Principes de la théorie des tests

21 août 2017

Brève histoire des tests

Naissance des ≪ tests mentaux ≫ il y a 100 ans :

- En France (1905), avec le test de Binet pour détecter les difficultés d'apprentissage (ancêtre du QI)
- Aux USA (1904) avec Spearman et la mesure de l'intelligence (analyses factorielles, facteur g)
- Histoire des tests : Gould (1997), ≪ The big test ≫ (1999)

Un mot sur la docimologie :

- Enquête Carnegie (1936) sur les notes au baccalauréat
- Nécessité d'une mesure objective, standardisée

Les « niveaux » de mesure

- l'**échelle nominale** : la mesure est un simple ensemble d'étiquettes (exemple : la nomenclature des PCS) ;
- l'**échelle ordinale** : elle impose l'existence d'une relation d'ordre (exemples : les classes d'âge, les niveaux scolaires, ?)
- l'**échelle d'intervalle** : les écarts entre les valeurs ont une signification (exemples : les températures °C ou °F, la position sur un axe de coordonnées, la plupart des variables psychologiques, ?)
- l'**échelle de rapport** : non seulement les écarts, mais aussi les rapports, doivent être comparables. Cela impose l'existence d'un 0 absolu (exemples : multiplication par une constante pour passer du mètre à l'inch, de la livre au kilo, de la seconde à l'année, de l'euro au dollar).

Une caractéristique latente

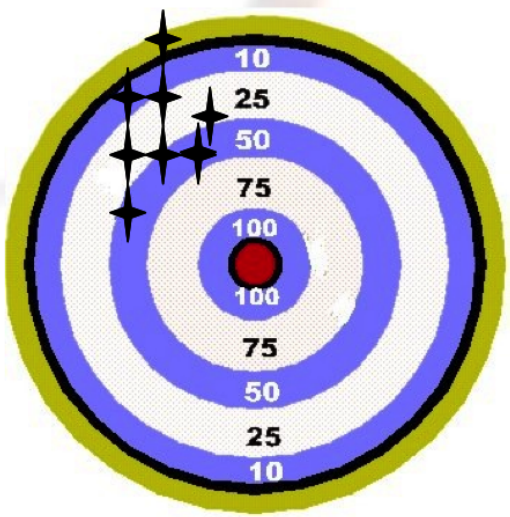
Le processus de mesure s'appuie sur des observations manifestes pour réaliser la mesure d'une caractéristique latente
Le Saint Graal de la psychométrie :

- Performance (score observé) = Compétence latente (score ≪ vrai ≫) + erreur de mesure
- L'objet de la mesure est un construit (≪ c'est ce que mesure mon test ≫)

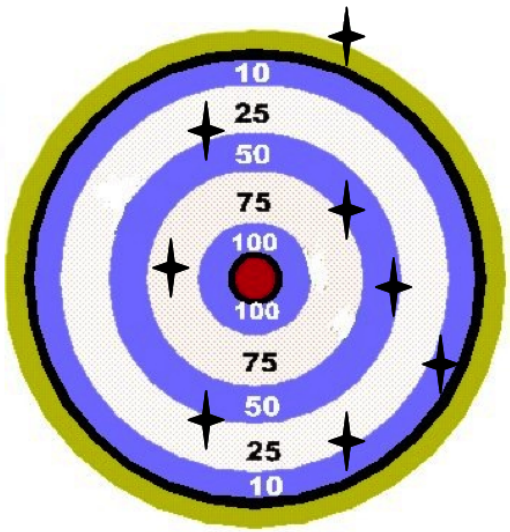
Deux questions :

- Mesure-t-on ce que l'on souhaite mesurer ? (validité)
- Avec quel degré de précision ? (fidélité)
- L'objectif de la psychométrie est de construire une mesure fiable, c'est-à-dire fidèle et valide.

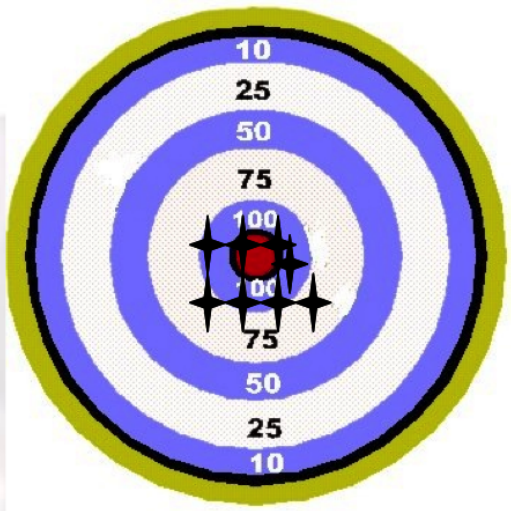
Une mesure fidèle et valide



A



B



C

La validité

La validité correspond au degré de signification de l'instrument de mesure (test) et des données qu'il permet d'obtenir

C'est le point crucial de l'évaluation d'un instrument

Différents types de validité :

- Validité apparente
- Validité de contenu
- Validité critériée convergente
- Validité critériée prédictive

La validité (suite)

Contexte d'utilisation	Question posée	Type de validité
Evaluation des apprentissages	A quel point l'élève maîtrise-t-il la matière présentée dans son enseignement de maths ?	De contenu
Prédiction de la performance ultérieure	Comment l'élève réussira-t-il en maths au secondaire	Critériée prédictive
Diagnostic des difficultés d'apprentissage	Quelles sont les difficultés spécifiques de l'élève en maths ?	Critériée convergente
Evaluation des habiletés mathématiques	Quelles sont les habiletés mathématiques de l'élève	De construit

La fidélité

Fidélité = reproductibilité des résultats obtenus à ce test
Méthodes :

- Il existe différentes méthodes pour évaluer la fidélité d'un test :
- La méthode des formes parallèles
- La méthode test-retest
- La méthode du split-half
- La consistance interne : α de Cronbach renvoie à la moyenne des corrélations entre items Il augmente avec le nombre d'items « Règle » habituelle : $\alpha > 0.70$ = fidélité acceptable

Quel(le)s test(évaluation)s ?

Fonction du test (à quoi ça sert ?) :

- à sélectionner des élèves
- à diagnostiquer des difficultés d'apprentissage
- à contrôler si les compétences attendues sont maîtrisées
- à déterminer si les élèves ont les compétences pour suivre une formation particulière ?

Utilité du test (à qui ça sert ?) :

- à l'élève pour obtenir une certification (test certificatif)
- à l'élève pour apprendre (évaluation formative)
- à l'enseignant pour connaître ses élèves (évaluation diagnostique)
- aux politiques pour piloter le système

L'élaboration des tests

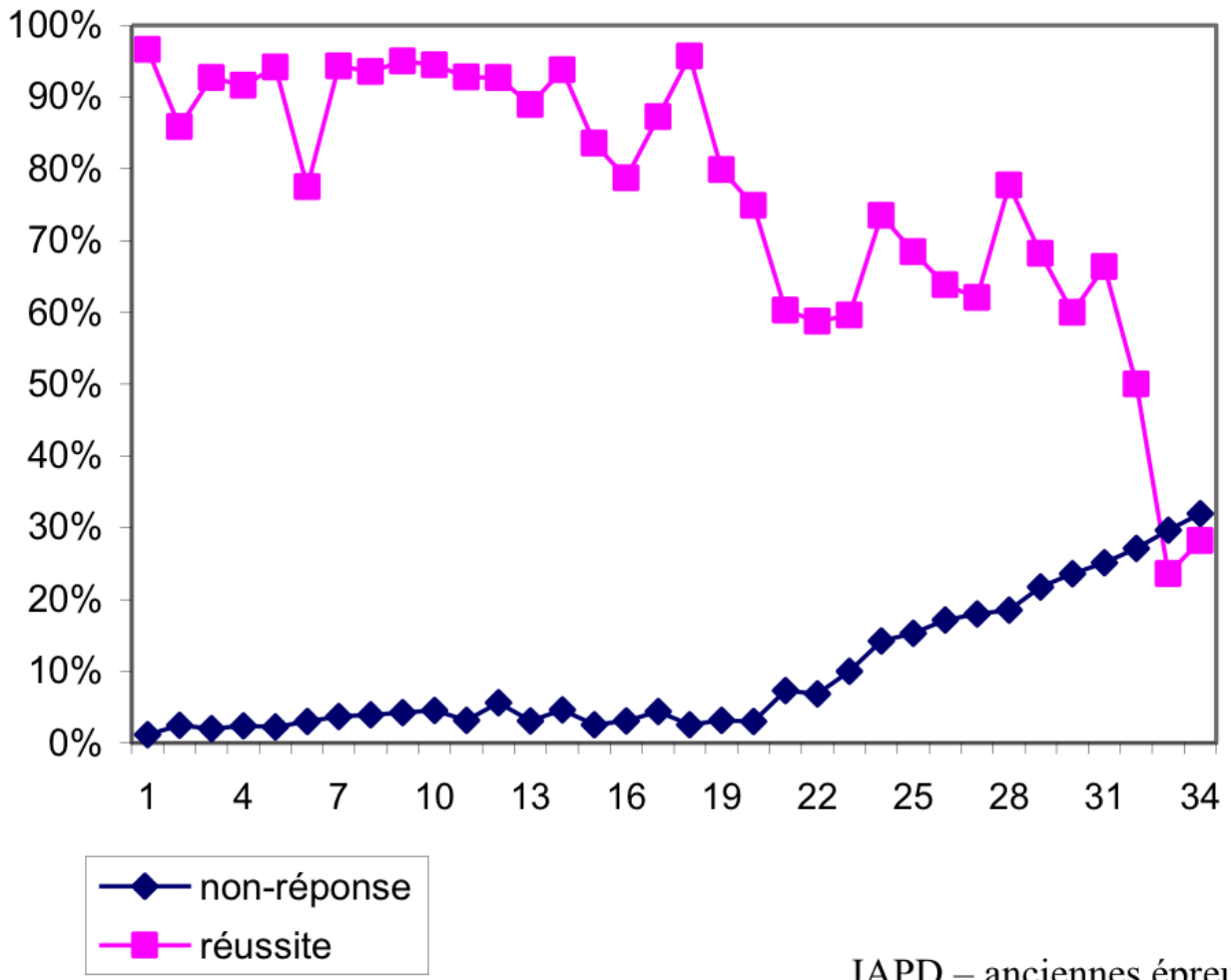
Quelques grands principes :

- Donner des consignes claires et précises
- Garantir au maximum l'indépendance locale : limiter si possible le nombre d'items relatifs au même stimulus
- Maîtriser les paramètres des supports : lisibilité, longueur, familiarité, etc.
- Varier la difficulté des items (sauf exceptions, cf. base)
- Soigner l'édition

Quelques effets connus

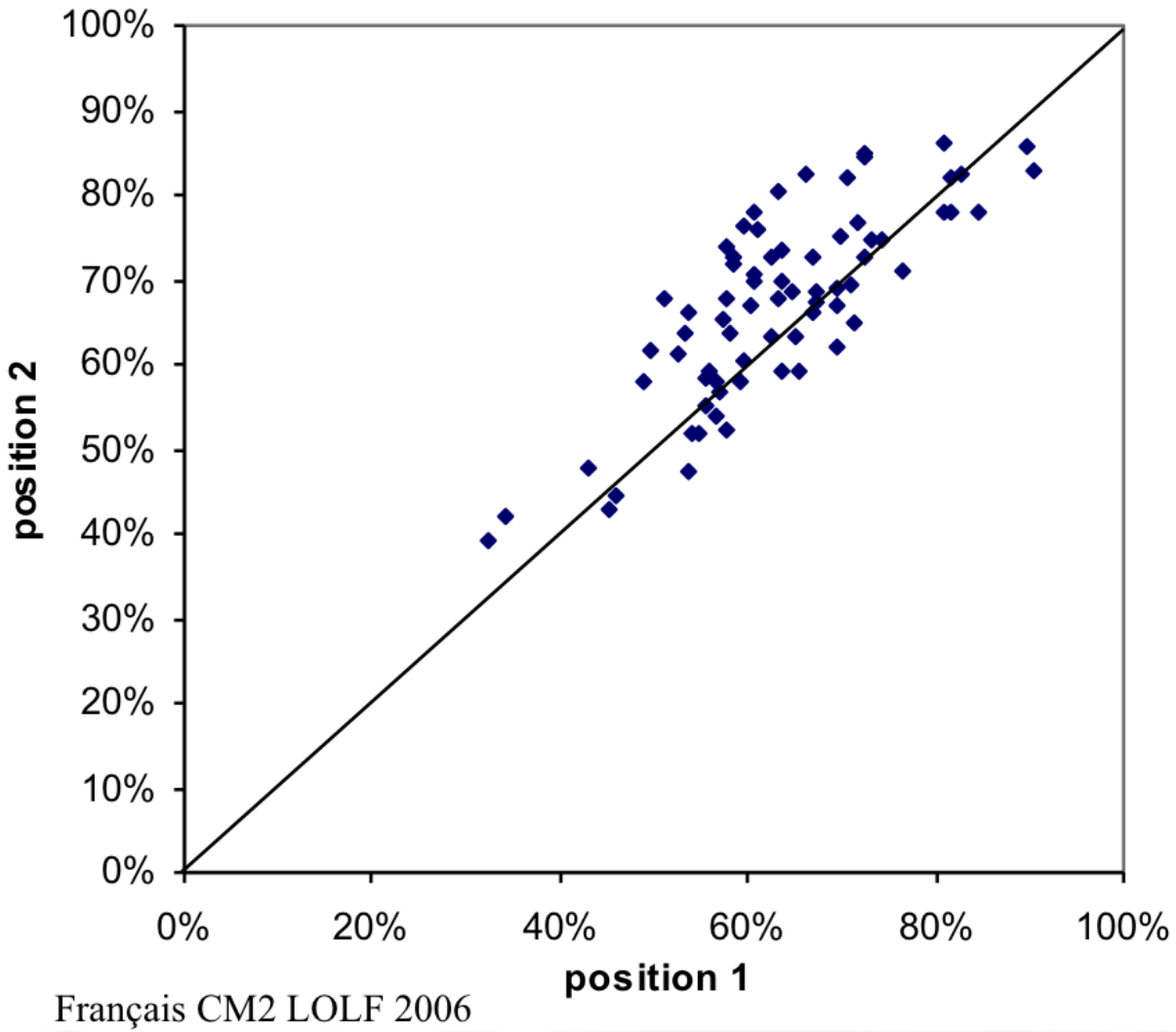
- L' « environnement » des items
- Exemples : évaluation 2nde (item de mathématiques dans une évaluation de géographie), PISA (mineures/majeures)
- L'effet de fatigue
- L'effet d'entraînement

Brève histoire des tests



JAPD – anciennes épreuves

Brève histoire des tests



Français CM2 LOLF 2006

Mesurer une variable latente

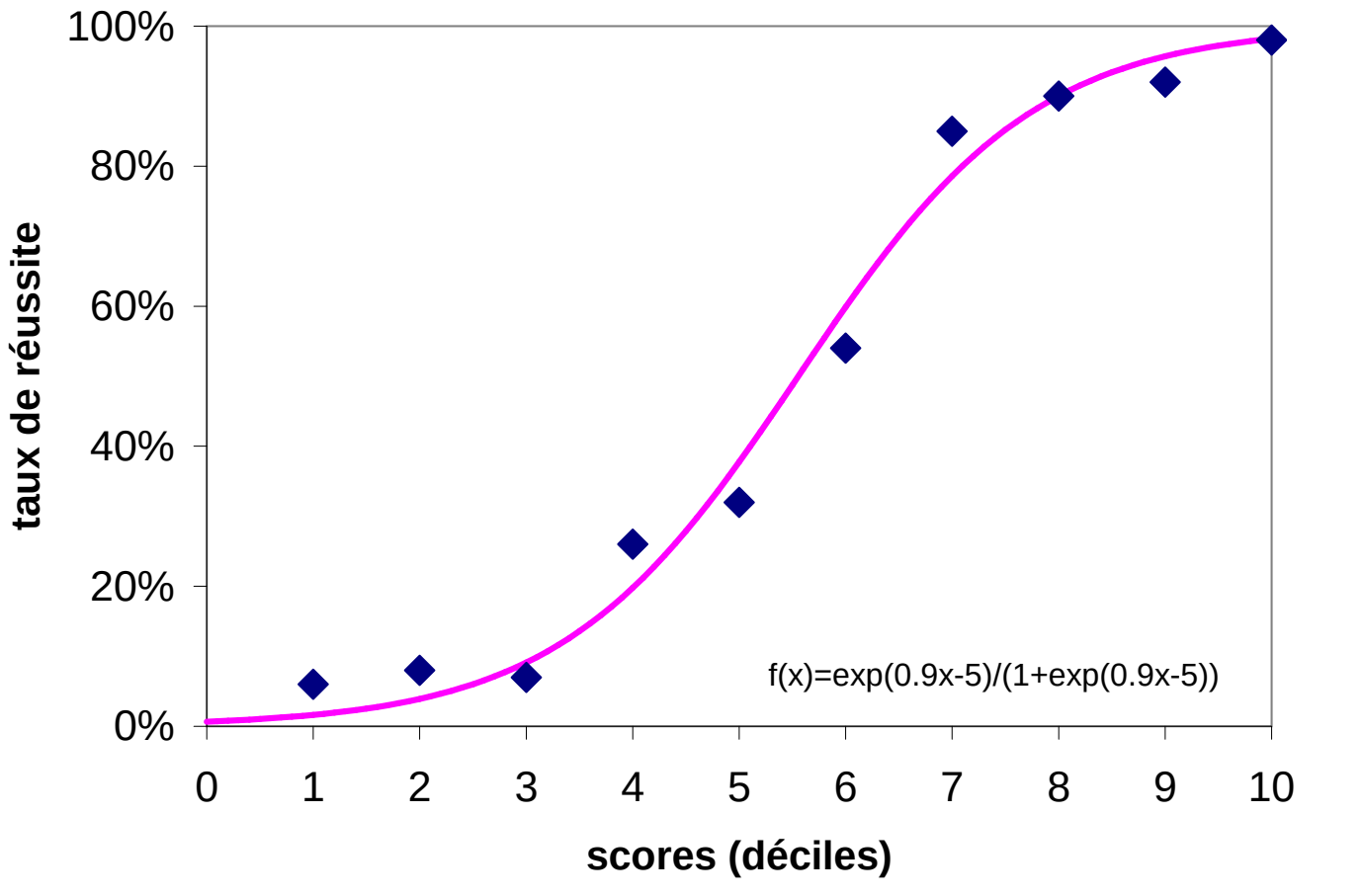
Mesurer la taille des individus... avec un questionnaire

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

...

Modélisation

Ajuster un modèle aux données



Modèles de réponse à l'item

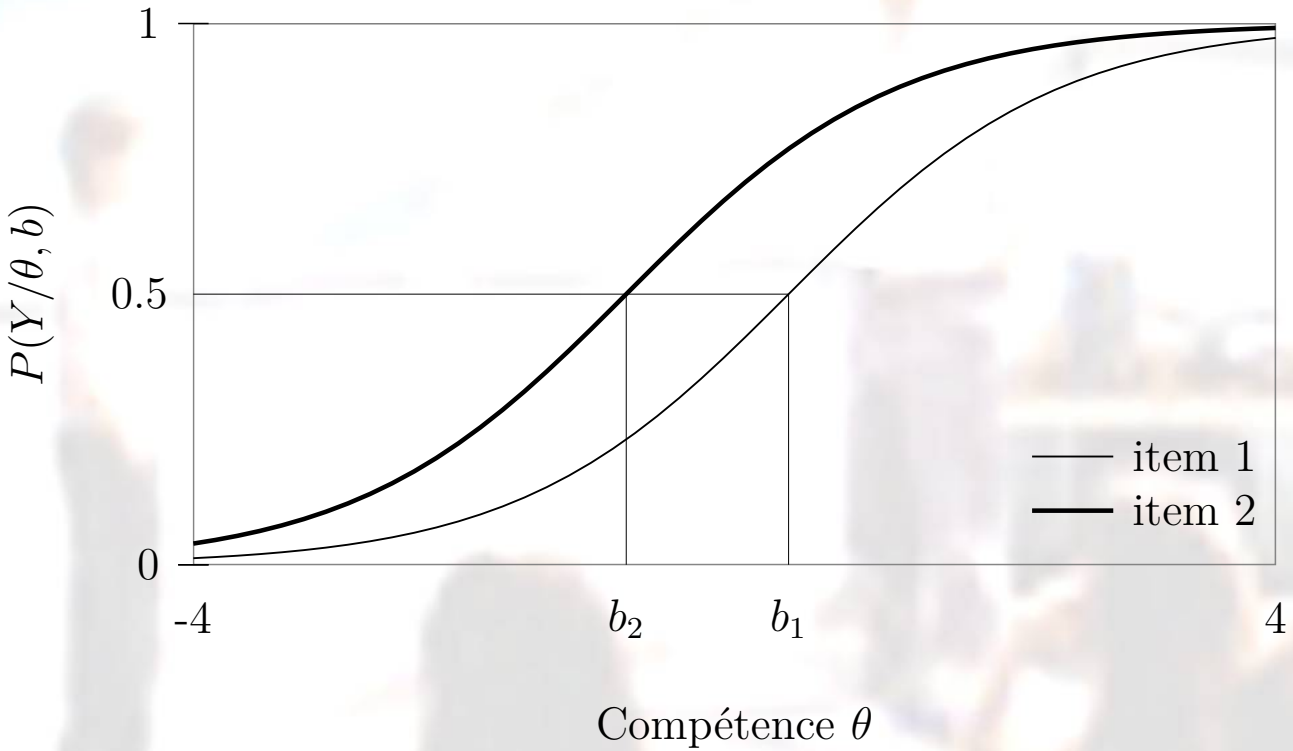
De manière générale :

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

- 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

DEPP : MRI unidimensionnels à deux paramètres avec fonction de lien logisitique (2PL) pour des items dichotomiques (bientôt *graded*)

Modèle de Rasch (1960), 1 paramètre



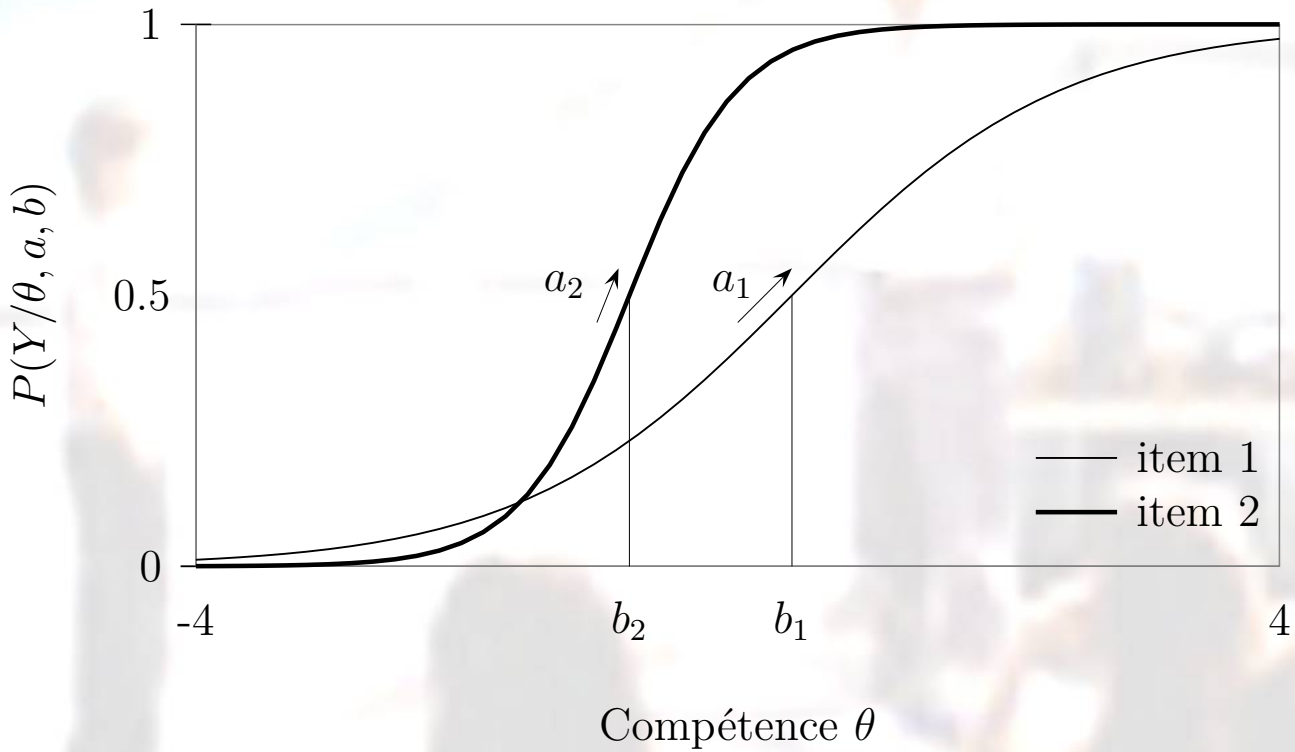
$$P(Y_i^j = 1/\theta_i, b_j) = \frac{exp(\theta_i - b_j)}{1 + exp(\theta_i - b_j)}$$

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

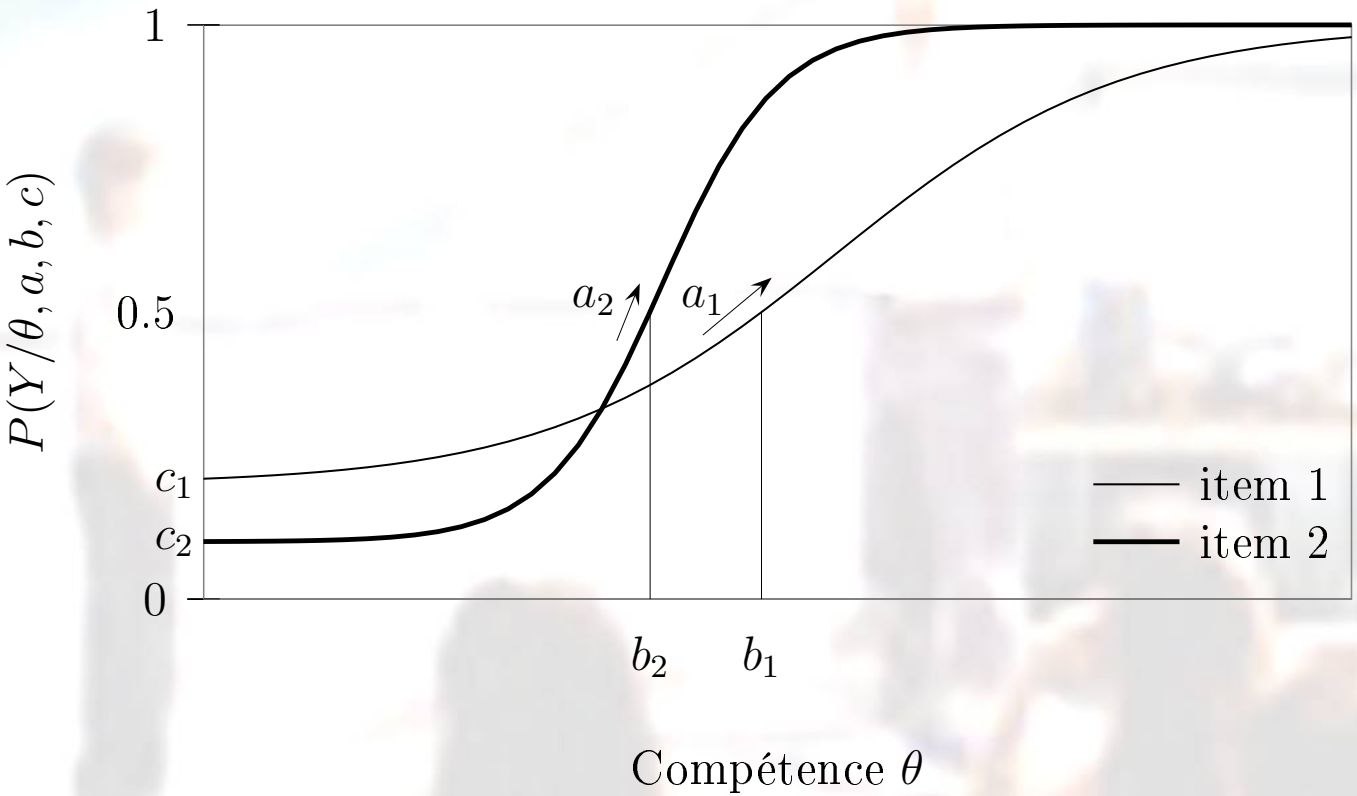


$$P(Y_i^j = 1/\theta_i, a_j, b_j) = \frac{exp(a_j(\theta_i - b_j))}{1 + exp(a_j(\theta_i - b_j))}$$

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

Modèle à 3 paramètres



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

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
c_j : paramètre de pseudo-chance j

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

Ancrage

- l'estimation des niveaux de compétence est relative à chaque évaluation.
- ajustement des métriques (*equating*) via des items communs

Equating

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
- niveaux de compétence des élèves : variables

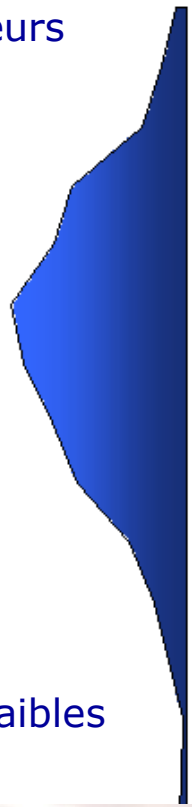
Hypothèse

- 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
- unidimensionnalité ?

Echelle

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

élèves les meilleurs



items difficiles

10

5

7

6 9

2 11

8

1 3

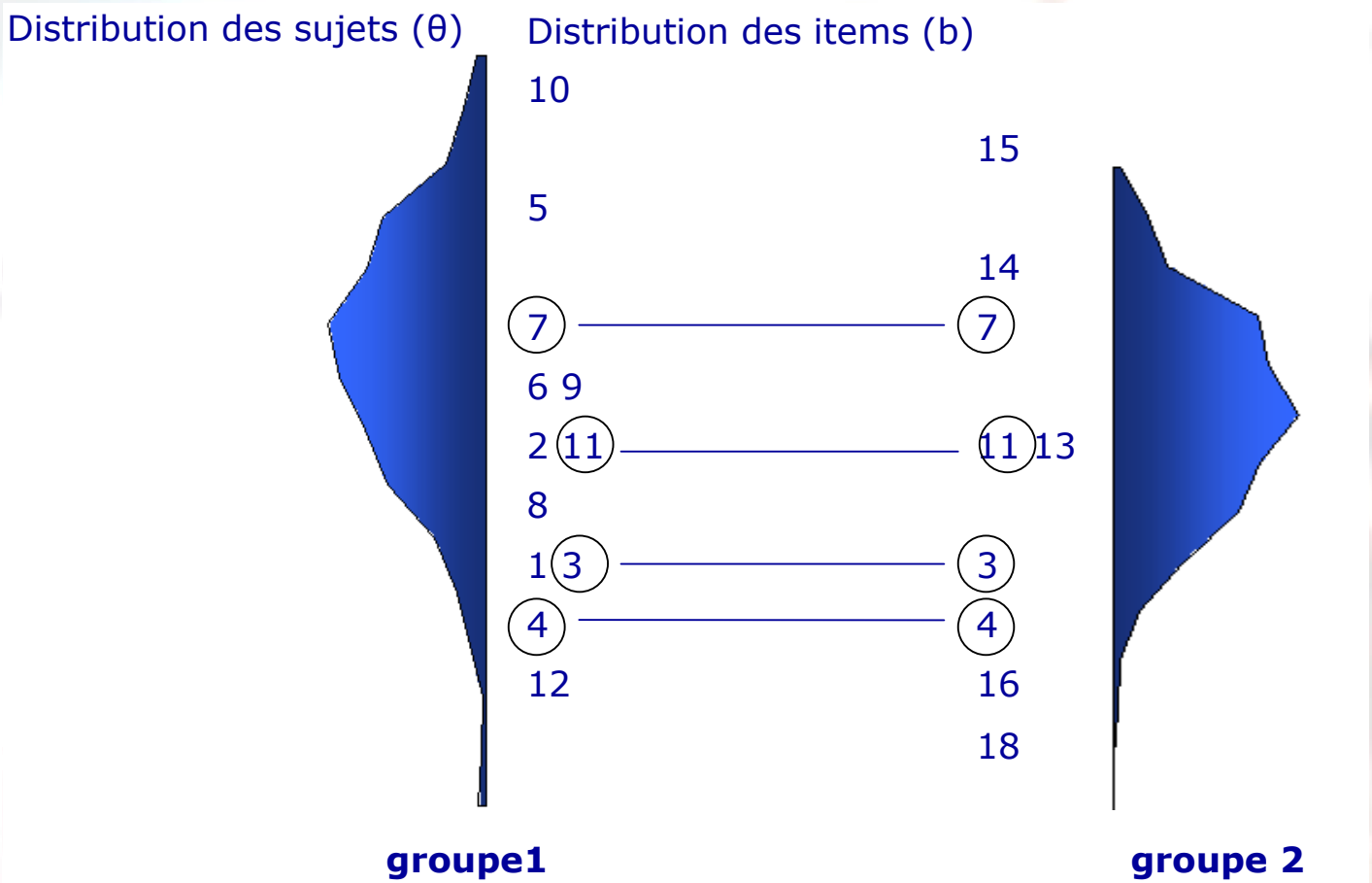
4

12

élèves les plus faibles

items faciles

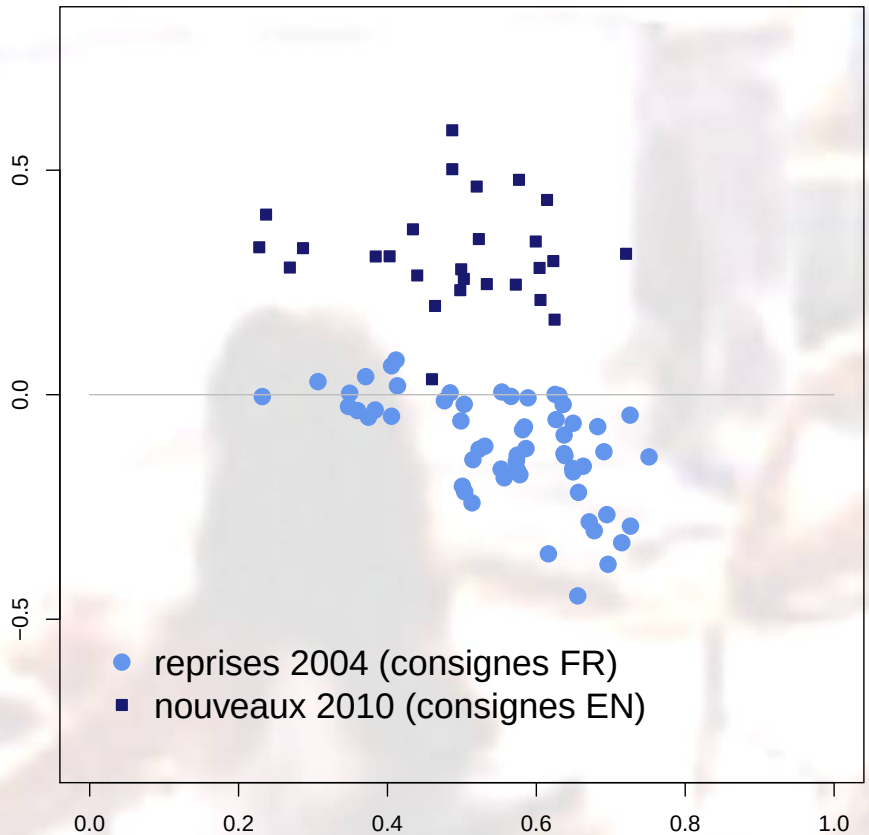
Ancrage


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

Unidimensionnalité

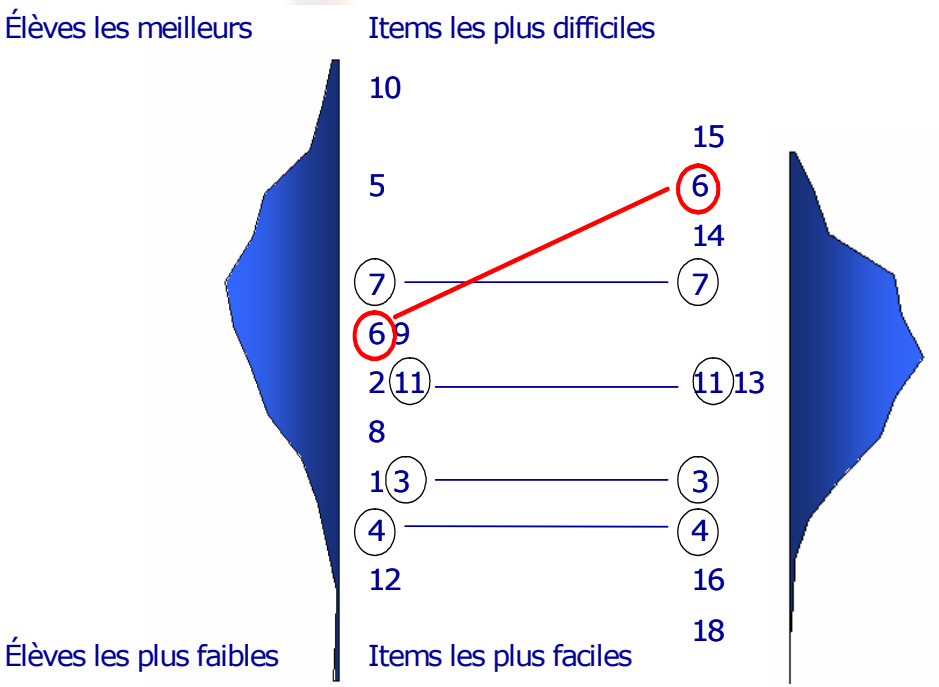
Seule la compétence θ explique la réussite à l'item

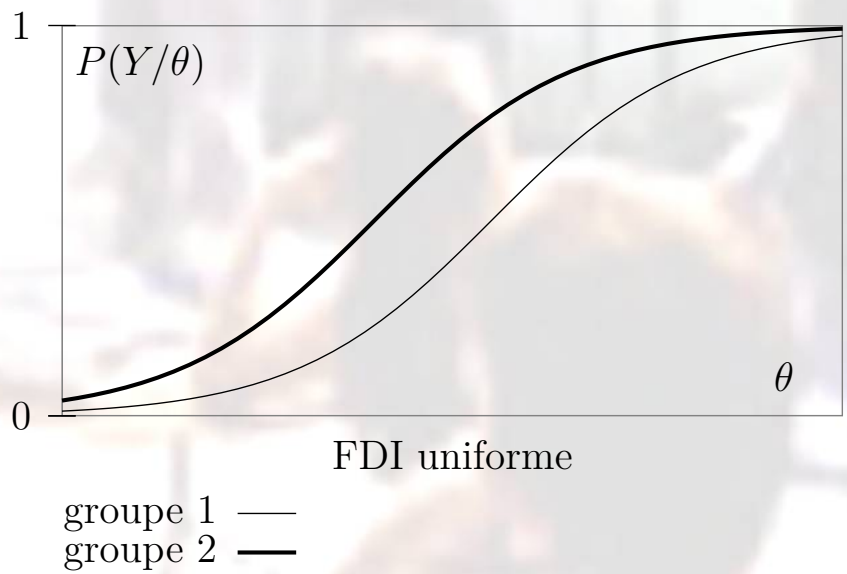
$\rightarrow$ Contre-exemple (format des items CEDRE anglais 3e 2010)



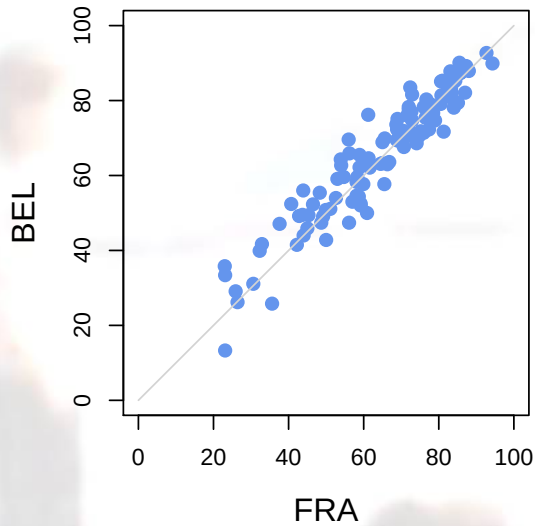
Fonctionnements différentiels

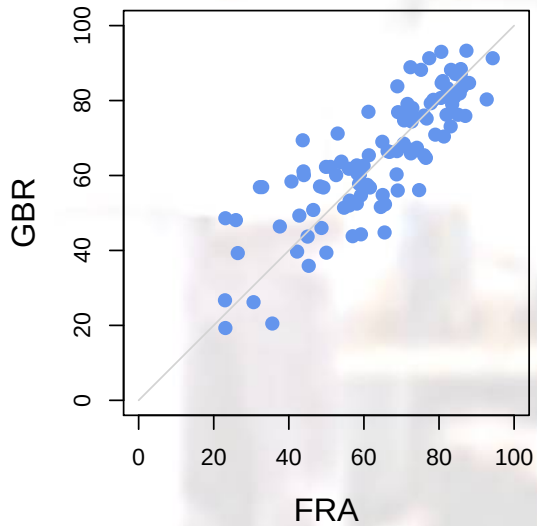
Un item est dit biaisé selon une caractéristique de groupe d'élèves (garçons/filles, français/anglais, etc.) dès lors qu'on observe, à niveau de compétence égal, des réussites différentes selon les groupes : un autre facteur que celui visé est impliqué dans la réussite.

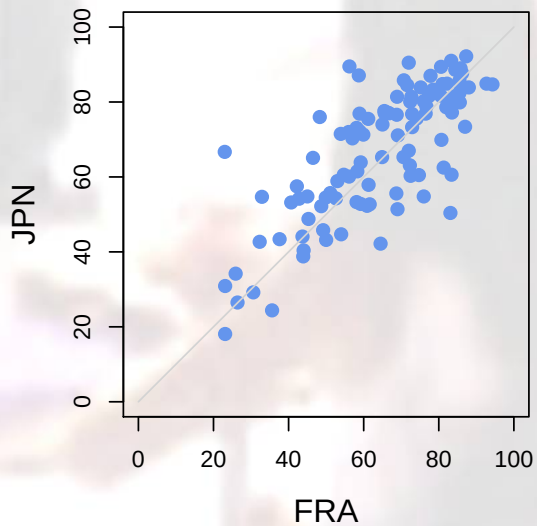


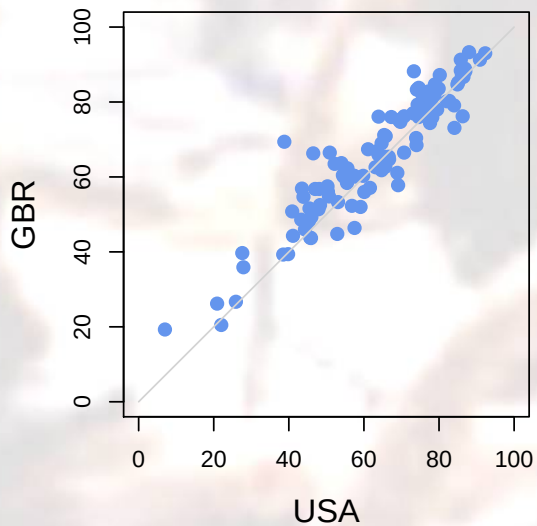


Exemple : PISA

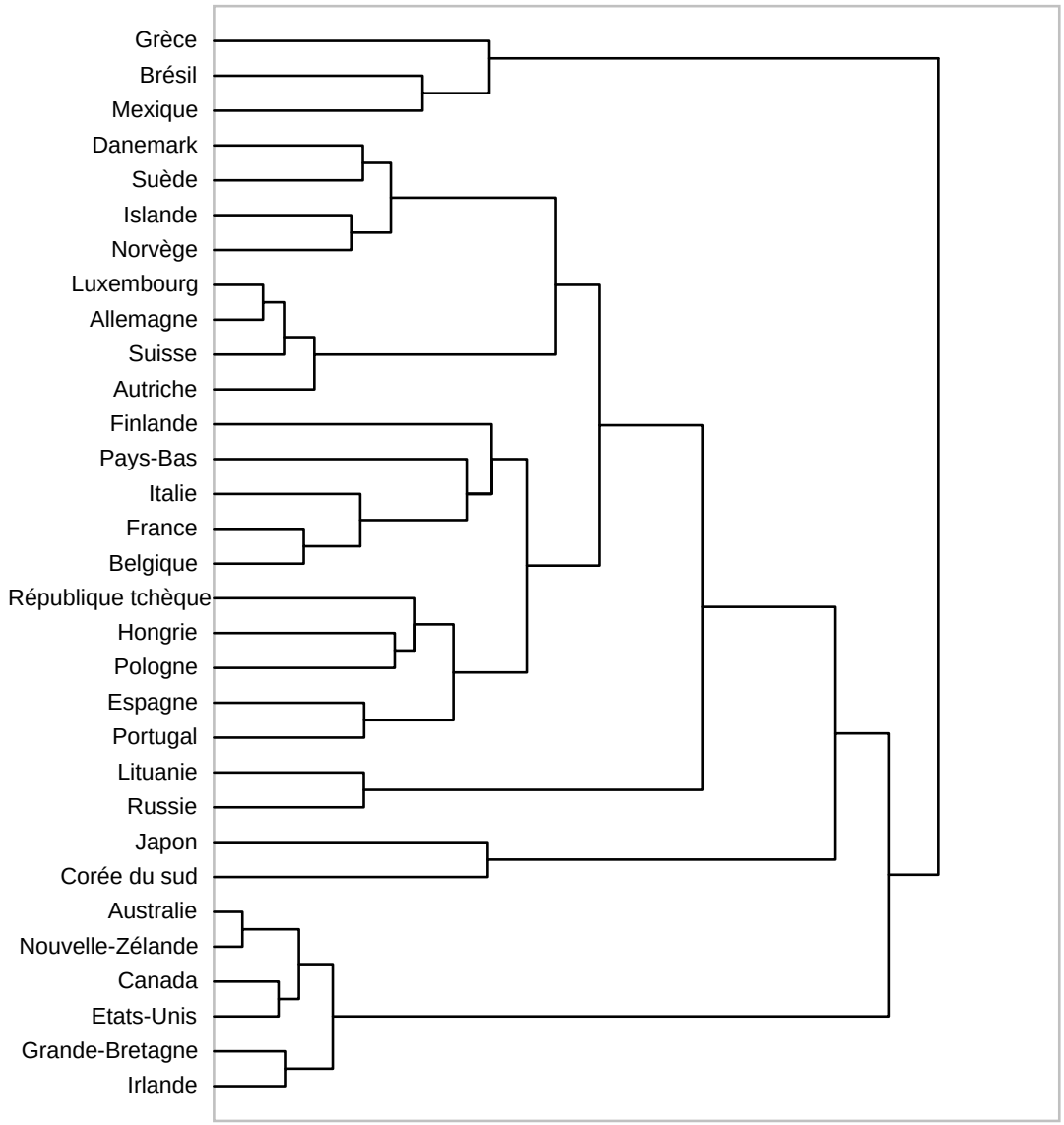






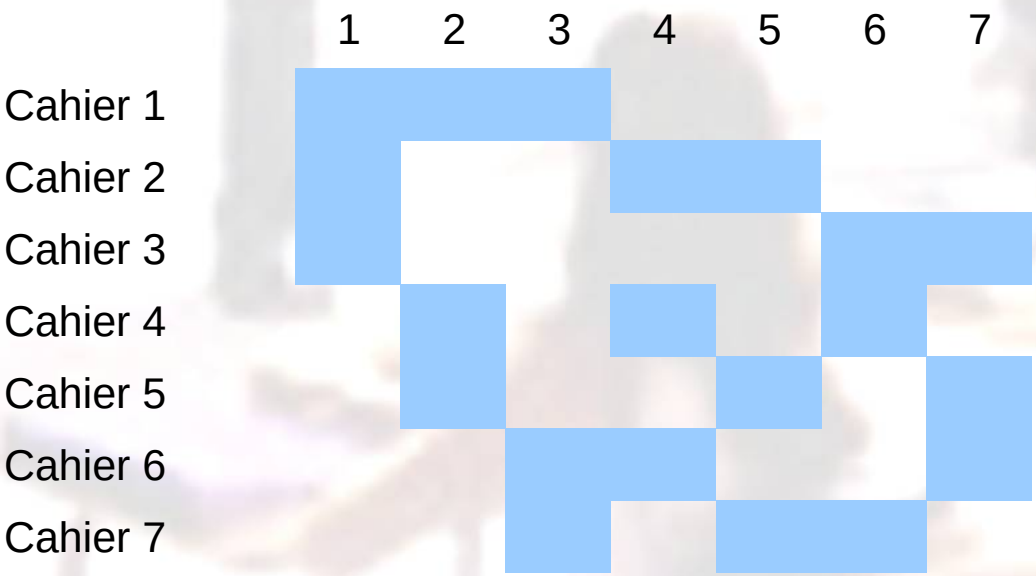


Exemple : PISA



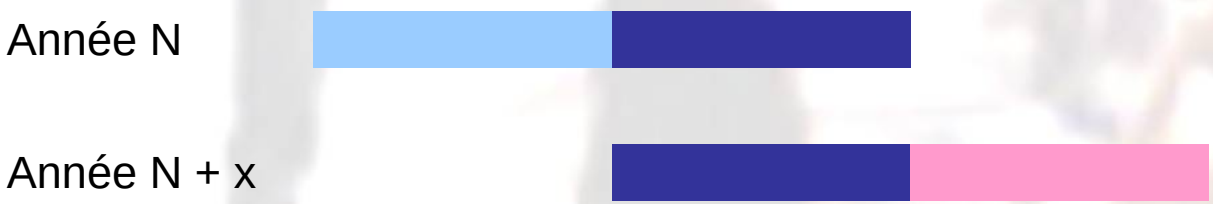
Cahiers tournants

- 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, ...)



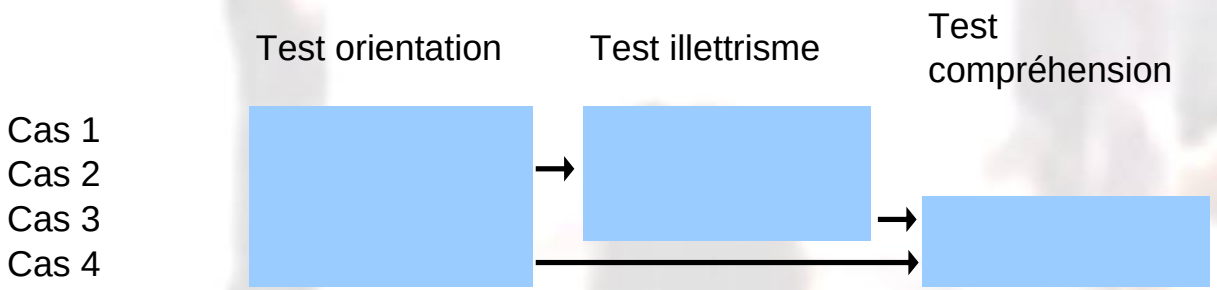
Evaluations multiniveaux

- 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), supports papier (IVQ, ESLC)

IVQ :



Test adaptatif LSE :

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

TEST ADAPTATIF ET HANDICAP: DE NOUVELLES MÉTHODES POUR L'ÉVALUATION DES POLITIQUES PUBLIQUES

R. Dos Santos, T. Rocher

Ministère de l' Education, France

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE L'ÉDUCATION NATIONALE,
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE

aea
EUROPE
November 3 - 5, 2016
Limassol, Cyprus

CONTEXTE

- En France, la loi sur les personnes handicapées du 11 février 2005 oblige à assurer l'inscription de tous les enfants handicapés dans le système scolaire ordinaire, aussi près que possible de leur domicile, afin d'assurer la continuité de leurs études et l'égalité des chances aux examens.
- Dix ans plus tard, le temps est venu pour une évaluation de cette politique: les enfants en situation de handicap bénéficient-ils de cette politique inclusive d'éducation?
- Afin d'observer leur inclusion réelle au cours de leur carrière scolaire, une enquête longitudinale a été mise en place.
- Par la suite, un fort souhait d'évaluer leurs compétences a émergé.

Une enquête à grande échelle…

- Un échantillon de 5000 enfants handicapés nés en 2005 a été tiré.
- Ces enfants ne sont jamais allés à l'école avant la loi car ils ont 11 ans.
- La majorité de ces enfants sont à la fin du cycle primaire dans la scolarité française (CM2).

… Qui soulève des questions de méthode

- Quelle est la meilleure façon d'évaluer une population fortement hétérogène, en raison de la grande diversité des handicaps?
- Comment pouvons-nous transformer nos outils d'évaluation traditionnels en fonction du handicap de chaque enfant, afin de réduire efficacement le biais?

Test adaptatif numérique: une révolution

Ces questions nous ont amené à considérer rapidement les tests adaptatifs numériques comme une méthode d'évaluation appropriée.

Ce type d'évaluation est assez nouveau dans l'histoire des programmes d'évaluation en France.

Test Multi-stage: un bon compromis

	Linéaire	Adaptatif
+	Facilement réalisable Facile à concevoir et à mettre en place	Convient à un public cible vaste
-	Inadapté pour une population aussi hétérogène	**Inutilement difficile à concevoir** Gourmand en calcul

Multi-stage test:

- Chaque module est composé d'une groupe d'items
- En fonction de ses résultats, l'élève se voit proposer un module plus simple ou plus difficile

Design du Test

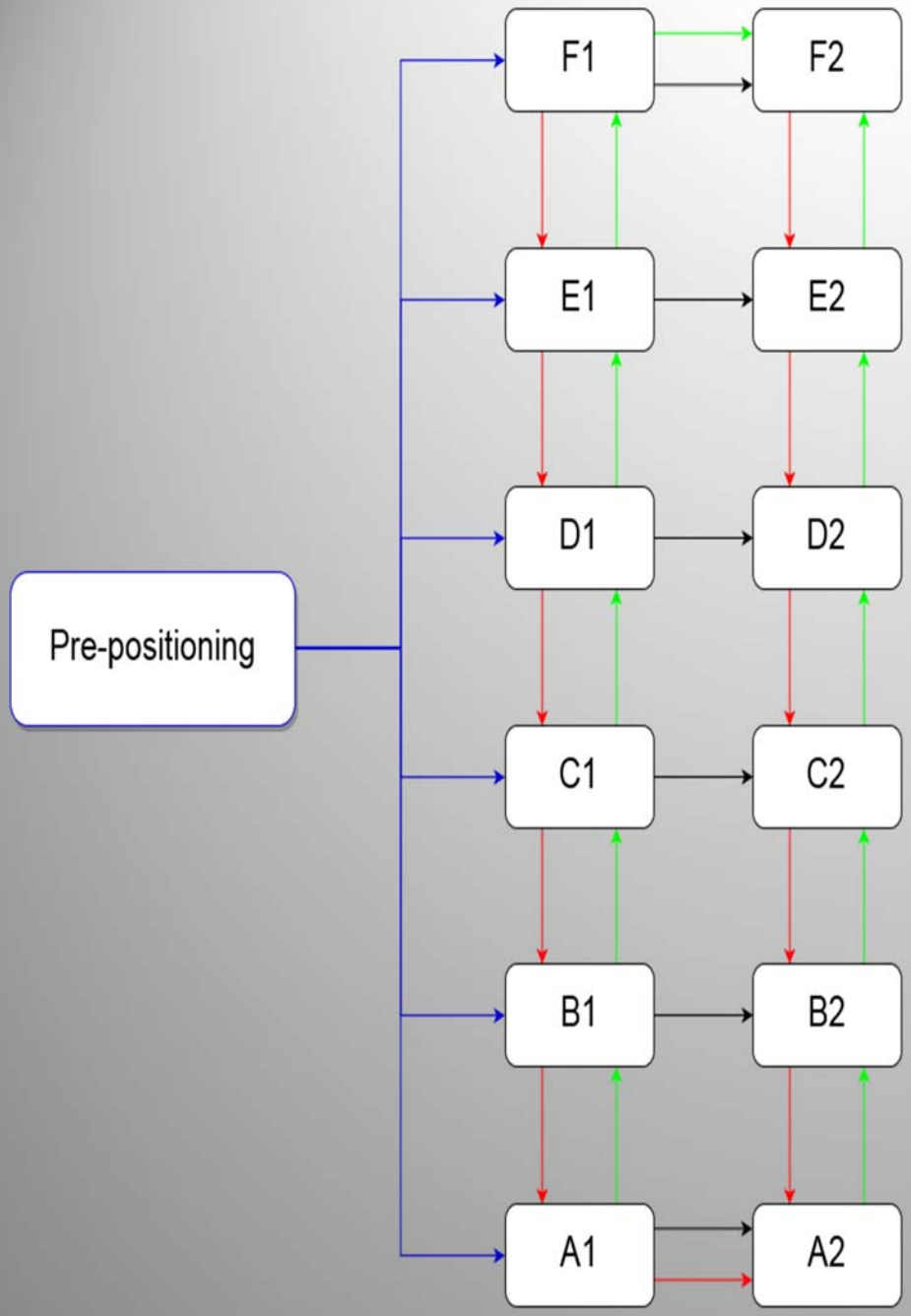
- Évaluation au format numérique
  - Permet d'adapter le test à chaque handicap, à un coût moindre que pour un test "papier crayon"
- En lecture et mathématiques
  - 4 tests distincts

    Lecture

    Compréhension de la langue

    Numération

    Calcul
- Questions fermées
  - Calcul automatique de score
- Items issus de la banque national d'items
  - Les métriques de chaque item sont calibrées sur d'autres enquêtes nationales.



Les modules sont construits par niveau croissant de difficulté (de A à F)

Critères d'orientation

- + de 80% de bonnes réponses: niveau supérieur
- - de 20% de bonnes réponses: niveau inférieur
- Sinon: second module du même niveau

The test se termine si:

- Il n'y a plus d'orientation possible
- Trois modules ont déjà été passés

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE
DE L'ÉDUCATION NATIONALE,
DE L'ENSEIGNEMENT SUPÉRIEUR
ET DE LA RECHERCHE

aea
EUROPE
November 3 - 5, 2016
Limassol, Cyprus

Résultats

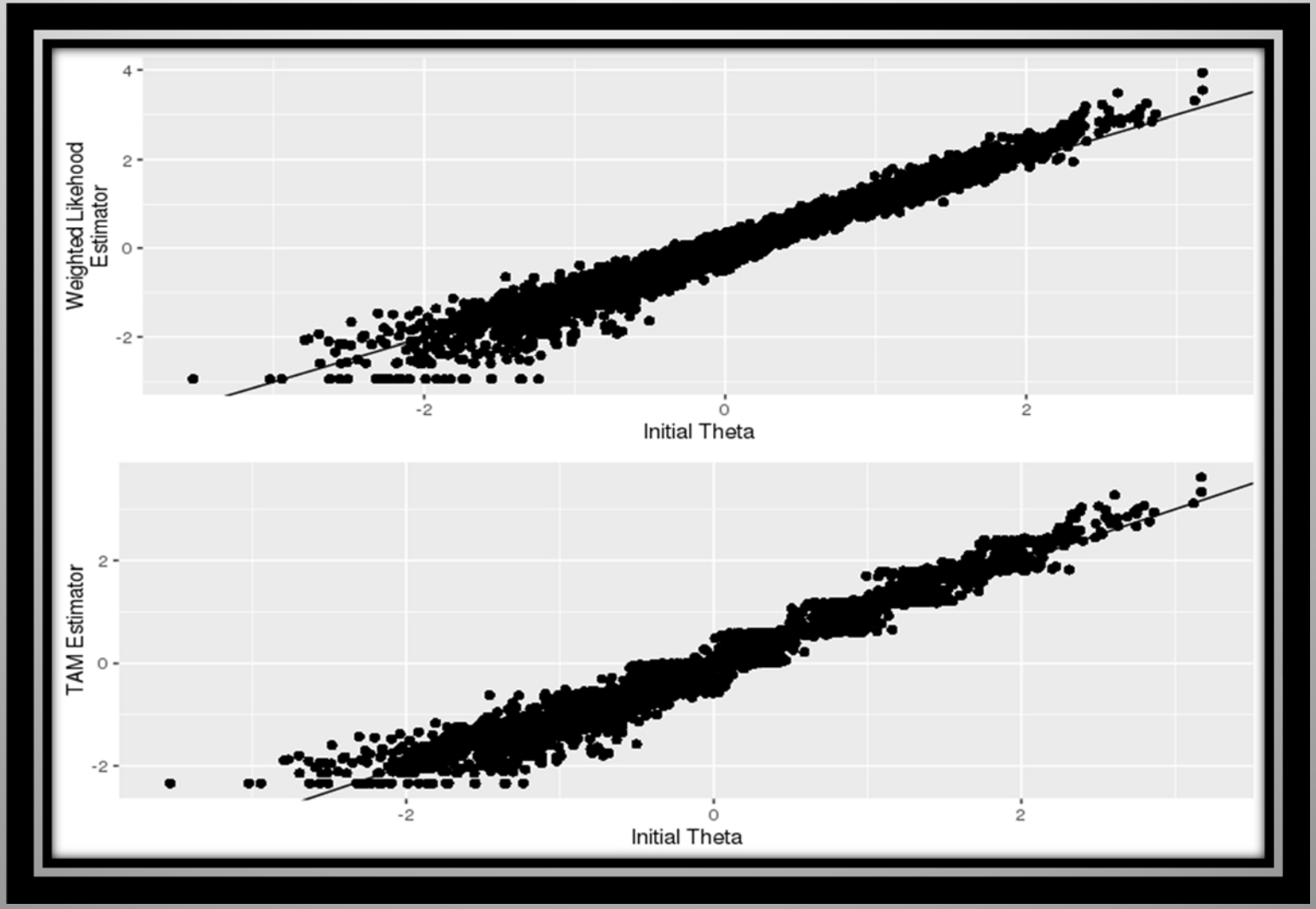
Les tests adaptatifs sont efficaces, mais fortement tributaires du pré-positionnement

Les tests adaptatifs sont efficaces…

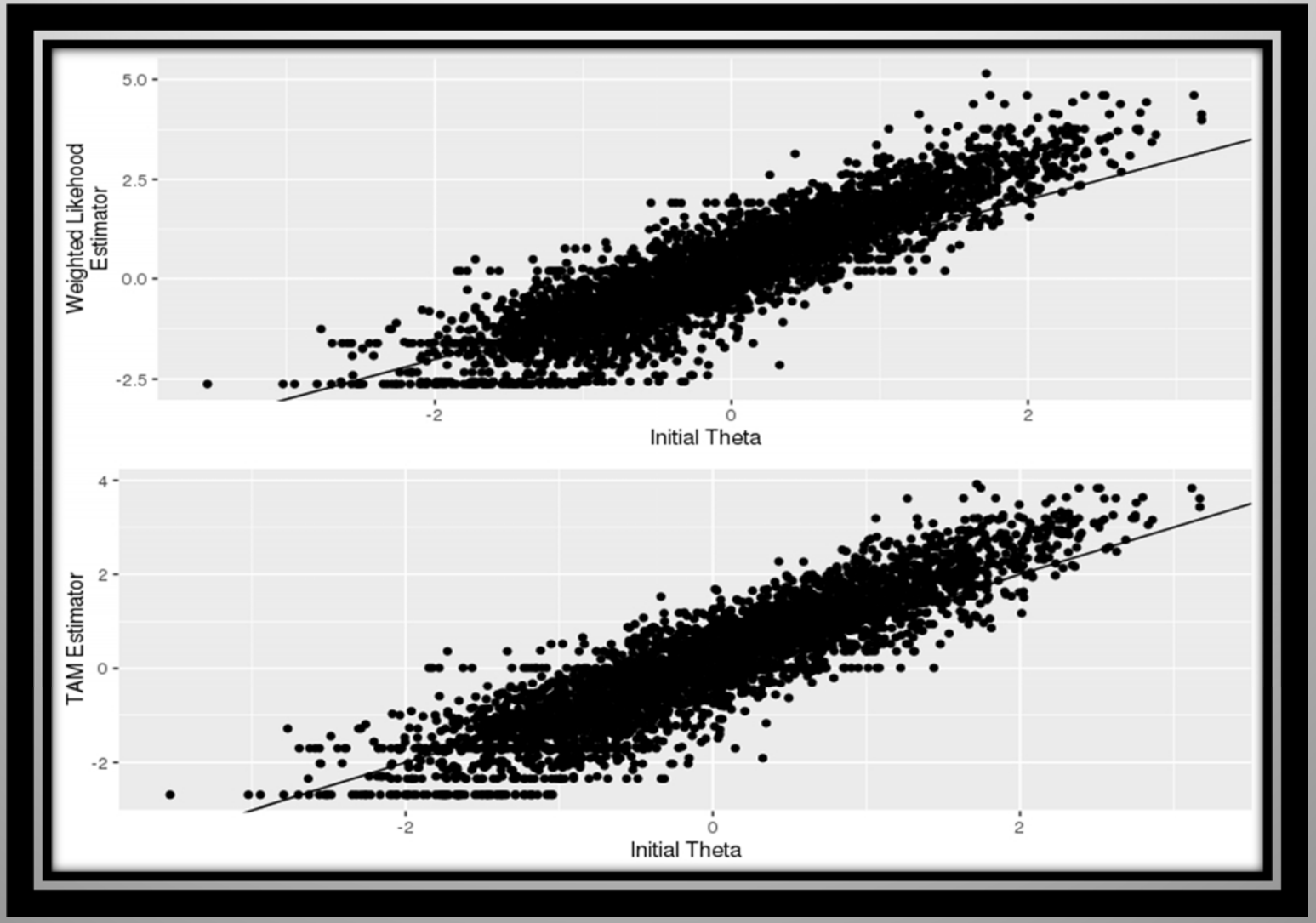
La théorie classique est fiable sur les paramètres d'items

- Nous avons construit un simulateur de test pour la Numération, respectant les mêmes critères que pour le test réel
- Nous avons soumis au test un échantillon aléatoire de 4200 « élèves », de niveaux très divers
  1. Test linéaire
  2. Test MultiStage
- Nous comparons le theta estimé à celui généré initialement

Efficacité du test linéaire



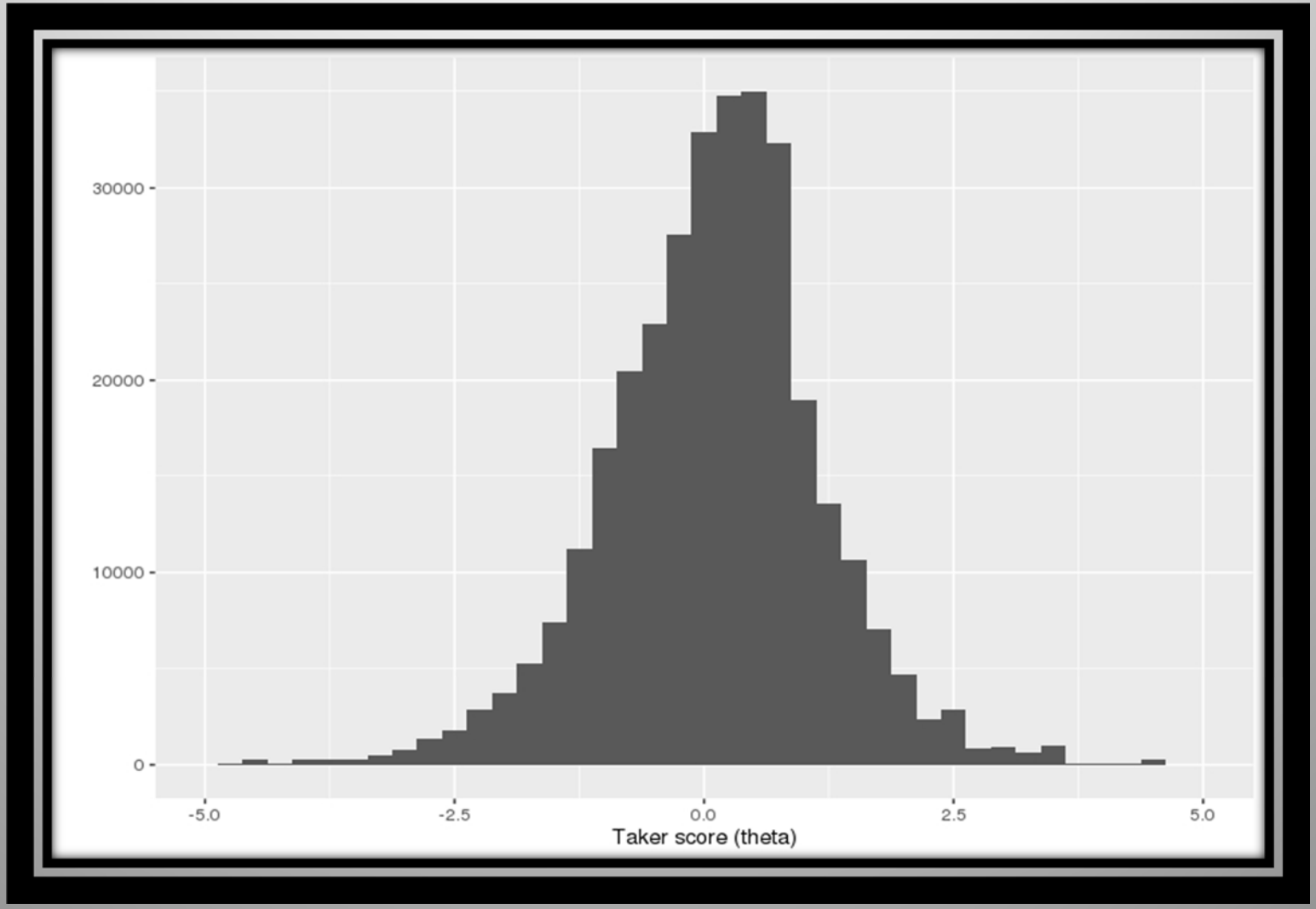
Efficacité du test MultiStage



Fiabilité du modèle de réponse à l'item?

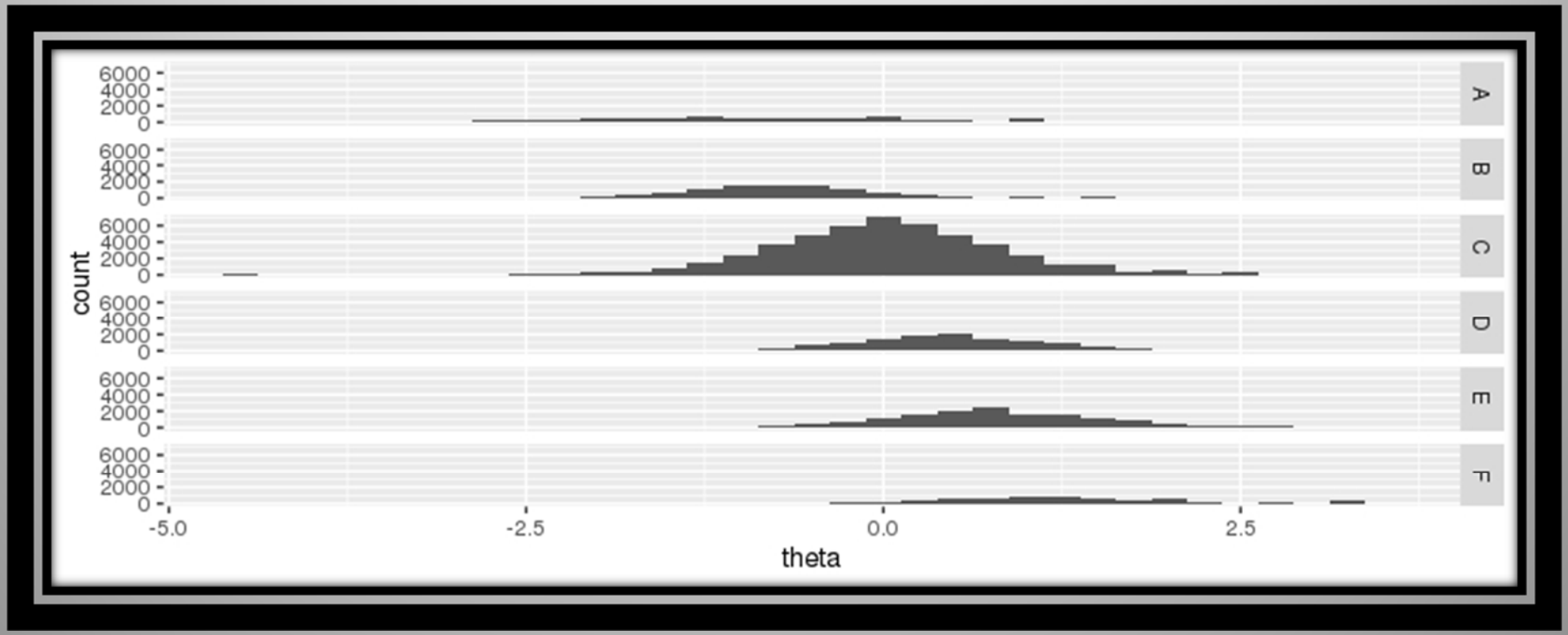
- Pour chaque test, le score (theta) des élèves se distribue ainsi:
  - moyenne = 0
  - écart-type = 1
- On peut donc les assembler pour vérifier la distribution dans son ensemble

Distribution des scores

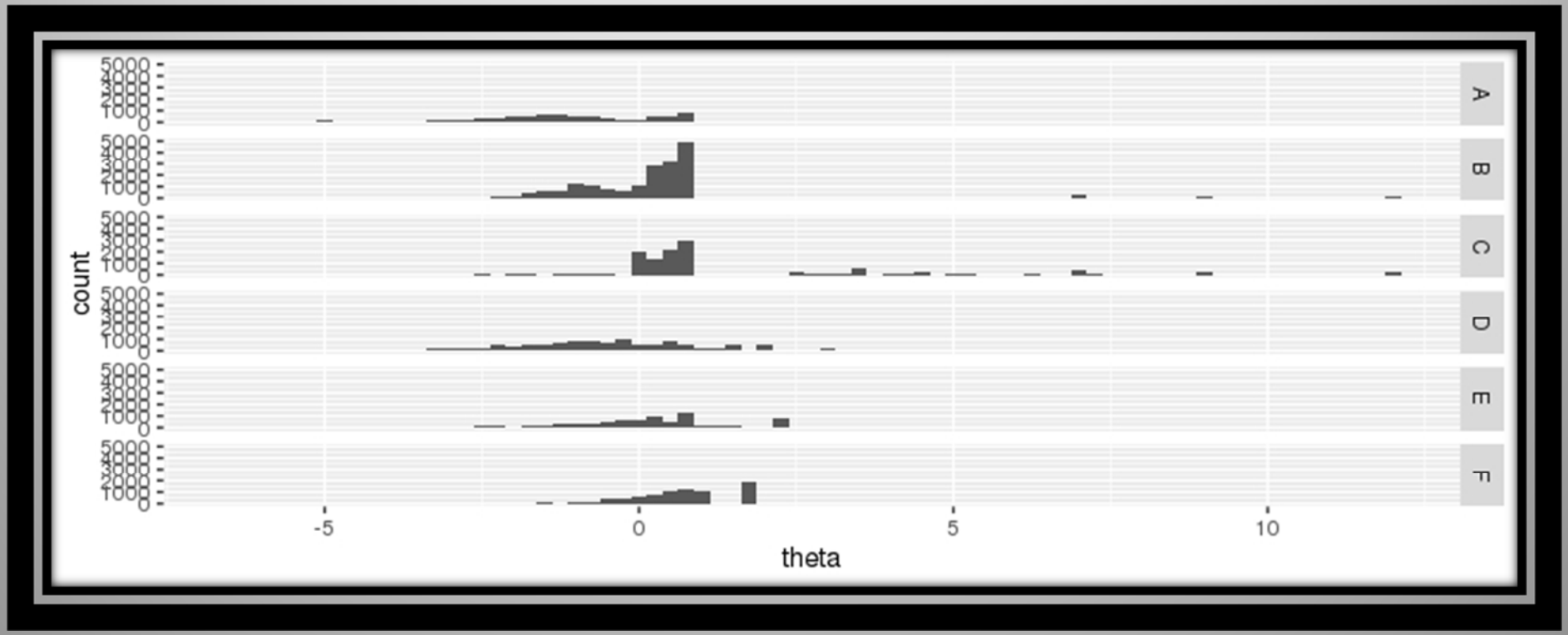


… mais dépend fortement du pré-positionnement

Fiable quand il est bon



Beaucoup moins quand il ne l'est pas



Possible explications

- Some students were assessed by inappropriate items
  - They failed most of them: theta underestimated
  - They succeeded most of them: theta overestimated
- Bad pre-positioning due to wrong estimation of the taker's level by the teacher
  - Level frequently underestimated: overprotection?
- Wrong composition of the modules
  - Modules not different enough
  - Modules not homogeneous enough
- Lack of depth in the test
  - Theory: test depth = number of paths

Quelle suite?

Amélioration du design

- **Améliorer l'accessibilité**
  - Libérer l'élève de la dépendance de son assistant
- **Améliorer la qualité des modules**
- **Renforcer la profondeur du test**

MERCI