Comment assurer la comparabilité des niveaux de compétences de populations n'ayant pas passé les mêmes évaluations ?

Thierry Rocher

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

Journées de Méthodologie Statistique
Paris, 24-26 janvier 2012

Le problème

Illustration :

- Quelle évolution dans le temps du niveau des élèves ?
- Dispositifs de tests standardisés organisés en cycle
- Les tests ne sont pas identiques d'un moment de mesure à l'autre (changement de programmes, exposition des items, ...)
- Assurer la comparabilité de résultats obtenus à des évaluations différentes $\rightarrow$ nécessité d'un ajustement des métriques
- $4 \times 7 \neq 7 \times 4$

Présentation de méthodes à partir d'exemples

1. Evaluations sur échantillons (nationales/internationales, élèves/adultes)
2. Evaluations exhaustives

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

...

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)

7 2 8 5 6 1 4 3

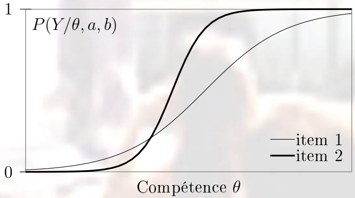
Alice Alex Jessica Bruno Rémi

Modèles de réponse à l'item

Modèle de réponse à l'item (2PL) :

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

Y_i^j : réponse de l'élève i à l'item j
θ_i : niveau de compétence de l'élève i
D : constante qui vaut 1.7
b_j : difficulté de l'item j
a_j : discrimination de l'item j



Séparation des concepts :

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

Indétermination

Modèle de réponse à l'item (2PL) :

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

Les paramètres sont définis à une trasnformation linéaire près :

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

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

Problème : comparer des élèves ayant passé deux évaluations différentes

Ancrage

L'estimation des niveaux de compétence est relative à chaque évaluation

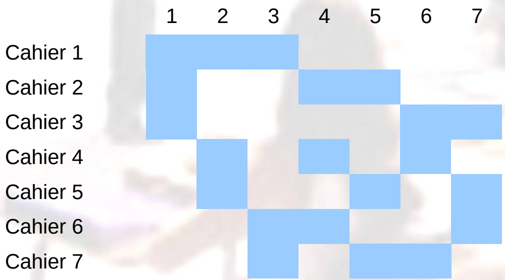
Ajustement des métriques (*equating*) via des items communs : positionner les niveaux de compétences sur la même échelle, grâce à des items repris à l'identique d'une évaluation à l'autre

De nombreuses méthodes ont été développées (littérature abondante en psychométrie)

Choix en fonction de l'architecture (*design*) et des contraintes de chaque évaluation

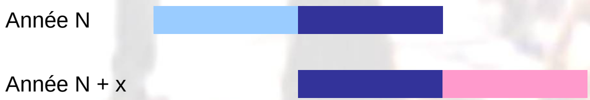
Cahiers tournants

- 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
- Application : évaluations d'élèves sur échantillons (PISA, CEDRE, LOLF-Socle, ...)



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 ...)
- Application : évaluations cycliques (PISA, CEDRE, IVQ, ...)



Evaluations reprises

- Cas particulier : ancrage via des élèves
- « Lire, écrire, compter » à vingt ans d'intervalle
- Calcul

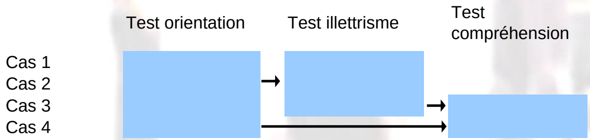
Evaluations multiniveaux

- Objectif : une même échelle de « développement »
- Application : banque d'items, suivis de cohorte (panels, internats d'excellence, ...)

Tests adaptatifs

- Objectif : adapter la difficulté de l'item au niveau de compétence de l'individu
- Application : tests sur supports électroniques (LSE), IVQ

IVQ :



Test adaptatif LSE :

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

Principe

Comparer des groupes d'élèves non équivalents, en terme de niveau de compétence, grâce à des items communs identique

D'une session d'évaluation à l'autre :

- paramètres des items : fixes
- niveau de compétence des élèves : variable

Hypothèse

- le fonctionnement des items est identique d'un moment de mesure à l'autre
- en particulier : la hiérarchie de difficulté des items est inchangée, sinon un autre facteur que θ a joué dans la réussite

Estimation

Estimation conjointe (≪ concourrante ≫) :

- Utilisation de toute l'information (toutes les sessions)
- Indétermination : on fixe μ_θ et σ_θ pour un des groupes

Limites :

- Contrainte sur la disponibilité des données (cf. PIAAC)
- Cohérence avec la diffusion des résultats et des données des sessions précédentes (ex : PISA, CEDRE)

Autre possibilité :

- fixation des paramètres des items estimés précédemment
- ex : IVQ-PIAAC

Estimations séparées

Les paramètres sont définis à une transformation linéaire près :

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

Un méthode « intuitive » :

$$\left| \begin{aligned} A &= \mu_a / \mu_{a^*} = \sigma_b / \sigma_{b^*} \\ B &= \mu_{b^*} - A\mu_b \end{aligned} \right.$$

A partir des items communs aux deux évaluations, deux procédures : *mean/mean* ou *mean/sigma*

Méthode de Stocking et Lord

On cherche les « meilleurs » A et B tels que les paramètres des items communs ($j \in C$) de la 2e évaluation, transformés sur l'échelle de la 1ère (a^*_{j2}, b^*_{j2}), soient les plus proches possibles des paramètres estimés pour la première évaluation (a_{j1}, b_{j1}).

Pour cela, on minimise l'écart entre les probabilités de réussir les items communs, à niveau de compétence égal :

$$min \sum_i (\xi_i - \xi^*_i)^2$$

où
$\xi_i = \sum_{j \in C} P(\theta_i, a_{j1}, b_{j1})$
et
$\xi^*_i = \sum_{j \in C} P(\theta_i, a^*_{j12}, b^*_{j2})$

Evaluations nationales CE1/CM2

Contraintes :

- Changement complet des épreuves chaque année, pour éviter le bachotage
- Conception et sélection des items sous contraintes
- Scores calculables localement par les enseignants
- Le « rendu » est sous forme de scores bruts
- Fournir une grille de correspondance entre les scores obtenus l'année N et ceux obtenus l'année N-1

Le cadre précédent (*common-items non-equivalent groups design*) n'est plus adapté.

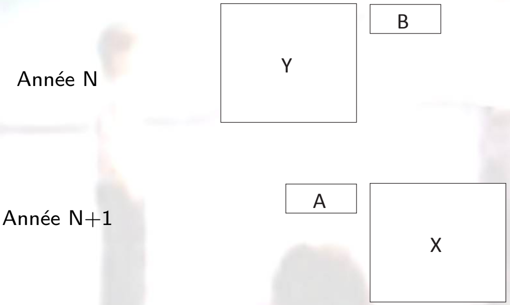
Nouveau *design*

Pré-tests et post-tests :

Année N

Année N+1

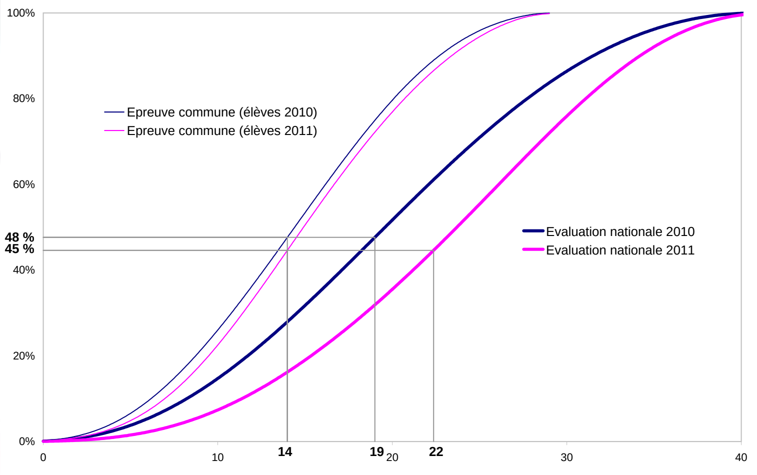
Stratégies



Deux stratégies d'ajustement (A et B)

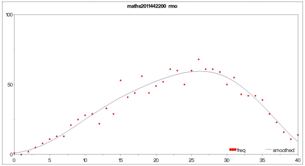
- A interne à Y, externe à X
- B externe à Y, interne à X

Equipercentile chaîné (*Chained Equipercentile*)

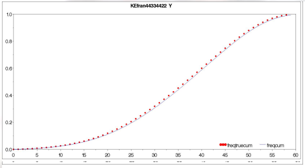


Procédure *Kernel equating* (Von Davier et al., 2004)

1. Lissage des distributions
   Modèle log-linéaire



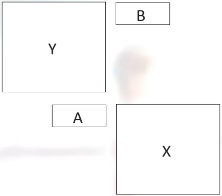
2. Continuité des fonctions de répartition
   Méthode du noyau (non paramétrique)



3. Equating

$$\hat{e}_Y(x) = \hat{G}_{2010}^{-1}\left(\hat{H}_{2010}\left(\hat{H}_{2011}^{-1}\left(\hat{F}_{2011}(x)\right)\right)\right)$$

Cas de la stratégie B



- Problème de la stratégie A : l'exposition des items
- Problème de la stratégie B : les items de l'année N proviennent d'items testés l'année N-1 sur des échantillons différents
  $\Rightarrow$ Il n'est pas possible de calculer directement un score brut sur les items B l'année N-1
  $\Rightarrow$ Reconstruction de la distribution des scores observés sur B à partir des paramètres d'un MRI (2PL)

MRI - Scores observés

Idée : « reconstruire » la distribution des scores observés sur B à partir des paramètres des items estimés selon un modèle MRI (2PL)

Si on note r le nombre d'items répondus, x le score observé et p_r la probabilité de réussir l'item r, on a la relation de récurrence suivante concernant la probabilité d'obtenir un score de x sur r items à un certain niveau de compétence θ fixé :

$$\begin{cases} f_r(x/\theta) = f_{r-1}(x/\theta)(1 - p_{r/\theta}) & \text{si } x = 0 \\ f_r(x/\theta) = f_{r-1}(x/\theta)(1 - p_{r/\theta}) + f_{r-1}(x - 1/\theta)p_{r/\theta} & \text{si } 0 < x < r \\ f_r(x/\theta) = f_{r-1}(x - 1/\theta)p_{r/\theta} & \text{si } x = r \end{cases}$$

Cette méthode permet d'estimer une distribution de scores bruts à partir d'items qui n'ont pas été passés par les mêmes élèves

Ajustement linéaire (Tucker)

Population *synthétique* S : année N et année N+1

Sur S, les scores centrés-réduits sont égaux pour les deux tests, après avoir ajusté les scores X sur l'échelle de Y

$$l_Y(x) = \frac{\sigma_S(Y)}{\sigma_S(X)}\left[x - \mu_S(X)\right] + \mu_S(Y)$$

où $l_Y(x)$ est l'ajustement du score x au test X sur l'échelle des scores Y, μ la moyenne des scores et σ leur écart-type.

Deux hypothèses pour calculer les μ et les σ :

- Régression de X en A - respectivement de Y en A - est la même pour les deux populations visées (N et N+1)
- Covariances conditionnelles $cov(X/A)$ et $cov(Y/A)$ sont également identiques quelle que soit la population

Rasch (1)

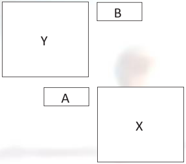
Une approche « classique » consiste à utiliser les propriétés du modèle de Rasch :

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

où Y_i^j est la réponse de l'élève i à l'item j, θ_i le niveau de compétence de l'élève i, b_j le niveau de difficulté de l'item j

Une propriété intéressante de ce modèle est l'exhaustivité du score brut $S_i = \sum_j (Y_i^j = 1)$, c'est-à-dire que θ_i et S_i sont liés par une relation bijective

Rasch (2)



Exemple stratégie A :

1. Estimation des b_j^A et des θ_i^Y à partir des données Y
2. Estimation des b_j^X du test X à partir des données $(X \cup A)$ en fixant les paramètres b_j^A
3. Estimation des θ_i^X du test X à partir des données X en fixant les paramètres b_j^X estimés au point précédent
4. Deux séries : (θ_i^Y, S_i^Y) et (θ_i^X, S_i^X) mise en équivalence des S_i^Y et les S_i^X à partir des θ_i les plus proches

Rasch (3)

- Avantage : simplicité de mise en œuvre

- Inconvénient : adéquation au modèle de Rasch

  Hypothèse très forte : discriminations des items homogènes.

  Cette hypothèse est si forte qu'en général, les items sont sélectionnés de manière à ce qu'ils soient conformes à ce modèle.

Ajustement MRI - Scores vrais

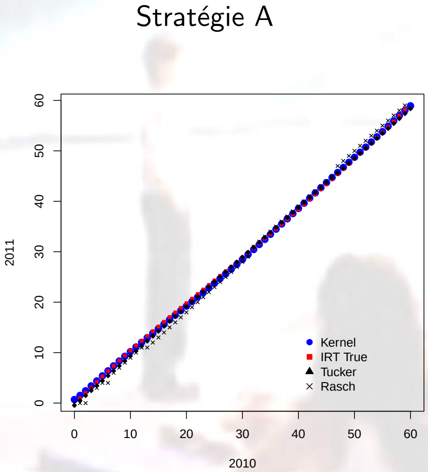
Relation entre « scores vrais » = relation entre scores observés

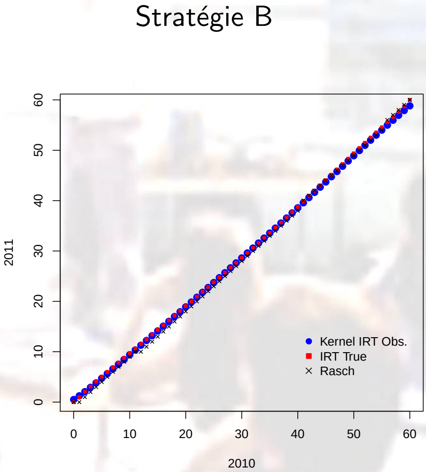
1. Estimation conjointe MRI (2PL) de tous les paramètres des items des tests (X, Y, A)
2. Pour un score vrai - nombre entier - fixé T_{Y_i} obtenu au test Y, on cherche le θ_i qui conduit à ce score.
   Le « score vrai » (au sens des MRI) :

   $$T_{Y_i} = \sum_{j \in Y} P_j(\theta_i)$$

   Equation non linéaire (algorithme de type Newton-Raphson)
3. A partir du θ_i trouvé, on calcule le score vrai aux items du test X.

Français


Stratégie A
2011
2010
0
10
20
30
40
50
60
Kernel
IRT True
Tucker
Rasch


Stratégie B
2011
2010
0
10
20
30
40
50
60
Kernel IRT Obs.
IRT True
Rasch

Mathématiques

Stratégie A