,\ \beta_j) = \frac{e^{x_j(\theta_i - \beta_j)}}{1 + e^{x_j(\theta_i - \beta_j)}}. \tag{11}$$

Avant de détailler précisément la démarche à mettre en œuvre pour construire un instrument qui fournisse des mesures objectives (§ 3), nous aimerions exposer les résultats que l'on obtient lorsqu'on applique cette « nouvelle » méthode aux données que nous avons précédemment traitées de manière tout à fait classique.
Rappelons que nous avions construit trois tests de difficulté croissante et que nous avions évalué avec chacun de ces tests quatre groupes homogènes.

2.2 Estimation de la difficulté des items

Les données que nous avions générées satisfaisaient les postulats d'unidimensionnalité et d'indépendance locale et la fonction caractéristique de chaque item était bien une fonction logistique qui ne dépendait que de la compétence θ_i des sujets et de la difficulté β_j de l'item :

$$P(X_{ij} = 1|\theta_i,\ \beta_j) = \frac{e^{(\theta_i - \beta_j)}}{1 + e^{(\theta_i - \beta_j)}}. \tag{12}$$

Les trois échantillons de 2000 individus qui nous permirent d'établir les normes des tests 1, 2, 3 et 4 (§ 1.2) peuvent être utilisés pour estimer, à l'aide de la méthode du maximum de vraisemblance conjointe [10], la difficulté des items des trois tests :

$$\hat{\beta}(1) = \left(\hat{\beta}_1(1),\ \hat{\beta}_2(1),\ \ldots,\ \hat{\beta}_k(1)\right) \tag{13}$$

$$\hat{\beta}(2) = \left(\hat{\beta}_1(2),\ \hat{\beta}_2(2),\ \ldots,\ \hat{\beta}_k(2)\right) \tag{14}$$

$$\hat{\beta}(3) = \left(\hat{\beta}_1(3),\ \hat{\beta}_2(3),\ \ldots,\ \hat{\beta}_k(3)\right). \tag{15}$$

2.3 Estimation des compétences des individus

À partir de l'estimation de la difficulté des items de chaque test, il est possible d'estimer la compétence des individus des quatre groupes homogènes. Ici aussi, nous recourrons à la méthode du maximum de vraisemblance [8].
Par commodité, nommons P_{jl}, la probabilité qu'un individu ayant la capacité θ réponde correctement à la question j du test l :

$$P_{jl} = P\left(X_{jl} = x_{jl}|\theta,\ \beta_j(l)\right) = \frac{e^{x_{jl}(\theta - \beta_j(l))}}{1 + e^{x_{jl}(\theta - \beta_j(l))}}. \tag{16}$$

En vertu de l'hypothèse d'indépendance locale, la probabilité que cette personne réponde $\mathbf{x} = (x_{1l},\ x_{2l},\ \ldots,\ x_{kl})$ au test l égale :

$$P\left(\mathbf{X} = \mathbf{x}|\theta,\ \beta(l)\right) = \prod_j P_{jl}^{x_{jl}} \left(1 - P_{jl}\right)^{(1 - x_{jl})}. \tag{17}$$

Si l'on connaît la valeur de $\beta(l)$ ou si l'on en possède une estimation, il est alors facile de calculer l'estimateur $\hat{\theta}$ du maximum de vraisemblance. Il suffit de résoudre l'équation suivante en fonction de θ :

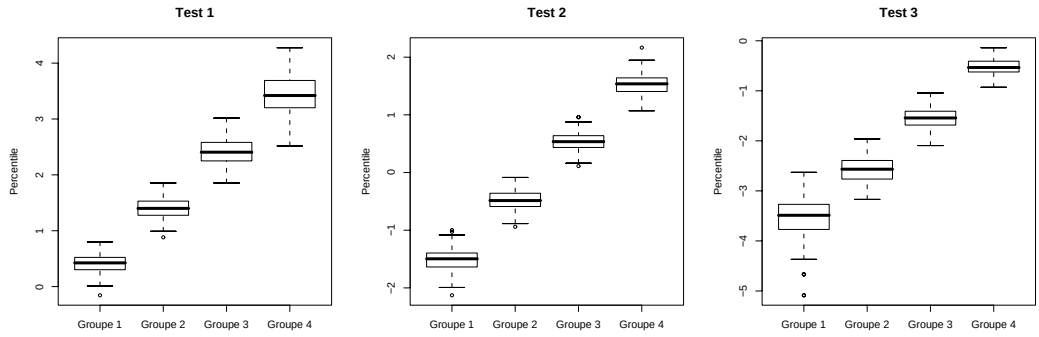
$$\frac{\partial \log P(\mathbf{X} = \mathbf{x}|\theta,\ \beta(l))}{\partial \theta} = 0. \tag{18}$$

2.4 Résultats

2.4.1 Premier test

Les distributions des compétences estimées à partir des résultats au premier test sont représentées dans la figure 7 (à gauche). Les compétences des individus du groupe 1 valent en moyenne 0.41. Celles des groupes 2, 3 et 4 valent 1.41, 2.42 et 3.42 respectivement.
Le trait latent sur lequel nous plaçons la difficulté des items et les compétences des individus ne possède pas de zéro absolu. Par convention, on place souvent la moyenne des items à l'origine. C'est ce que nous avons fait. La coordonnée du groupe 1 est donc égale à la distance qui sépare le groupe 1 de la moyenne des items. Théoriquement cette distance aurait dû être égale à 0.5 car les individus du groupe 1 ont tous une compétence de -1.5 et les items du test 1 ont en moyenne une difficulté égale à -2 (figure 3). Pour la même raison, le groupe 2 aurait dû se trouver à 1.5, le groupe 3 à 2.5 et le groupe 4 à 3.5 (tableau 4).

FIGURE 7 – *Distribution des scores obtenus par les individus des groupes 1, 2, 3 et 4 soumis aux tests 1, 2 et 3 respectivement. Les scores ont été calculés selon les préceptes de Rasch.*



2.4.2 Deuxième test

Les individus des groupes 1, 2, 3 et 4 ont été soumis au test 2. À partir de leurs performances et de l'estimation des difficultés des items du test 2, nous avons estimé leurs compétences. Les distributions des estimations sont représentées dans la figure 7 (au centre). La compétence moyenne du groupe 1 vaut -1.50, celle des groupes 2, 3 et 4 vaut -0.48, 0.54 et 1.53 respectivement. Ces résultats sont parfaitement en accord avec nos attentes (tableau 4).
Contrairement à ce que nous avions obtenu dans le cadre de l'approche classique (§ 1.5), nous constatons que les distances entre les groupes sont invariantes. La distance entre deux groupes ne dépend plus de l'instrument utilisé pour mesurer les positions des groupes.

TABLEAU 4 – *Position moyenne des groupes évalués à l'aide de trois tests différents.*

	Test 1		Test 2		Test 3	
	Score moyen		Score moyen		Score moyen	
	empirique	théorique	empirique	théorique	empirique	théorique
Groupe 1	0.41	0.5	−1.50	−1.5	−3.60	−3.5
Groupe 2	1.41	1.5	−0.48	−0.5	−2.56	−2.5
Groupe 3	2.42	2.5	0.54	0.5	−1.55	−1.5
Groupe 4	3.42	3.5	1.53	1.5	−0.53	−0.5

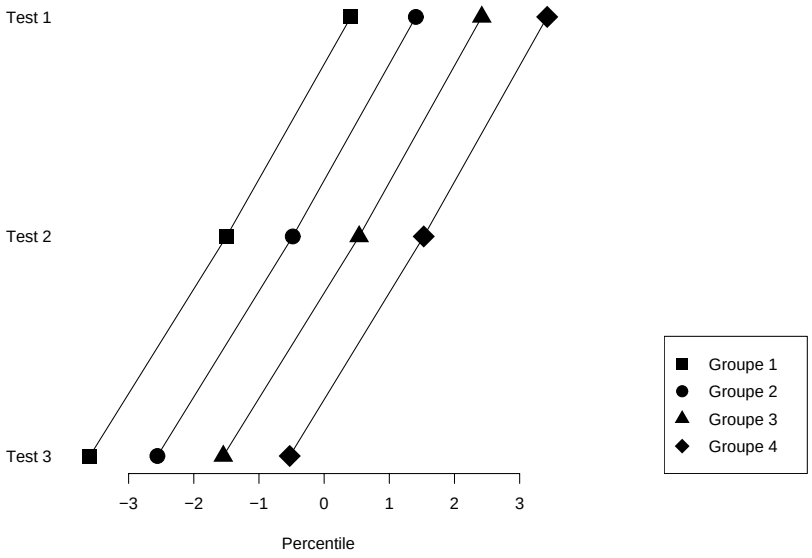
2.4.3 Troisième test

Les estimations des compétences des individus des groupes 1, 2, 3 et 4 faites à partir de leurs performances au test 3 sont représentées dans la figure 7 (à droite). La compétence moyenne du groupe 1 vaut −3.60, celle des groupes 2, 3 et 4 vaut −2.56, −1.55 et −0.53 respectivement. Encore une fois ces résultats sont conformes à nos attentes (tableau 4). Le constat que nous faisions précédemment peut être réitéré : les distances entre les groupes sont effectivement invariantes.

2.5 Bilan

La figure 8 fait apparaître toute la supériorité de l'approche de Rasch sur l'approche classique. Les instruments ont beau ne pas être les mêmes, les écarts entre les groupes restent les mêmes. À une translation près, le résultat des mesures est identique. Entre la première et la deuxième mesure, le décalage est de 2 ; entre la deuxième et la troisième aussi. Ce résultat était prévisible : ces décalages correspondent exactement aux différences de difficulté moyenne entre les trois tests. Rappelons que la difficulté moyenne du test 1 est égale à −2, que celle du test 2 est égale à 0 et celle du test 3 à 2.
Si nous avions voulu supprimer les décalages entre les trois tests, il aurait fallu incorporer dans le test 2 quelques items du test 1 et dans le test 3 quelques items du test 2. Ces items communs auraient permis d'ancrer toutes les mesures à une même origine.
À ce stade, nous pensons avoir montré que l'approche classique souffre de quelques limites qui dans certaines circonstances peuvent être palliées grâce à une approche innovée par Rasch fondée sur l'utilisation d'un modèle de la réponse à l'item.
Nous allons maintenant décrire la démarche à mettre en œuvre concrètement pour construire un instrument qui permette véritablement de mesurer objectivement un trait latent.

FIGURE 8 – *Comparaison des mesures faites à l'aide de trois instruments différents mais s'appuyant tous sur le modèle de Rasch.*



3 Construction d'un trait latent

La démarche que nous allons décrire s'appuiera sur un exemple simple que nous allons générer artificiellement à l'aide du logiciel statistique `R` [9, 11] :

```
> set.seed(3517)

> n <- 1000
> k <- 20

> theta0 <- rnorm(n)
> beta0 <- rnorm(k)

> diff <- outer(theta0, beta0, FUN="-")
> Prob.succes <- exp(diff)/(1 + exp(diff))

> data <- matrix(runif(n*k), ncol=k)
> data <- data < Prob.succes
> data[data==TRUE] <- 1
> dimnames(data) <- list(1:n, 1:k)
```

Les données `data` représentent les réponses de 1000 sujets soumis à un test composé de 20 items. Les sujets sont tirés aléatoirement d'une distribution normale centrée et réduite, les items également :

$$\theta_i \sim \mathcal{N}(0,\ 1) \qquad \text{et} \qquad \beta_j \sim \mathcal{N}(0,\ 1). \tag{19}$$

Les réponses des sujets sont générées conformément aux hypothèses du modèle de Rasch. Pour analyser ces données, nous procéderons en trois étapes : lors de la première étape, nous estimerons les paramètres du modèle (*i.e.* la compétence des sujets et la difficulté des items) ; lors de la deuxième, nous vérifierons que les conditions d'application du modèle de Rasch sont satisfaites et lors de la troisième, nous évaluerons la qualité de l'ajustement des données au modèle.

3.1 Estimation des paramètres

Les paramètres sont estimés à l'aide de la méthode du maximum de vraisemblance conjointe [10]. L'algorithme de calcul que nous avons programmé est enregistré dans le fichier `Rasch.R`. La librairie `Rasch.R`, qui contient toutes les fonctions nécessaires à la construction d'un trait latent, peut être téléchargée librement à partir de la page `http ://wwwpeople.unil.ch/jean-philippe.antonietti/`. Pour pouvoir utiliser les fonctions de la librairie `Rasch.R`, il suffit de taper dans la console de `R` l'instruction suivante :

```
> source("Rasch.R")
```

L'estimation de la valeur des paramètres du modèle s'effectue grâce à la commande :

```
> fit <- Rasch(data)
```

En sortie, la fonction `Rasch` fournit une liste de quatre objets qui se nomment `beta`, `SE.beta`, `theta` et `SE.theta`.

`beta`	est un vecteur formé des estimations de la difficulté des items $\hat{\beta}_j$.
`SE.beta`	est le vecteur des erreurs standard $\hat{\sigma}_j$ associées aux estimations de la difficulté des items.
`theta`	est un vecteur dont les valeurs sont égales aux estimations de la compétence des sujets $\hat{\theta}_i$.
`SE.theta`	est le vecteur des erreurs standard $\hat{sigma}_i$ associées aux estimations de la compétence des sujets.

Les erreurs standard $\hat{\sigma}_i$ et $\hat{\sigma}_j$ se calculent de la manière suivante [1, 17] :

$$\hat{\sigma}_i = \frac{1}{\sqrt{\sum_{j=1}^{k} \hat{P}_{ij}(1-\hat{P}_{ij})}} \tag{20}$$

$$\hat{\sigma}_j = \frac{1}{\sqrt{\sum_{i=1}^{n} \hat{P}_{ij}(1-\hat{P}_{ij})}} \tag{21}$$

avec

$$\hat{P}_{ij} = \frac{e^{(\hat{\theta}_i - \hat{\beta}_j)}}{1 + e^{(\hat{\theta}_i - \hat{\beta}_j)}}. \tag{22}$$

Notons qu'avec la méthode du maximum de vraisemblance conjointe, il n'est pas possible d'estimer la compétence des individus qui n'ont fourni aucune réponse juste ni celle de ceux qui ont répondu correctement à toutes les questions.
La figure 9 représente graphiquement le résultat des estimations. Cette image s'obtient comme suit :

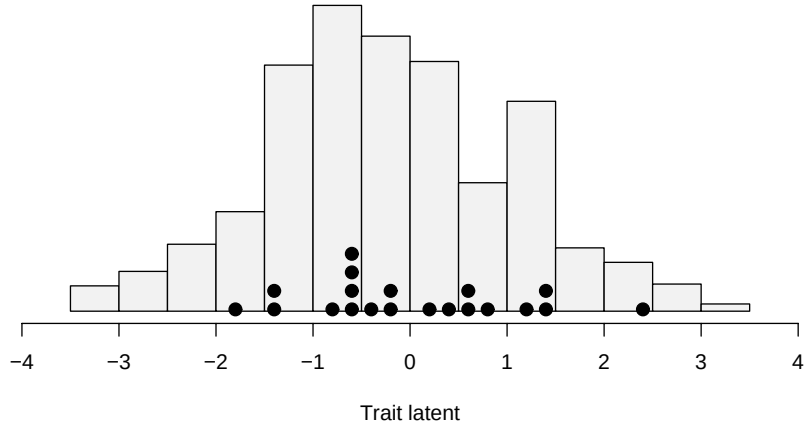
```
> beta <- fit$beta
> theta <- fit$theta
> plot.parameters(beta, theta)
```

Nous constatons que le test est bien approprié aux sujets : les items ne sont ni trop faciles, ni trop difficiles et, dans l'ensemble, ils recouvrent bien les compétences des sujets.

3.2 Vérification des postulats de base

Deux hypothèses vont être vérifiées (§ 2.1) : la première concerne l'unidimensionnalité et la seconde l'indépendance locale.

FIGURE 9 – *Compétence des individus et difficulté des items. Alors que les compétences sont représentées par un histogramme, chaque item est représenté par un point* (•).



3.2.1 Unidimensionnalité

La méthode que nous allons utiliser pour vérifier l'hypothèse d'unidimensionnalité est une analyse parallèle modifiée [3, 4]. En voici le schéma général :

1. L'on effectue l'analyse en composantes principales normée de la matrice des données brutes. Les valeurs propres λ_j de la matrice des corrélations des items sont ainsi calculées ; elles représentent, rappelons-le, les variances des composantes principales. En cas d'unidimensionnalité, l'on s'attend à ce que la première valeur propre soit très grande et que les autres soient négligeables puisque tous les items sont sensés contribuer à une seule et même dimension.
2. On génère de nouvelles données dichotomiques ayant une structure identique aux données brutes : la matrice générée possède un nombre de lignes égal au nombre de sujets observés et un nombre de colonnes égal au nombre d'items contenus dans le test. Les valeurs de chaque colonne X_j sont tirées aléatoirement d'une distribution de Bernoulli de paramètre égal à n_j/n où n_j est le nombre de sujets ayant répondu correctement à l'item j :
$$X_j \sim \mathcal{B}(1,\ n_j/n). \tag{23}$$
Deux colonnes quelconques X_j et $X_{j'}$ $(j \neq j')$ sont indépendantes.
L'on effectue alors l'analyse en composantes principales normée de ces données. Toutes les valeurs propres sont *grosso modo* égales à 1 étant donné que tous les items sont orthogonaux les uns aux autres.
3. Les opérations décrites au point 2 sont réitérées 500 ou 1000 fois. Il est ainsi possible d'établir la distribution des valeurs propres successives correspondant à l'analyse d'une matrice d'items indépendants mais de difficultés fixées.

4. La dimension des données observées est déterminée en comparant les valeurs propres empiriques à la distribution des valeurs propres théoriques correspondant à l'indépendance des items. Seules les dimensions pour lesquelles la valeur propre empirique est supérieure au quantile d'ordre $1 - \alpha$ de la distribution théorique associée sont retenues.

Pratiquement l'instruction à fournir pour réaliser cette analyse est la suivante :

```
> dimensionality(data)
```

Par défaut le nombre de matrices constituées d'items indépendants générées est égal à 500 et le seuil de signification α est fixé à 0.05. Ces deux paramètres peuvent être aisément modifiés :

```
> dimensionality(data, alpha=0.01, n.iter=1000)
```

La fonction `dimensionality` livre la liste des valeurs propres empiriques, la liste des valeurs propres critiques ainsi que le diagramme représentant le résultat des calculs (figure 10) :

```
$lambda.obs
 [1] 4.148 1.041 1.012 0.973 0.957 0.931 0.921 0.905 0.883 0.852
[11] 0.834 0.800 0.773 0.759 0.747 0.745 0.715 0.712 0.664 0.628

$lambda.crit
 [1] 1.314 1.260 1.213 1.184 1.153 1.126 1.098 1.075 1.053 1.029
[11] 1.008 0.988 0.966 0.945 0.923 0.903 0.882 0.860 0.838 0.808
```

Dans notre exemple, une seule valeur propre empirique dépasse la valeur propre critique ($\alpha = 0.01$). Ce résultat corrobore l'hypothèse d'unidimensionnalité.

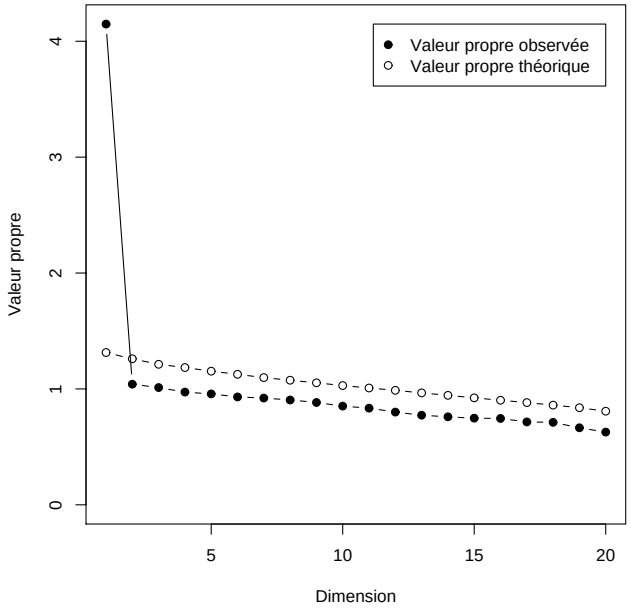
3.2.2 Indépendance locale

Lorsque l'hypothèse d'indépendance locale est satisfaite, les items pris deux à deux sont indépendants conditionnellement à θ_i :

$$\begin{aligned} &\text{Si} \qquad P(\mathbf{X} = \mathbf{x}|\theta_i) = \prod_{j=1}^{k} P(X_j = x_j|\theta_i), \\ &\text{alors} \qquad P(X_j = x_j \;\&\; X_{j'} = x_{j'}|\theta_i) = P(X_j = x_j|\theta_i) \cdot P(X_{j'} = x_{j'}|\theta_i). \end{aligned} \tag{24}$$

L'indépendance conditionnelle des paires d'items est donc une condition nécessaire à l'indépendance locale. Nous nous limiterons à l'examen de cette condition. Au cas où cette condition ne serait pas satisfaite, il en irait de même pour la condition d'indépendance locale. Dans le cas contraire, cela signifierait que l'hypothèse d'indépendance locale est plausible.
Pour se faire une idée de la validité de l'hypothèse d'indépendance locale des paires d'items, il suffit d'appliquer la démarche suivante :

FIGURE 10 – *Dimension des données.*



1. Isoler tous les individus ayant répondu correctement à r items exactement. Ces individus ont tous la même compétence $\hat{\theta}_r$.
2. Pour tous les couples d'items $(j,\ j')$ construire la table de contingence croisant les réponses à l'item j aux réponses à l'item j' :

	item j'		
item j	0	1	Total
0	$n_{r\bar{j}\bar{j'}}$	$n_{r\bar{j}j'}$	$n_{r\bar{j}\cdot}$
1	$n_{rj\bar{j'}}$	$n_{rjj'}$	$n_{rj\cdot}$
Total	$n_{r\cdot\bar{j'}}$	$n_{r\cdot j'}$	$n_{r\cdot\cdot}$

   Calculer la valeur du khi carré $\chi^2_{rjj'}$ associée à chacune de ces tables[1] :

$$\chi^2_{rjj'} = \frac{n_{r\cdot\cdot}(n_{r\bar{j}\bar{j'}} \cdot n_{rjj'} - n_{r\bar{j}j'} \cdot n_{rj\bar{j'}})^2}{n_{r\bar{j}\cdot} \cdot n_{rj\cdot} \cdot n_{r\cdot\bar{j'}} \cdot n_{r\cdot j'}}. \tag{25}$$

   Comparer $\chi^2_{rjj'}$ au quantile d'ordre $1-\alpha$ de la distribution du χ^2 à un degré de liberté.
3. Réitérer les deux étapes précédentes pour toutes les valeurs possibles de $r \in \{1,\ 2,\ \ldots,\ k-1\}$.

Pour réaliser ces opérations, il faut saisir dans la console de `R` la commande :

[1]Si l'un des effectifs attendus en cas d'indépendance est inférieur à 5, le $\chi^2_{rjj'}$ n'est pas calculé.

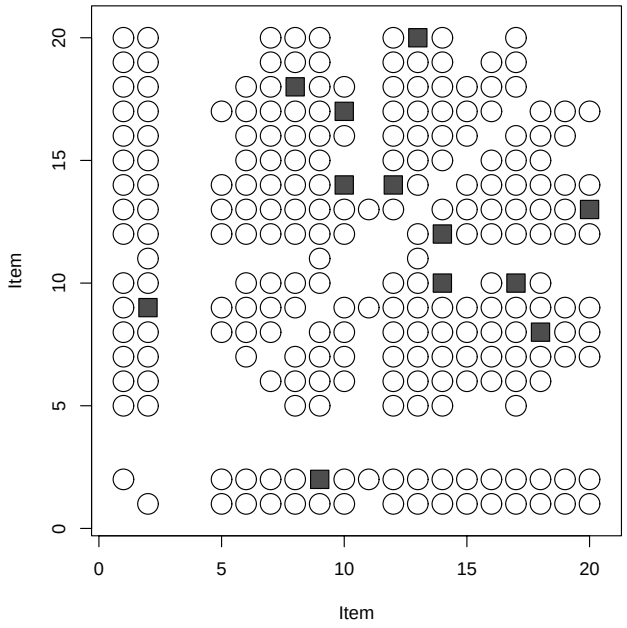
```
> R <- local.independence(data)
```

L'objet `R` est une liste de $k-1$ matrices. La première matrice est formée des $\chi^2_{1jj'}$, la deuxième est formée des $\chi^2_{2jj'}$, la r-ième est formée des $\chi^2_{rjj'}$, et la dernière est formée des $\chi^2_{k-1\ jj'}$.
Il est possible de représenter graphiquement le résultat des comparaisons entre $\chi^2_{rjj'}$ et $\chi^2_{1-\alpha}[1]$. Si l'on veut, par exemple, visualiser l'ensemble des tests fait à partir des réponses des individus ayant obtenu un score global de 8, l'instruction à donner est la suivante :

```
> local.ind.plot(R, 8)
```

Par défaut le seuil de signification α est fixé à 0.05 mais il peut facilement être modifié en assignant une autre valeur à l'argument `alpha` de la fonction `local.ind.plot`.
La figure 11 représente l'état des dépendances au sein du groupe des individus ayant obtenu un score de 8.

FIGURE 11 – *Dépendance au sein du groupe ayant obtenu un score de 8. Si à la position* $(j,\ j')$*, il y a un cercle* (◯) *cela signifie que les items* j *et* j' *sont indépendants au seuil* α *; s'il y a un carré* (■)*, cela signifie que les items* j *et* j' *sont dépendants et si aucun symbole n'est dessiné, cela signifie que le nombre d'observations est trop restreint pour autoriser la mise en application d'un test d'indépendance.*



En première approximation, pour que l'on puisse accepter l'hypothèse d'indépendance conditionnelle des paires d'items, il ne faudrait pas que la proportion des tests significatifs dépasse le seuil α. Dans l'exemple que nous traitons, cette proportion est inférieure à $\alpha = 0.05$. Ce constat rend l'hypothèse d'indépendance locale plausible.

3.3 Tests d'ajustement

Une fois les postulats de base corroborés – c'est le cas –, il importe d'évaluer la qualité de l'ajustement des données au modèle. Nous proposerons deux procédures : l'une, générale, relativement peu informative, et l'autre, plus particulière, qui nous livrera des informations précises concernant chaque item du test et chaque individu interrogé.

3.3.1 Test général

Le test global se fonde sur une statistique du χ^2. Wright et Panchapakesan [16] montrèrent que lorsque les données se conforment au modèle de Rasch la statistique

$$\chi^2 = \sum_{r=1}^{k-1} \sum_{j=1}^{k} \frac{(n_{rj} - n_r \hat{P}_{rj})^2}{n_r \hat{P}_{rj}(1 - \hat{P}_{rj})} \tag{26}$$

avec

$$\hat{P}_{rj} = \frac{e^{(\hat{\theta}_r - \hat{\beta}_j)}}{e^{(\hat{\theta}_r - \hat{\beta}_j)} + 1} \tag{27}$$

se distribue asymptotiquement selon une loi du χ^2 à $(k-2)(k-1)$ degrés de liberté.
Pour réaliser ce test, il suffit d'inscrire sur la console de `R` la commande :

```
> test1 <- Rasch.model.test(data, beta, theta)

  chi2 = 313.3818, df = 342, p.value = 0.8645
```

La fonction `Rasch.model.test` crée six objets qui sont :

`n.rj`	la matrice des n_{rj} où n_{rj} est le nombre d'individus qui ont obtenu le score global de r et ont répondu correctement à l'item j.
`prob.rj`	la matrice des $\hat{P}_{rj}$ où $\hat{P}_{rj}$ est la probabilité qu'un individu ayant la compétence $\hat{\theta}_r$ réponde correctement à la question j.
`n.r`	le vecteur des n_r où n_r est égal au nombre d'individus ayant obtenu le score global r.
`chi2`	la valeur empirique χ^2 de la variable de décision du test d'ajustement.
`df`	le nombre de degrés de liberté $(k-2)(k-1)$.
`p.value`	la probabilité critique $p = P(\chi^2[(k-2)(k-1)] \geq \chi^2 \mid H_0)$ avec H_0 l'hypothèse selon laquelle les données s'ajustent au modèle de Rasch.

Grâce à ce test, nous pouvons conclure – sans surprise – que les données sont adéquatement décrites par un modèle de Rasch ($p > 0.05$).
Lorsqu'un test est composé de nombreux items et que la taille de l'échantillon interrogé est relativement modeste, les groupes formés par les individus ayant tous obtenu le même score global sont souvent petits. Dans ce cas, il est courant de réunir les groupes voisins.

Supposons que nous ne voulions former dans notre échantillon que trois grands groupes ($G = 3$) : le premier rassemblant les individus ayant obtenu un score compris entre 1 et 6, le deuxième rassemblant les individus ayant obtenu un score compris entre 7 et 13 et le troisième ceux ayant obtenu un score compris entre 14 et 19. Avec ce nouveau regroupement le test d'adéquation s'effectue ainsi :

```
> G.class <- rep(1:3, c(6, 7, 6))
> test2 <- G.class.test(G.class, data, beta, theta)

  chi2 = 49.75461, df = 38, p.value = 0.096
```

Lorsque les données s'ajustent au modèle de Rasch la statistique

$$\chi^2 = \sum_{g=1}^{g} \sum_{j=1}^{k} \frac{(n_{gj} - n_g \hat{P}_{gj})^2}{n_g \hat{P}_{gj}(1 - \hat{P}_{gj})} \tag{28}$$

avec

$$\hat{P}_{gj} = \frac{e^{(\hat{\theta}_g - \hat{\beta}_j)}}{e^{(\hat{\theta}_g - \hat{\beta}_j)} + 1} \qquad \text{et} \qquad \hat{\theta}_g = \frac{\sum\limits_{g \in G} n_r \hat{\theta}_r}{\sum\limits_{g \in G} n_r} \tag{29}$$

se distribue asymptotiquement selon une loi du χ^2 à $(G-1)(k-1)$ degrés de liberté. Dans notre exemple, nous voyons que la conclusion de ce deuxième test est la même que celle que nous avions obtenue à l'issue du premier ($p > 0.05$).

3.3.2 Tests particuliers

Ajustement des items Pour apprécier la qualité de l'ajustement des données à la courbe caractéristique théorique d'un item j – conforme donc au modèle de Rasch –, nous commençons par calculer l'erreur standardisée z_{ij} associée à la réponse x_{ij} d'un sujet i :

$$z_{ij} = \frac{x_{ij} - \hat{P}_{ij}}{\sqrt{\hat{P}_{ij}(1 - \hat{P}_{ij})}} \tag{30}$$

où $\hat{P}_{ij}$ est, comme précédemment, l'estimation de la probabilité que le sujet i réponde correctement à la question j :

$$\hat{P}_{ij} = \frac{e^{(\hat{\theta}_i - \hat{\beta}_j)}}{e^{(\hat{\theta}_i - \hat{\beta}_j)} + 1}. \tag{31}$$

Asymptotiquement la distribution de l'erreur standardisée suit une loi normale centrée-réduite [17] :

$$Z_{ij} = \frac{X_{ij} - \hat{P}_{ij}}{\sqrt{\hat{P}_{ij}(1 - \hat{P}_{ij})}} \sim \mathcal{N}(0,\ 1). \tag{32}$$

Le carré de l'erreur standardisée suit, à son tour, une loi du χ^2 à un degré de liberté :

$$Z_{ij}^2 \sim \chi^2[1]. \tag{33}$$

Il est possible de montrer que, sur l'ensemble des réponses fournies à la question j, la variable

$$V_j = \frac{\sum_{i=1}^{n} Z_{ij}^2}{n-1} \tag{34}$$

suit une distribution de Fisher-Snedecor à $n-1$ et ∞ degrés de liberté :

$$V_j \sim F[n-1,\ \infty]. \tag{35}$$

Par commodité nous appliquons la transformation de Wilson-Hilferty [14] à cette variable V_j :

$$T_j = (\ln(V_j) + V_j - 1) \cdot \sqrt{\frac{n-1}{8}}. \tag{36}$$

Lorsque les données s'ajustent au modèle, la variable T_j suit une loi normale centrée-réduite :

$$T_j \sim \mathcal{N}(0,\ 1). \tag{37}$$

Il suffit donc de calculer la valeur empirique t_j de cette variable T_j puis de voir si cette valeur appartient bien à la zone d'acceptation de l'hypothèse H_0 affirmant que les données sont correctement décrites par le modèle. On accepte H_0 si $u_{\alpha/2} < t_j < u_{1-\alpha/2}$ où $u_{\alpha/2}$ et $u_{1-\alpha/2}$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$, respectivement, d'une distribution normale centrée-réduite. Souvent, l'on choisi un seuil α de 5% ; on conclut donc que les données s'ajustent bien à la courbe caractéristique théorique de l'item lorsque t_j est *grosso modo* compris entre -2 et $+2$.
La commande qui permet calculer la valeur des t_j est la suivante :

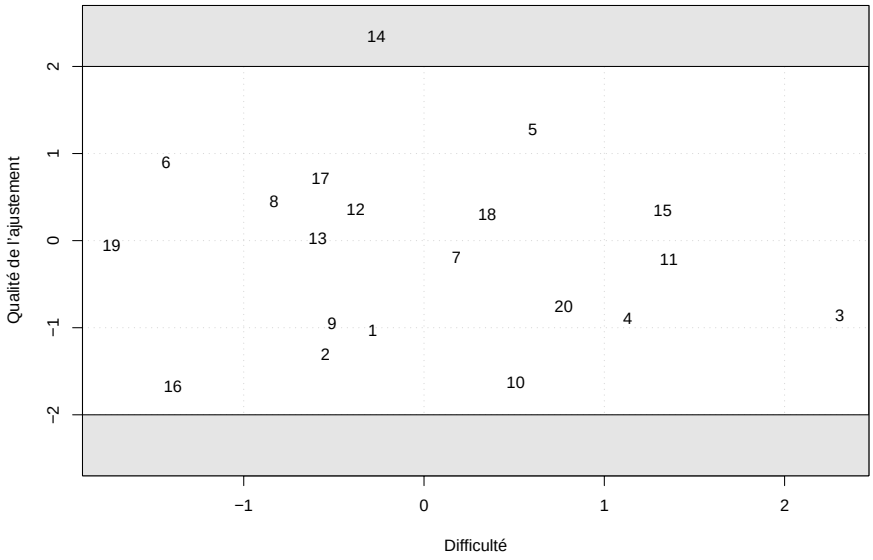
```
> misfit.item(data, beta, theta)
```

Pour visualiser le résultat de tous les tests d'ajustement, il suffit d'exécuter la commande :

```
> plot.item.fit(data, beta, theta)
```

Nous constatons, à partir de l'examen de la figure 12, que les réponses à quasiment tous les items s'ajustent aux contraintes imposées par le modèle de Rasch. Un seul item, le quatorzième, se trouve à l'extérieur de la bande horizontale définissant les items bien ajustés.

FIGURE 12 – *Ajustement des items selon leur difficulté. Les deux horizontales d'ordonnées* -2 *et* $+2$ *définissent une bande de confiance dans laquelle se trouvent les items qui satisfont les hypothèses du modèle de Rasch.*



Montrons brièvement comment varie la valeur du coefficient d'ajustement t_j en fonction de la discrimination α_j d'un item. Arbitrairement, fixons la difficulté de l'item j à 0. Le modèle de la réponse à l'item que nous allons utiliser ici satisfait l'équation suivante :

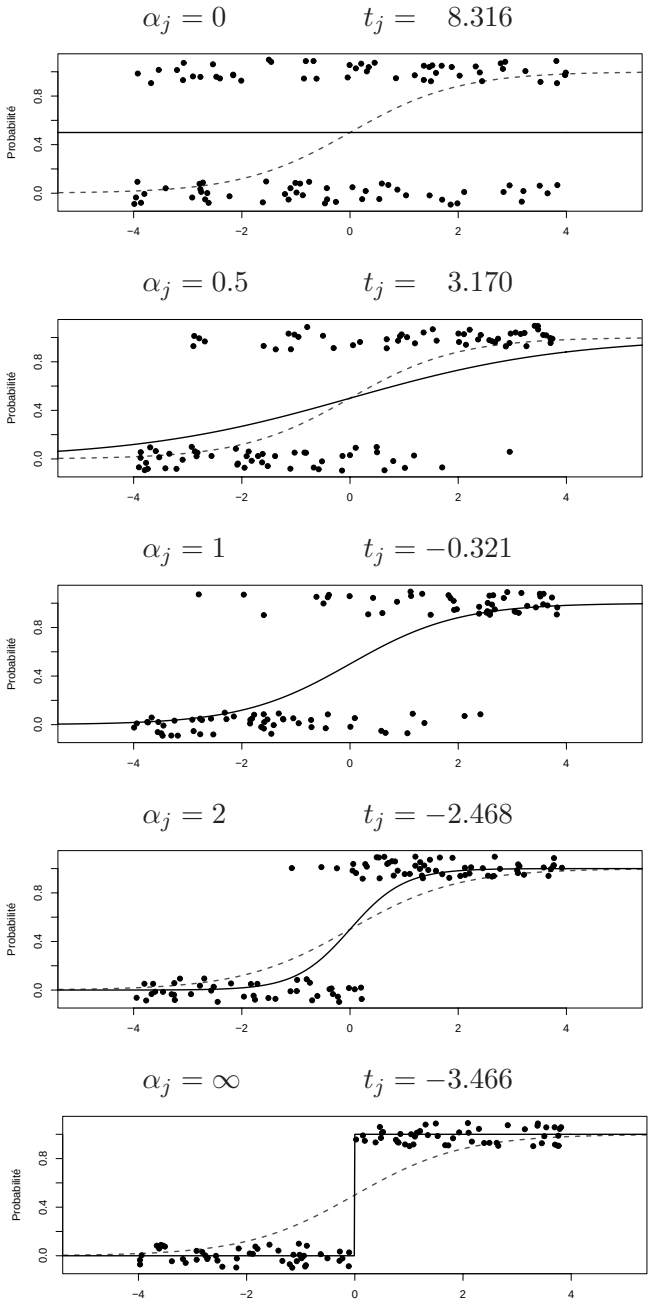
$$P(X = 1|\theta,\ \alpha_j) = \frac{e^{\alpha_j \theta}}{1 + e^{\alpha_j \theta}}. \tag{38}$$

Pour chaque α_j, nous allons :

1. Tirer d'une population uniforme sur l'intervalle des compétences allant de -4 à $+4$ un échantillon de taille 100.
2. Soumettre chaque individu de l'échantillon à l'item j.
3. Calculer, à partir des réponses enregistrées, la valeur du coefficient d'ajustement t_j.

Assignons à α_j les valeurs suivantes : 0, 0.5, 1, 2 et ∞. Le résultat des simulations est représenté dans la figure 13. Nous voyons que lorsque la discrimination α_j est proche de 1, le coefficient d'ajustement est proche de 0. Lorsque la discrimination de l'item diminue, le coefficient d'ajustement augmente et risque de dépasser 2. Au contraire, lorsque la discrimination augmente, le coefficient d'ajustement diminue et peut, dans ce cas, prendre une valeur inférieure à -2.

FIGURE 13 – *Influence de la discrimination sur l'adéquation. Dans chaque graphique l'on trouve les réponses des sujets (points noirs), la courbe caractéristique de l'item ayant une discrimination égale à α_j (ligne continue) et la courbe caractéristique d'un item de référence ayant une discrimination de 1 (ligne traitillée).*

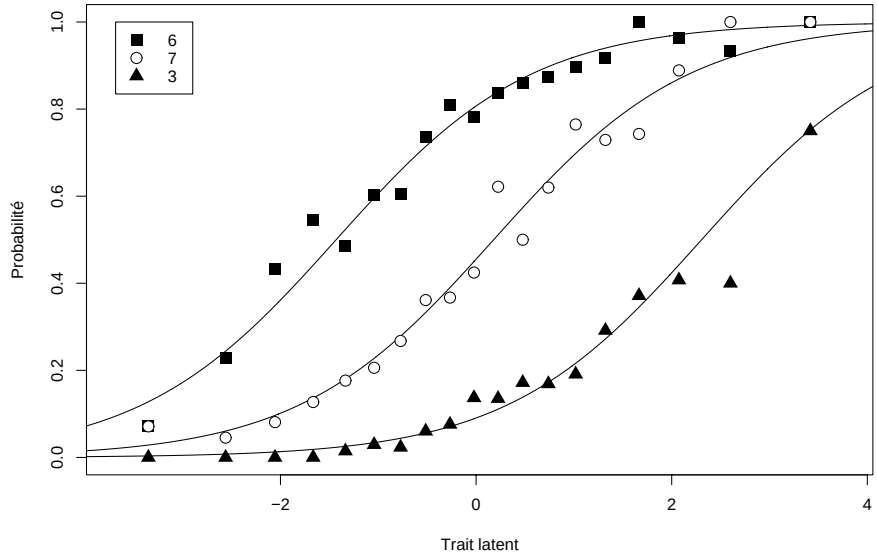


Pour se faire une meilleure idée de l'ajustement des données au modèle, il est judicieux de représenter dans un même graphique les courbes caractéristiques des items et les taux de bonnes réponses observés selon les compétences estimées. À titre d'exemple, représentons les courbes caractéristiques des items 6, 7, et 3 (le premier est un item facile, le deuxième est un item moyennement difficile et le troisième est difficile) :

```
> ICC(c(6, 7, 3), data, beta, theta)
```

Les observations s'ajustent au modèle : les taux de bonnes réponses épousent presque parfaitement les courbes caractéristiques des items (figure 14).

FIGURE 14 – *Courbes caractéristiques de trois items.*



Ajustement des personnes De manière similaire, il est possible d'examiner les motifs de réponses des personnes. Les commandes à utiliser sont :

```
> misfit.person(data, beta, theta)
> plot.person.fit(data, beta, theta)
```

Par curiosité, affichons les réponses des deux individus ayant les coefficients d'ajustement les plus extrêmes en prenant soin d'ordonner les items du plus facile au plus difficile :

```
> fit.person <- misfit.person(data, beta, theta)
> data[which.min(fit.person), order(beta)]
 19  6 16  8 13 17  2  9 12  1 14  7 18 10  5 20  4 15 11  3
  1  1  1  1  1  1  1  1  1  1  1  1  0  0  1  0  0  0  0  0
```

```
> data[which.max(fit.person), order(beta)]
19  6 16  8 13 17  2  9 12  1 14  7 18 10  5 20  4 15 11  3
 0  0  0  1  1  0  1  0  0  1  0  1  1  0  1  0  0  1  0  1
```

L'individu ayant le coefficient d'ajustement le plus petit ($t_i = -1.72$) répond systématiquement juste à tous les items les plus faciles puis, à partir d'un certain seuil, systématiquement faux. La transition entre ses réponses justes et ses réponses fausses se fait de manière très abrupte. L'on pourrait presque utiliser un modèle déterministe de Guttman [6, 7] pour décrire son comportement.
Au contraire, celui ayant le coefficient d'ajustement le plus grand ($t_i = 3.61$) a un comportement très instable, ses bonnes réponses fluctuent beaucoup et, bizarrement, il répond de manière erronée aux questions les plus faciles et tout à fait correctement aux questions les plus difficiles.

Nous proposons à notre lecteur d'effectuer une analyse similaire à celle que nous venons de réaliser sur un nouvel exemple dans lequel tous les items ne satisfont pas les contraintes du modèle de Rasch.

```
> set.seed(3517)
> n <- 1000
> k <- 20

> theta0 <- rnorm(n)
> beta0 <- rnorm(k)
> alpha0 <- c(0.5, rep(1, 18), 2)

> diff <- outer(theta0, beta0, FUN="-") %*% diag(alpha0)
> Prob.succes <- exp(diff)/(1 + exp(diff))

> data <- matrix(runif(n*k), ncol=k)
> data <- data < Prob.succes
> data[data==TRUE] <- 1
> dimnames(data) <- list(1:n, 1:k)
```

Conclusion

En psychologie, comme dans toute autre science, il est souhaitable de pouvoir effectuer des mesures objectives. De telles mesures nécessitent la définition d'une unité de mesure, de grandeur constante, reproductible tout au long de l'échelle et indépendante de l'objet mesuré. Comme nous l'avons montré, le modèle de Rasch permet de réaliser ce souhait à condition, bien évidemment, que les données observées satisfassent les prescriptions du modèle.

Dans ce fascicule, nous nous sommes limités au traitement des données dichotomiques mais il existe de nombreuses extensions au modèle de Rasch telles que le modèle d'échelle de classement ou le modèle du crédit partiel [2, 5, 13, 15]. Ces modèles peuvent facilement être construits et analysés dans `R` à l'aide de routines écrites et mises à disposition par Dimitri Rizopoulos. Ces procédures sont incluses dans la librairie `ltm` (*Latent Trait Models under IRT*) disponibles sur le site *The Comprehensive R Archive Network* dont l'adresse est `http ://cran.r-project.org/`.

Bibliographie

[1] D. Andrich, *Rasch models for measurement*, Sage, Newbury Park, London, New Delhi, 1988.

[2] A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders, eds., *Essays on item response theory*, Springer, New York, 2001.

[3] D. V. Budescu, Y. Cohen, & A. Ben-Simon, *A revised modified parallel analysis for the construction of unidimensional item pools*, Applied Psychological Measurement, 21 (1997), pp. 233–253.

[4] F. Drasgow & R. I. Lissak, *Modified parallel analysis : A procedure for examining the latent dimensionality of dichotomously scored item responses*, Journal of Applied Psychology, 68 (1983), pp. 363–373.

[5] G. H. Fischer & I. W. Molenaar, eds., *Rasch models : Foundations, recent developments, and applications*, Springer, New York, 1995.

[6] L. Guttman, *A basis for scaling qualitative data*, American Sociological Review, 9 (1944), pp. 139–150.

[7] ——, *The basis for scalogram analysis*, in Measurement and prediction : Studies in social psychology in World War II, S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen, eds., vol. 4, Princeton University Press, Princeton, NJ, 1950, pp. 60–90.

[8] H. Hoijtink & A. Boomsma, *On person parameter estimation in the dichotomous Rasch model*, in Rasch models : Foundations, recent developments, and applications, G. H. Fischer & I. W. Molenaar, eds., Springer, New York, 1995, pp. 39–51.

[9] R. Ihaka & R. Gentleman, *R : A language for data analysis and graphics*, Journal of Computational and Graphical Statistics, 5 (1996), pp. 299–314.

[10] 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.

[11] R Development Core Team, *R : A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, 2003.

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

[13] W. J. Van der Linden & R. K. Hambleton, eds., *Handbook of modern item response theory*, Springer, New York, 1997.

[14] E. B. Wilson & M. M. Hilferty, *The distribution of chi-square*, Proceedings of the National Academy of Sciences of the United States of America, 17 (1931), pp. 684–688.

[15] B. Wright & G. N. Masters, *Rating scale analysis*, MESA Press, Chicago, 1982.

[16] B. Wright & N. Panchapakesan, *A procedure for sample-free item analysis*, Educational and Psychological Measurement, 29 (1969), pp. 23–48.

[17] B. Wright & M. H. Stone, *Best test design*, MESA Press, Chicago, 1979.

Table des matières