Évaluation des compétences : les apports des modèles de réponse à l'item

Thierry Rocher

DEPP - Direction de l'évaluation, de la prospective et de la performance
Ministère de l'éducation nationale

Les rencontres de statistiques appliquées
INED, 9 juin 2017

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
ministère éducation nationale

Contexte

Dispositifs d'évaluations des élèves conduits par la DEPP

- Évaluations standardisées
- Nationales et internationales
- Principalement sur échantillons
- Objectif : mesurer et suivre les résultats du système

Corpus méthodologique

- Mesure en éducation
- Psychométrie

Mesurer une variable latente

Mesurer la taille des individus...

Mesurer une variable latente

Mesurer la taille des individus... avec un questionnaire

Mesurer une variable latente

Mesurer la taille des individus... avec un questionnaire

1. Je dois souvent faire attention à ne pas me cogner la tête
2. Pour les photos de groupe, on me demande souvent d'être au premier rang
3. On me demande souvent si je fais du basket-ball
4. Dans la plupart des voitures, je suis mal assis(e)
5. Je dois souvent faire faire les ourlets quand j'achète un pantalon
6. Je dois souvent me baisser pour faire la bise
7. Au supermarché, je dois souvent demander de l'aide pour attraper des produits en haut des gondoles
8. A deux sous un parapluie, c'est souvent moi qui le tiens

...

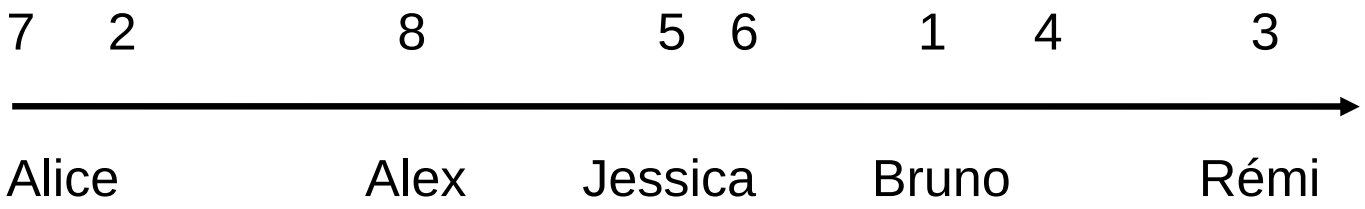
Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
ministère éducation nationale

Illustration

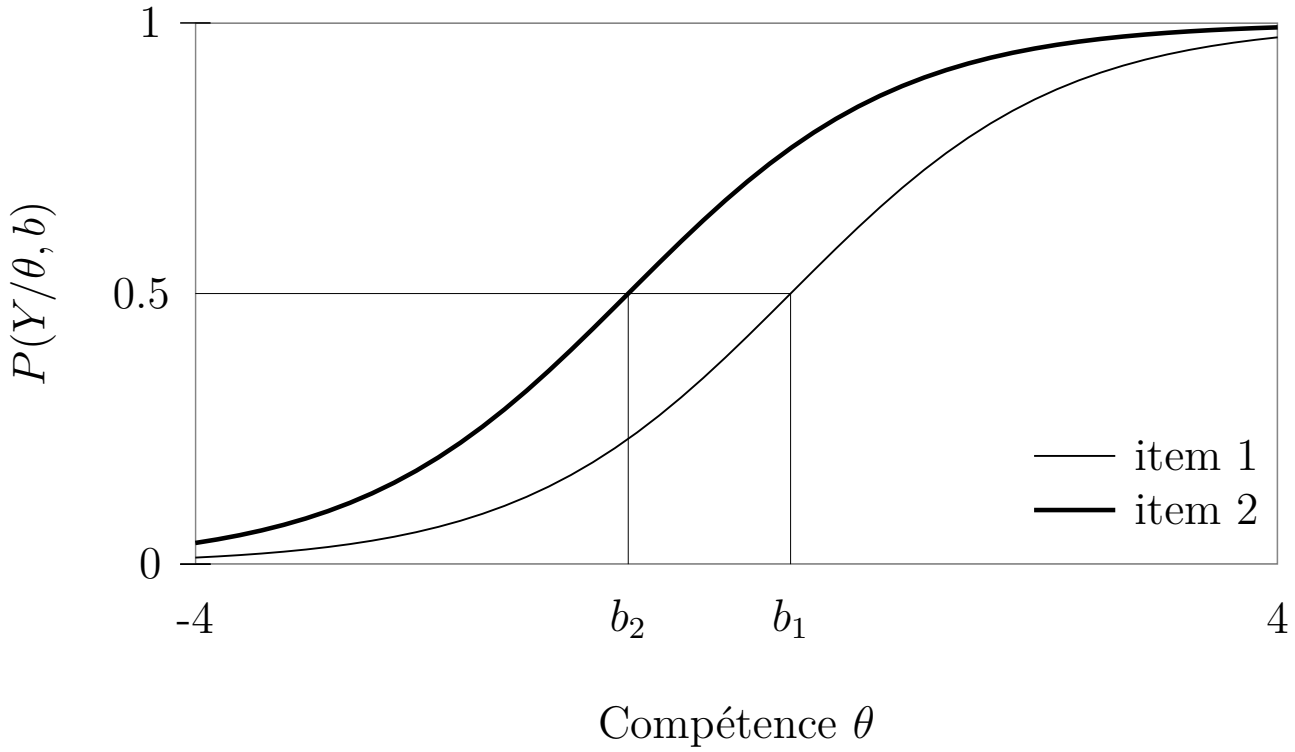
Passation : 24 items, 276 individus

Quelques notions de psychométrie :

- Validité : corrélation(score construit, taille réelle), r=0.85
- Fidélité : à quelques exceptions près, les items forment un ensemble homogène
- Fonctionnements différentiels : quelques items selon le genre
- Echelle d'intervalle : classement des individus + métrique
- Métrique : l'échelle n'a pas d'unité pré-définie ni de 0 absolu ($\approx$ échelles de température)



Modèle de Rasch (1960), 1 paramètre

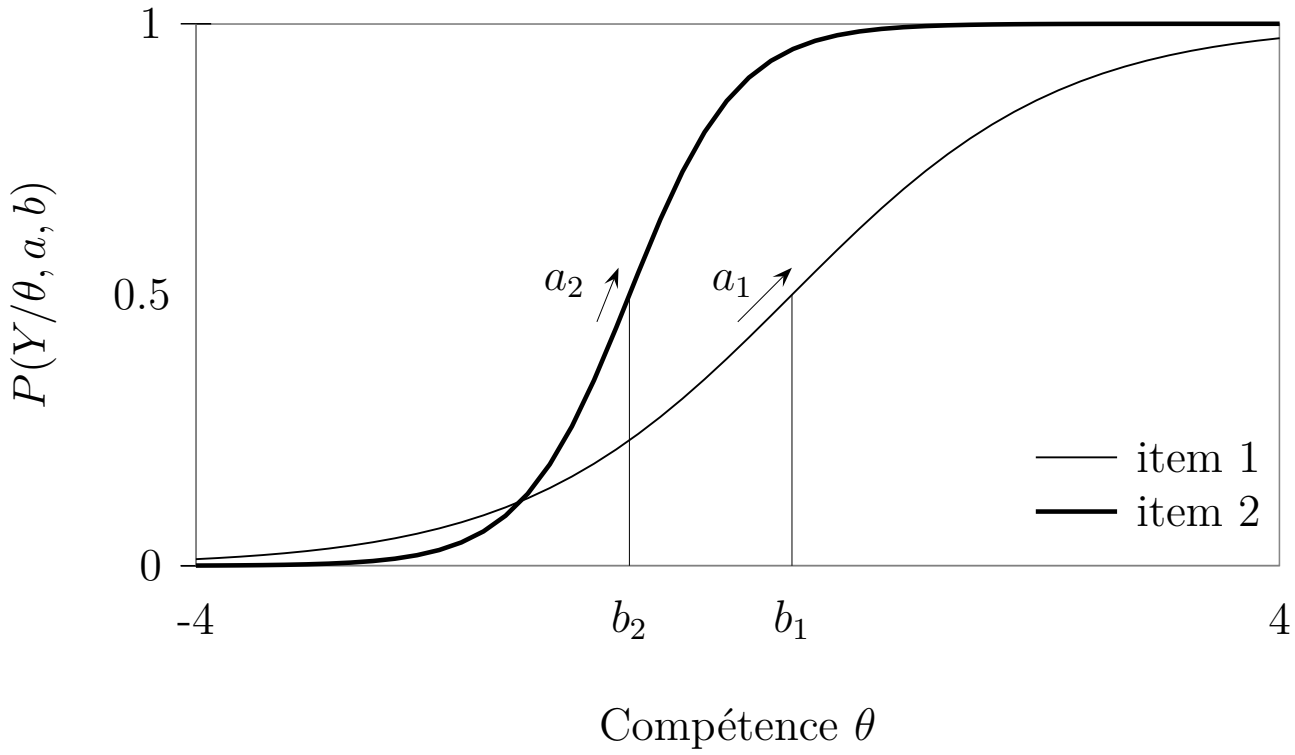


$$P(Y_i^j = 1/\theta_i, b_j) = \frac{exp(\theta_i - b_j)}{1 + exp(\theta_i - b_j)}$$

Y_i^j : réponse de l'élève i à l'item j
θ_i : niveau de compétence de l'élève i

b_j : difficulté de l'item j

Modèle de Birnbaum (1968), 2 paramètres



$$P(Y_i^j = 1/\theta_i, a_j, b_j) = \frac{exp(a_j(\theta_i - b_j))}{1 + exp(a_j(\theta_i - b_j))}$$

Y_i^j : réponse de l'élève i à l'item j
θ_i : niveau de compétence de l'élève i

b_j : difficulté de l'item j
a_j : discrimination de l'item j

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
ministère éducation nationale

Cadre général

De manière générale :

$$P(Y = k/\boldsymbol{\theta}, \boldsymbol{\xi}) = F(\boldsymbol{\theta}, \boldsymbol{\xi}, k)$$

- Item (Y) : dichotomique ou polytomique (k = 1, 2, 3, ...)
- Compétence ($\boldsymbol{\theta}$) : unidimensionnelle ou multidimensionnelle, continue ou catégorielle, structure simple ou complexe
- Paramètres des items ($\boldsymbol{\xi}$) : 1, 2, 3, + temps, correcteurs
- Fonction de lien F : logistique, normale

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
ministère éducation nationale

Avantages des MRI

Séparation des concepts

- niveau de difficulté des items
- niveau de compétence des élèves

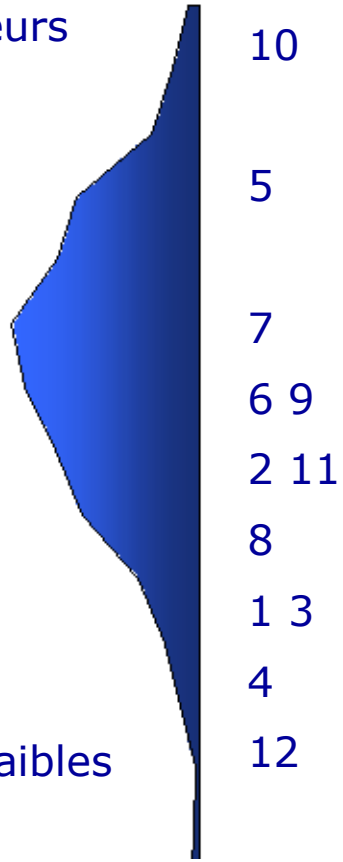
Construction d'échelle

- une même échelle pour les paramètres de difficulté des items et les niveaux de compétence des élèves
- analyse pédagogique : description des tâches maîtrisées en fonction des niveaux de compétence

Échelle

Distribution des élèves (θ) Distribution des items (b)

élèves les meilleurs



items difficiles

élèves les plus faibles

items faciles

Un mot sur l'estimation

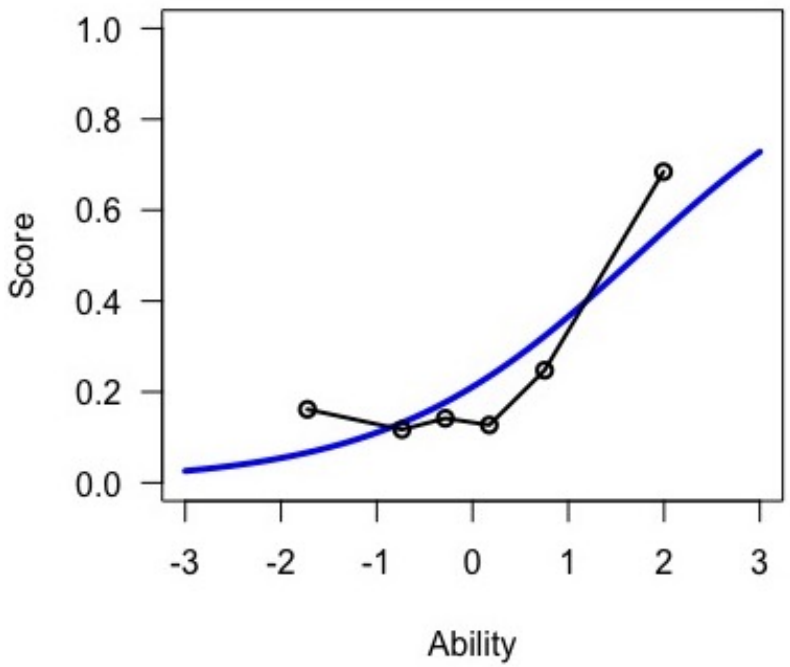
1. Estimation des paramètres des items :
   - *Marginal Maximum Likelihood* (MML)
   - La distribution des θ est fixée
   - Calculs d'intégrales un peu « coûteux »
2. Estimation des scores :
   - Les paramètres des items estimés sont considérés comme fixes
   - L'estimateur du maximum de vraisemblance est biaisé (!)
   - Estimation d'un maximum de vraisemblance pondéré (WML)

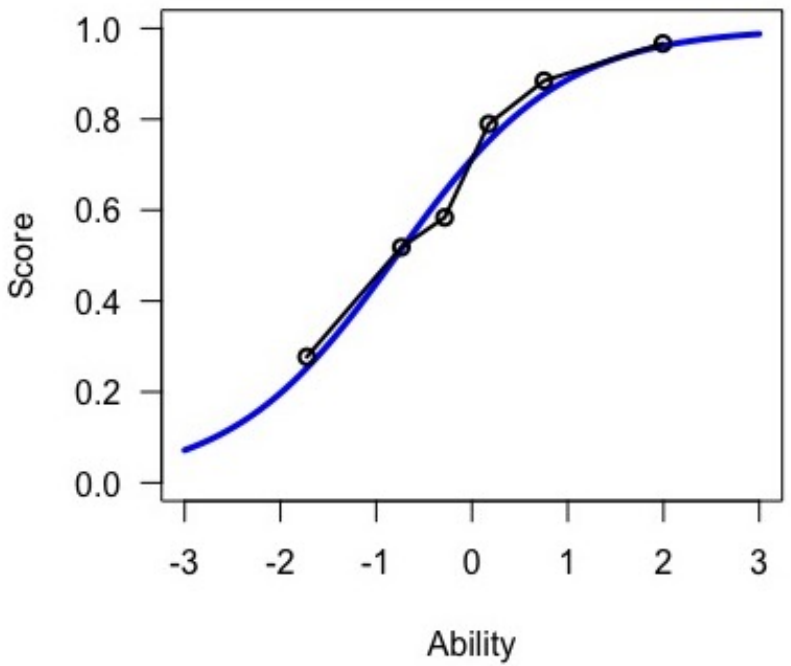
Note : les MRI peuvent également être vus comme des modèles à effets aléatoires (multiniveaux), ou encore comme des analyses factorielles confirmatoires

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
ministère éducation nationale

Ajustement du modèle

Pour chaque item, comparaison entre les données empiriques et le modèle :





Indétermination

Modèle de réponse à l'item (2PL) :

$$P(Y_i^j = 1/\theta_i, a_j, b_j) = \frac{exp(a_j(\theta_i - b_j))}{1 + exp(a_j(\theta_i - b_j))}$$

Les paramètres sont définis à une transformation linéaire près :

$$\left| \begin{array}{l} \theta_i^* = A\theta_i + B \\ a_j^* = a_j/A \\ b_j^* = Ab_j + B \end{array} \right.$$

Généralement, lors de l'estimation on fixe $\mu_\theta = 0$ et $\sigma_\theta = 1$

Problème : comparer des élèves ayant passé deux évaluations différentes

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
ministère éducation nationale

Ancrage

Ajustement des métriques (*equating*) via des items communs : positionner les niveaux de compétences sur la même échelle, grâce à des items repris à l'identique d'une évaluation à l'autre

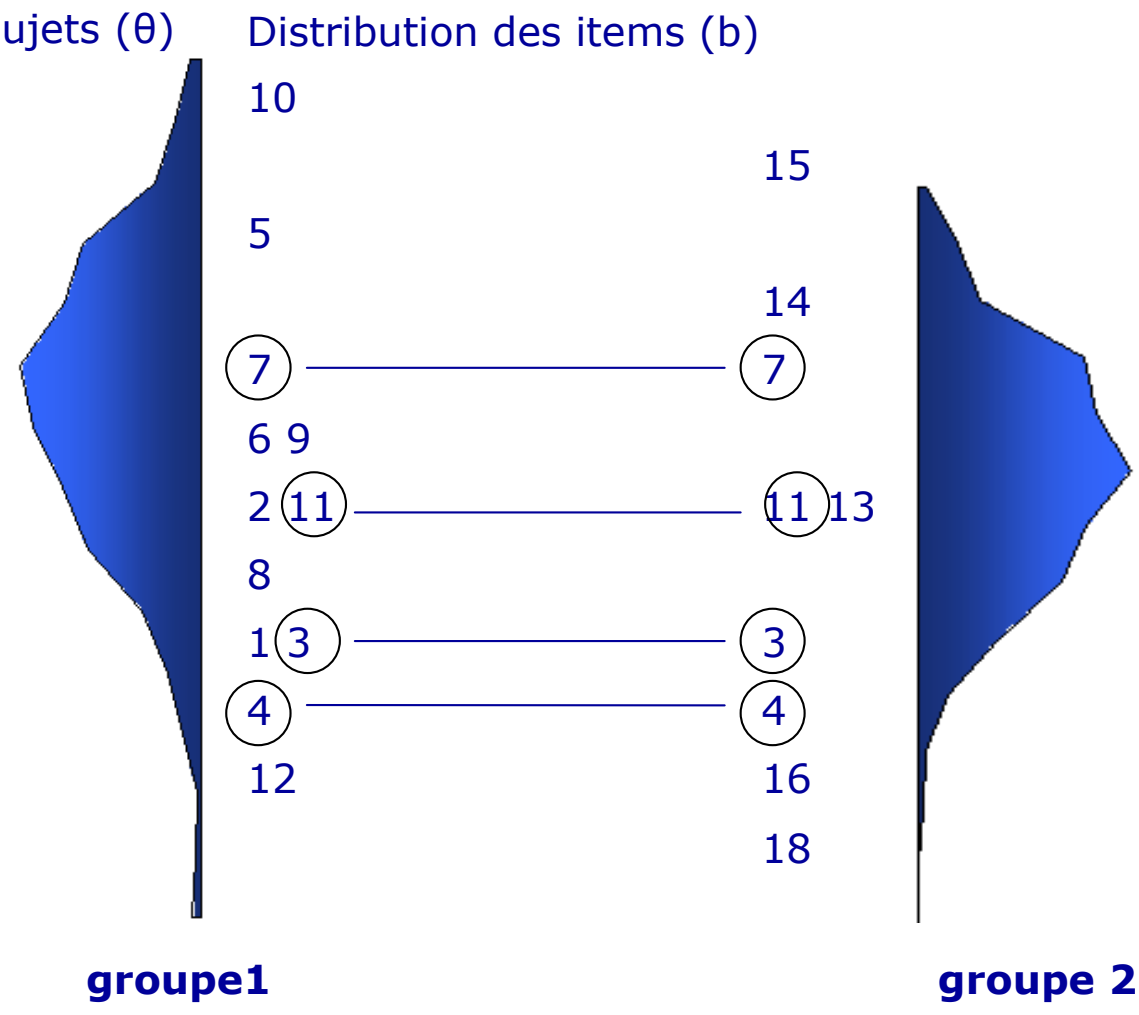
D'une évaluation à l'autre :

- paramètres des items : fixes
- distribution des compétences : variable

Hypothèse

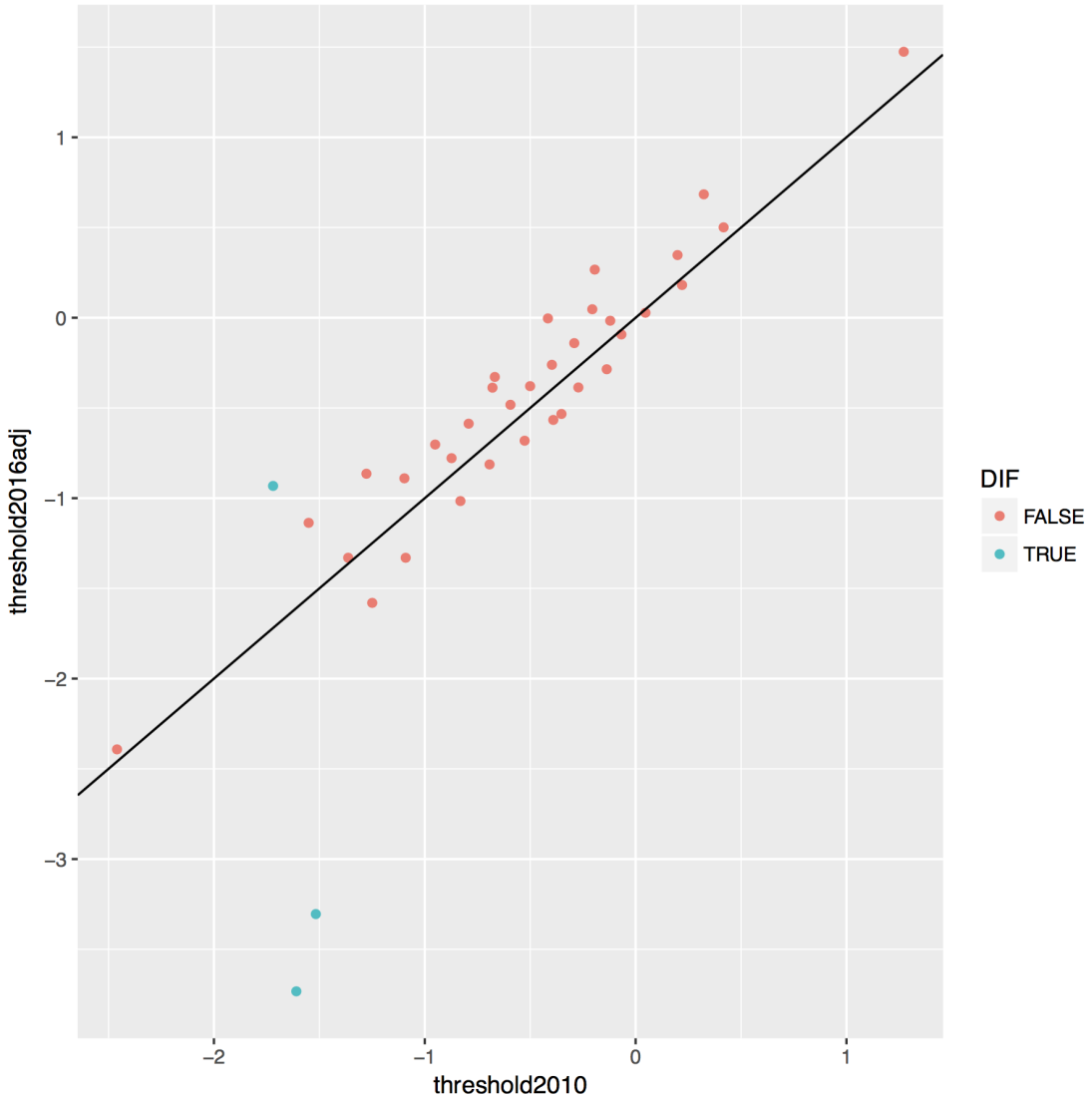
- le fonctionnement des items est identique d'une évaluation à l'autre
- en particulier : la hiérarchie de difficulté des items est inchangée

Ancrage


Distribution des sujets (θ)
Distribution des items (b)
10
15
5
14
7
7
6 9
2 11
11 13
8
1 3
3
4
4
12
16
18
groupe1
groupe 2

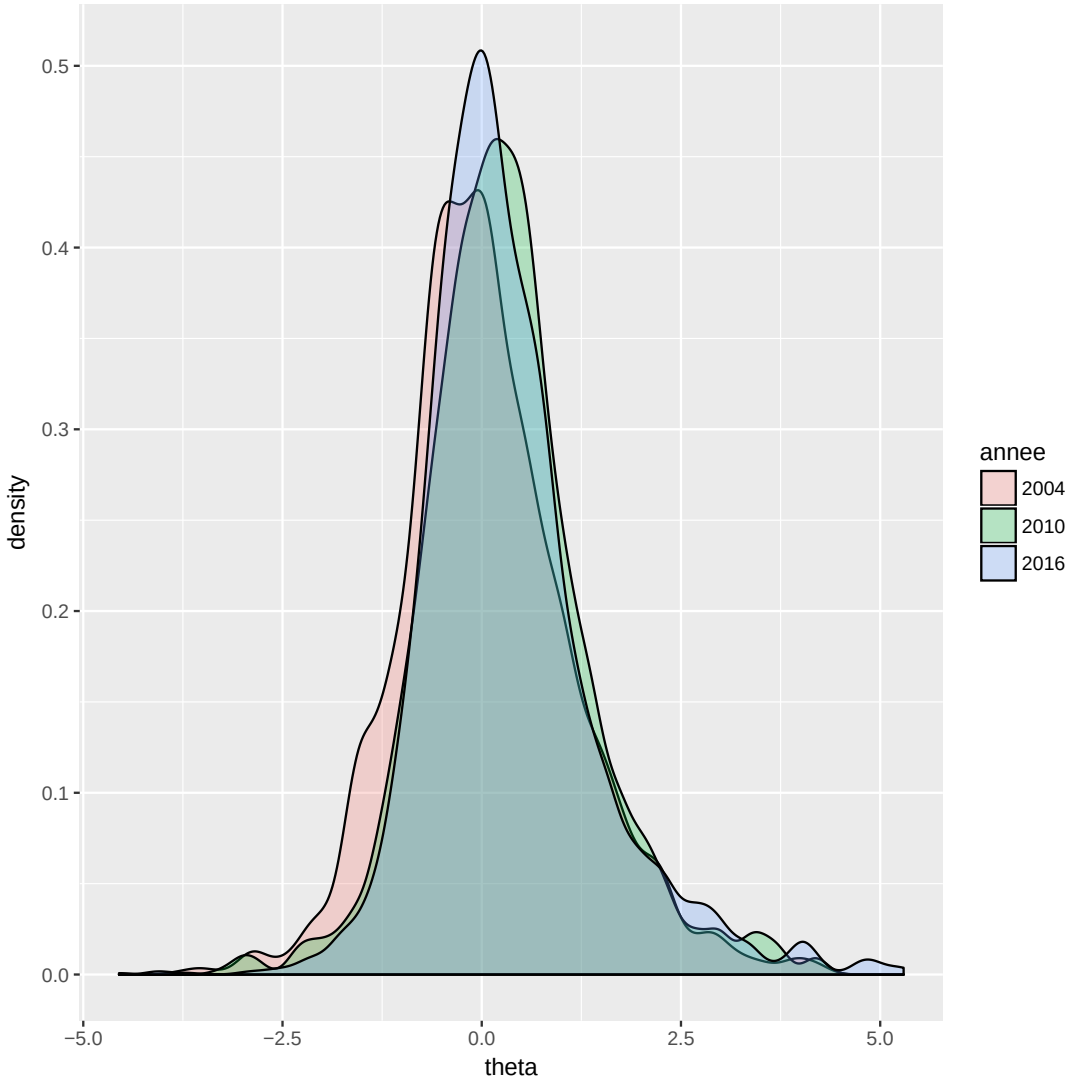
Fonctionnements différentiels d'items

Exemple CEDRE



Comparabilité des scores

Exemple CEDRE

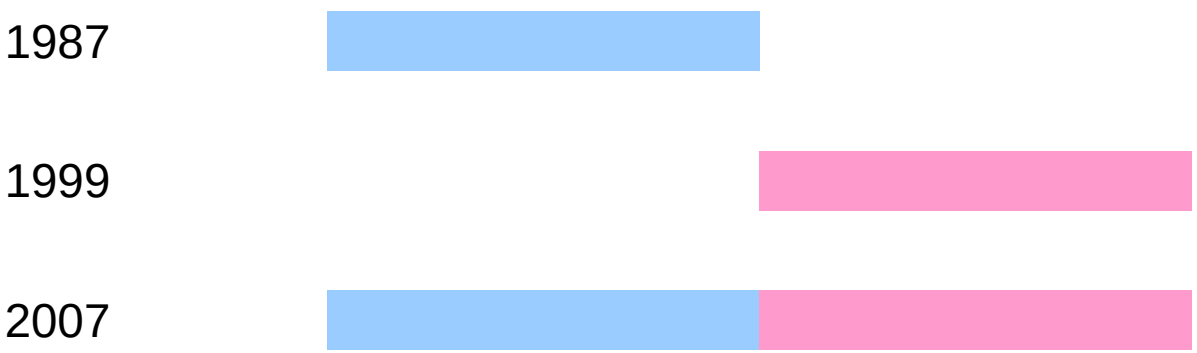


Evaluations cycliques

- Objectif : mesure de l'évolution dans le temps du niveau de compétence
- Principe : reprise à l'identique d'items communs ; la reprise complète peut s'avérer délicate (exposition des items, évolution des programmes, fonctionnement des items ...)
- Paramètres des items communs supposés fixes
- Application : évaluations cycliques (CEDRE, PISA, TIMSS ...)

Evaluations reprises

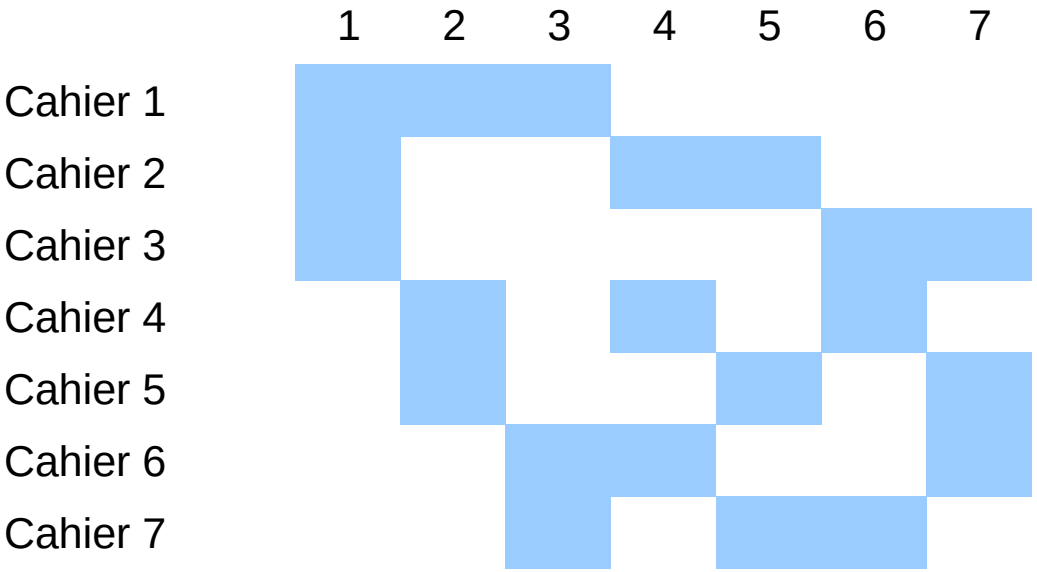
- Cas particulier : ancrage via des élèves
- Estimation des paramètres des items sur la même échelle rendue possible par les élèves ayant passé tous les items
- ≪ Lire, écrire, compter ≫ à vingt ans d'intervalle (CM2, 1987-2007), partie calcul : les trois années peuvent être comparées



Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
ministère éducation nationale

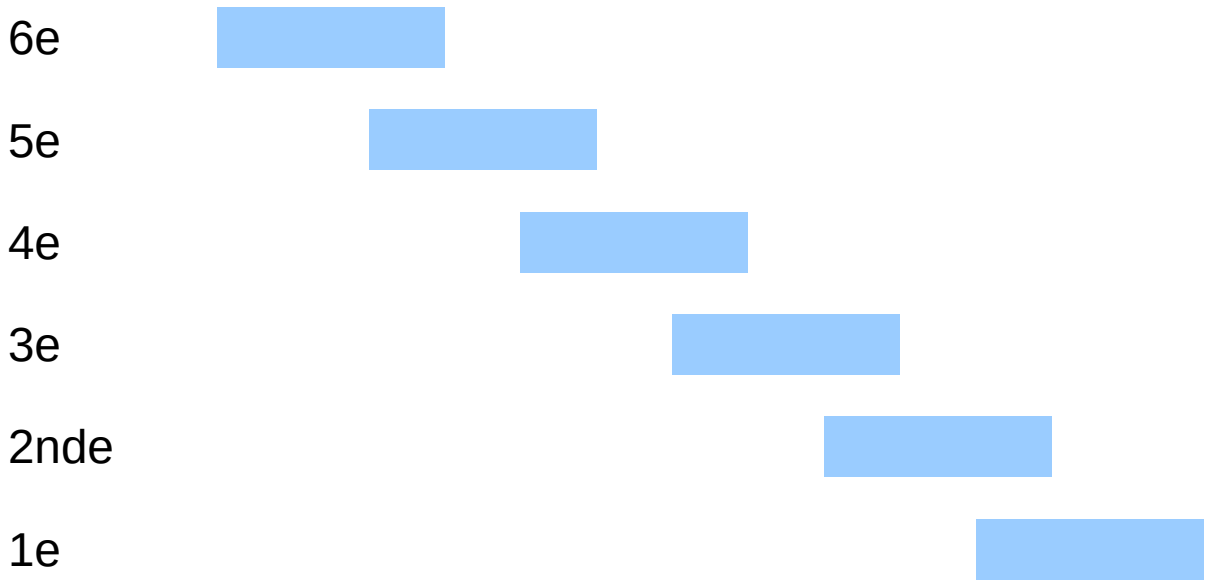
Cahiers tournants

- Objectif : évaluer de nombreux items sans augmenter le temps de passation
- Principe : découpage en « blocs », chaque paire de blocs est évaluée, contrôle de l'ordre de passation
- Les MRI prennent facilement en compte les valeurs manquantes aléatoires



Evaluations multiniveaux

- Objectif : une même échelle de « développement »
- Application : banque d'items, suivis de cohorte (ex : panel CP 2011)



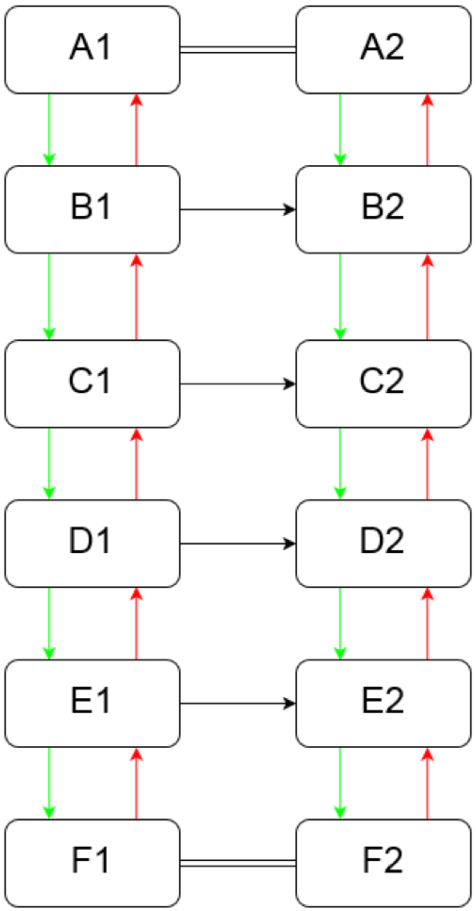
Renouvellement des items

- Application : évaluations CE1/CM2 (2009-2011), ou aussi tests de langue commerciaux
- Éviter les biais liés à l'≪ exposition ≫ des items : système de pré-tests et de post-tests

Tests adaptatifs

- Objectif : adapter la difficulté des items au niveau de compétence de l'individu

Application : handipanel



Autre exemple :

- Réponse $Y_i^j \rightarrow$ estimation $\hat{\theta}_i$ $\rightarrow$ proposition d'un item adapté (b_j proche de $\hat{\theta}_i$)

Pour conclure

Les MRI en résumé

- « All models are wrong, but some are useful »
- And these models are not so wrong ¨‿

Les MRI en pratique

- De nombreux avantages
- Un coût d'entrée

Pour conclure

Les MRI en résumé

- « All models are wrong, but some are useful »
- And these models are not so wrong ¨‿

Les MRI en pratique

- De nombreux avantages
- Un coût d'entrée

MERCI !