Développement et évaluation d'un test mathématique adapté, administré par ordinateur

Jang Schiltz

Université du Luxemburg

jang.schiltz@uni.lu

La théorie de réponse par item (IRT)

Les postulats de l'IRT

- On suppose que la probabilité d'une réponse correcte est attribuable à la position du sujet sur un nombre spécifique *k* de traits latents de compétence requise pour répondre au type d'items en question.

- On suppose qu'il y a indépendance locale, c'est-à-dire que la probabilité d'une réponse correcte d'un sujet à un item donné ne dépend pas de ses réponses aux autres items du test.

La fonction de réponse de l'item (IRF)

- La réponse d'un sujet donné à un item dichotomique *j* est codée par

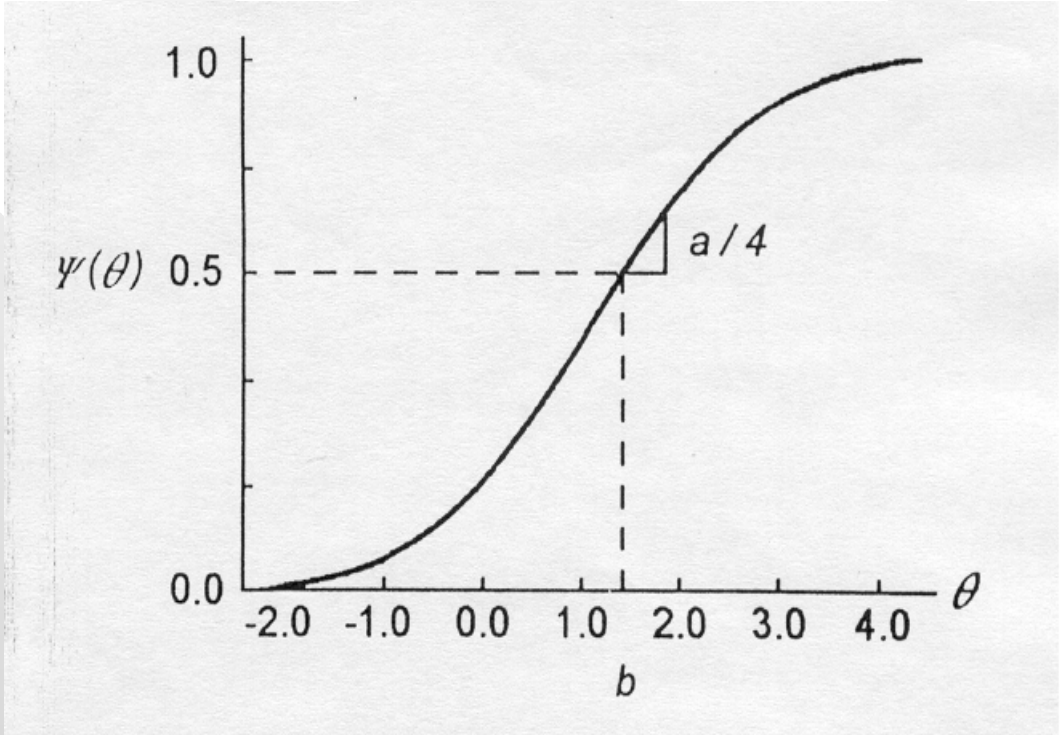
$$x_j = \begin{cases} 1, \text{ si la réponse est correcte} \\ 0, \text{ si la réponse est fausse} \end{cases}$$

- La probabilité d'une réponse correcte à l'item *j* pour une compétence donnée θ est égale à

$$P_j(\theta) = \frac{1}{1+e^{-a_j(\theta-b_j)}},$$

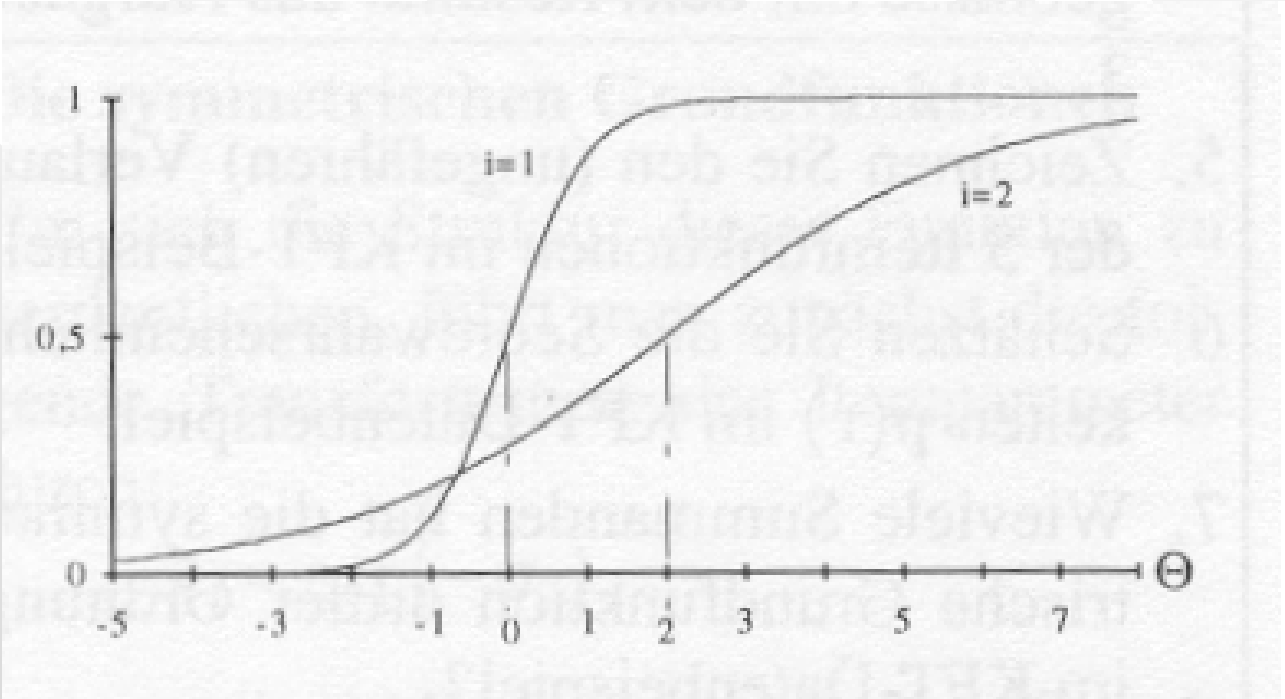
où a_j désigne la discrimination de l'item (*item slope*) et b_j sa difficulté (*item threshold*).

La fonction de réponse de l'item



Interprétation des paramètres du modèle IRT à 2 paramètres

La fonction de réponse de l'item



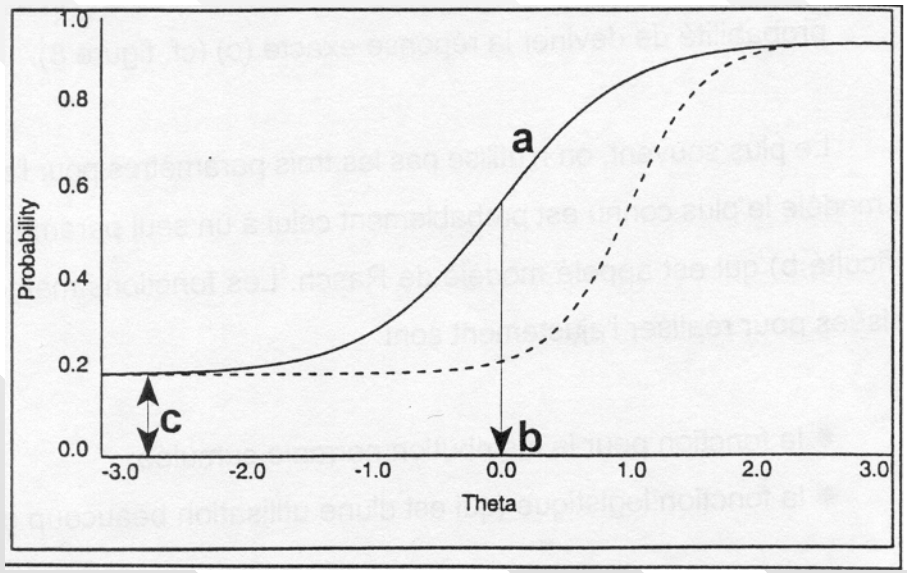
2 IRF avec les paramètres a_1=2, b_1=0,5, respectivement a_2=0, b_2=2

La fonction de réponse de l'item

La probabilité qu'un sujet avec une compétence égale à θ ne connaît pas la bonne réponse, mais répond correctement est $g_j[1-\Psi_j(\theta)]$, si g_j désigne la probabilité de deviner juste. Par conséquent,

$$\begin{aligned} P_j(\theta) &= g_j\left[1-\Psi_j(\theta)\right]+\Psi_j(\theta) \\ &= g_j+(1-g_j)\Psi_j(\theta). \end{aligned}$$

La fonction de réponse de l'item



- La difficulté *b* de l'item détermine la position de la courbe par rapport à l'axe des compétences.
- La discrimination *a* est proportionnelle à la pente de l'IRF.
- Le paramètre de „guessing" *c*, également appelé niveau du score de pseudo-chance, indique la probabilité de donner la réponse exacte par hasard.

La fonction de réponse de l'item

Un des grands avantages de l'IRT est le fait que l'estimation des paramètres ne dépend pas de la population à partir de laquelle cette estimation est faite. Autrement dit, la distribution des compétences dans l'échantillon utilisé pour calibrer les items n'a aucun effet sur les estimations des paramètres des items.

Ceci veut dire qu'une échelle de mesure semblable peut être utilisée dans des populations différentes et que les sujets peuvent être testés avec des items différents, appropriés à leur niveau de compétence tout en préservant la comparabilité de leurs scores .

Par conséquent, une IRF donne la probabilité qu'un individu répondra correctement à un item à un niveau donné de compétence, sans que cette probabilité ne dépende du nombre de sujets situés à ce niveau de compétence. Cette propriété d'invariance des IRF est une caractéristique importante des modèles de théorie de réponse par item.

La fonction d'information de l'item (IIF)

En IRT, il est possible de déterminer combien d'information chaque item fournit à chaque point du continuum de compétence.

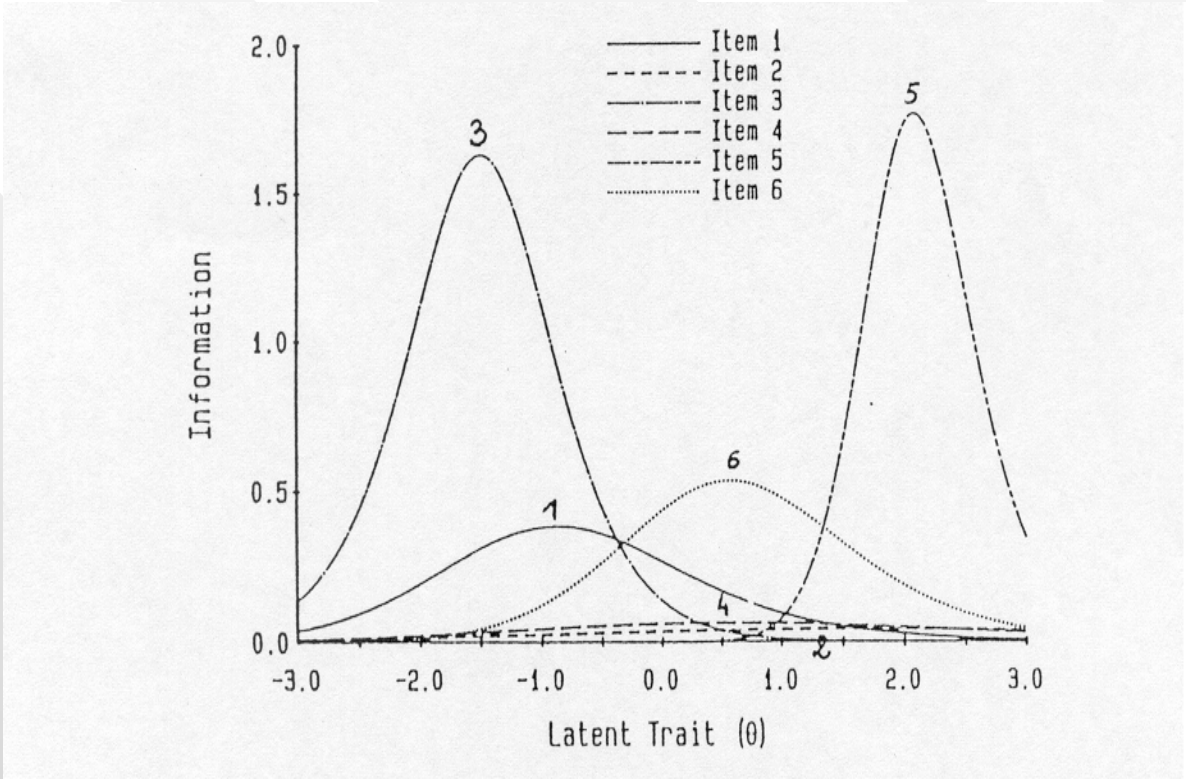
Plus on aura d'information à un certain niveau de compétence, plus précise sera l'évaluation des paramètres à ce niveau.

Un des avantages de l'IRT est donc de pouvoir choisir les items fournissant un maximum d'information à des niveaux spécifiques de compétence.

$$I(\theta) = a_j^2 P_j(\theta) \left[1 - \mathrm{P_j}(\theta)\right].$$

Plus cette pente est raide, plus la probabilité de répondre correctement à un item varie en fonction de la compétence, à condition de se trouver dans un voisinage du point de pente maximale de l'item.

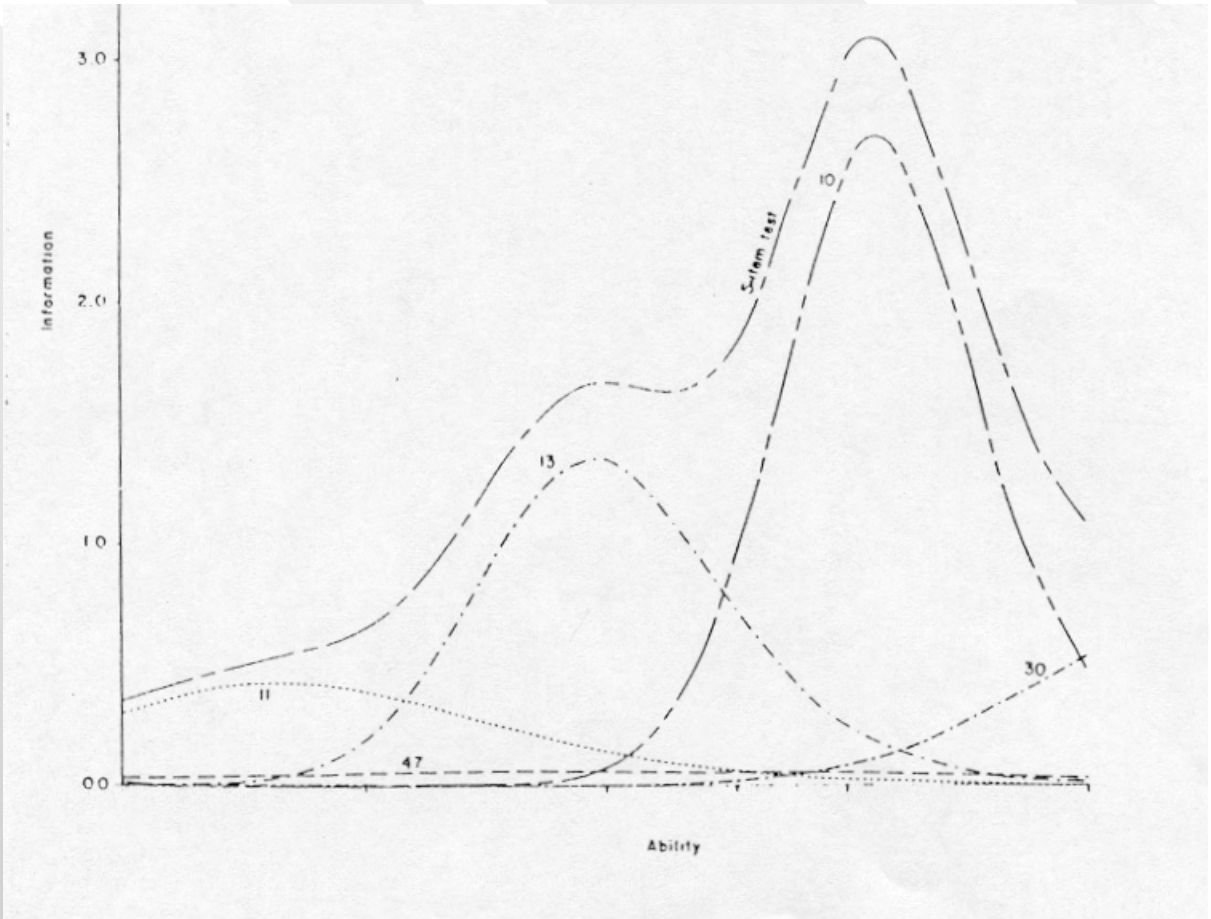
La fonction d'information de l'item (IIF)



Les IIF d'un test comportant 6 items

La fonction d'information du test (TIF)

Le postulat d'indépendance locale permet la sommation de l'information des différents items pour un niveau de compétence donné. On obtient ainsi la fonction d'information du test



La fonction d'information du test (TIF)

Alors que, dans la théorie classique des tests, la contribution de chaque item à la fidélité ou à la validité du test dépend de tous les autres items qui composent le test, dans la théorie des IRT, la contribution d'un item à l'efficacité d'un test ne dépend donc que de ses propriétés intrinsèques.

La TIF donne un moyen de connaître les niveaux du trait latent pour lesquels le test mesure le plus précisément et ceux pour lesquels il faut encore ajouter d'autres items. L'une des utilisations les plus importantes du concept d'information dans l'IRT réside dans le fait que les IIF et la TIF peuvent être développés avant que le test ne soit construit.

L’estimation des niveaux de compétence

Il y a deux méthodes courantes pour estimer les niveaux de compétence des sujets. Ces méthodes sont d’une part l’estimation du maximum de vraisemblance (*maximum likelihood estimation*) et d’autre part l’estimation modale bayesienne (*Bayesian modal estimation*).

Pour ces deux méthodes, on tient compte de toute l’information contenue dans la suite des réponses d’un sujet aux items du test. Le vecteur réponse pour chaque individu consiste en une suite de 1 et de 0, qui indiquent s’il a répondu correctement ou non aux différents items. L’indépendance locale implique que la probabilité pour un vectur donné est obtenue par multiplication.

L'estimation du maximum de vraisemblance

L'estimation du maximum de vraisemblance des réponses du sujet *i* est la valeur θ qui maximise la fonction L_i déterminée par

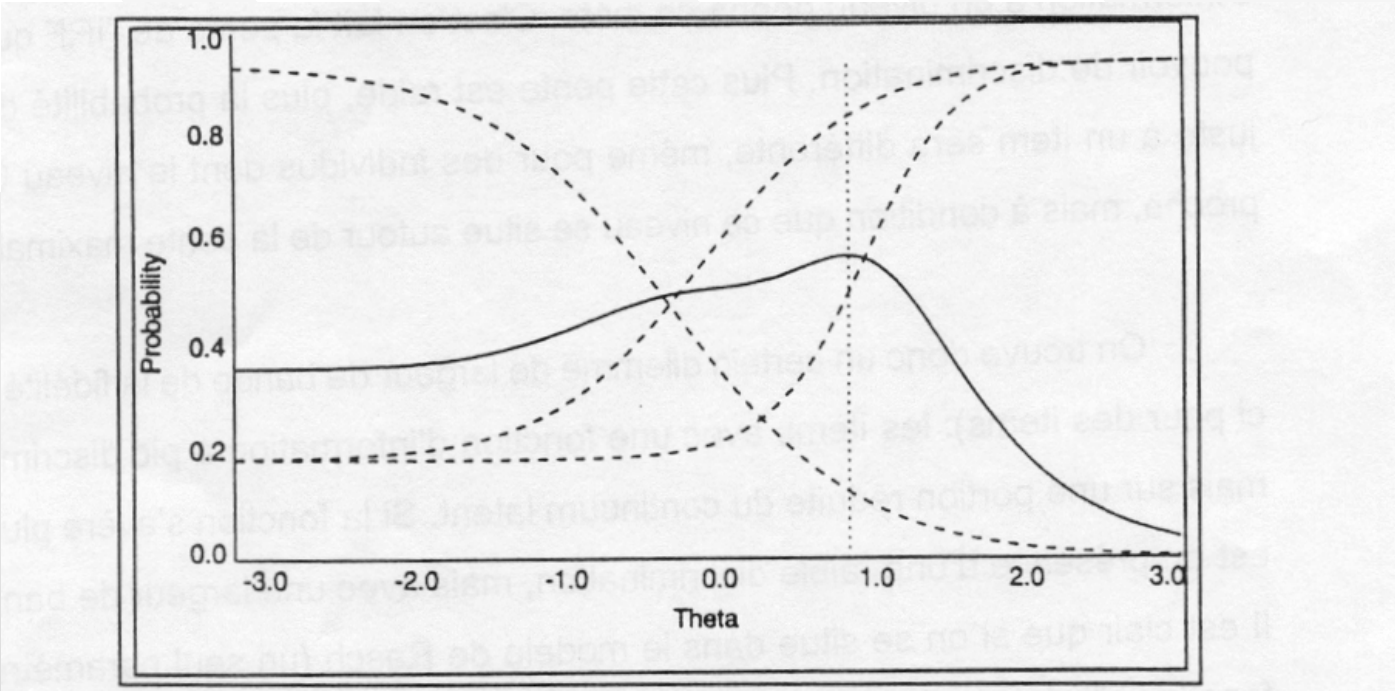
$$\log\ L_i(\theta) = \sum_j \left\{ x_{ij} \log\ P_j(\theta) + (1 - x_{ij}) \log\left[1 - P_j(\theta)\right] \right\},$$

où x_{ij} désigne la réponse du sujet *i* à l'item *j*.

Il faut alors résoudre l'équation de vraisemblance suivante

$$\frac{\partial \log L_i(\theta)}{\partial \theta} = \sum_j \frac{x_{ij} - P_j(\theta)}{P_j(\theta)\left[1 - P_j(\theta)\right]} \frac{\partial P_j(\theta)}{\partial \theta} = 0,$$

L’estimation du maximum de vraisemblance



Un problème avec l’estimation du maximum de vraisemblance est que, si un sujet répond correctement à tous les items (vecteur parfait) ou incorrectement à tous les items (vecteur zéro), alors la fonction de vraisemblance n’admet pas un seul maximum identifiable d’une manière univoque.

L'estimation modale bayesienne

L'estimation de la fonction postérieure de vraisemblance des réponses du sujet *i* est la valeur θ qui maximise la fonction

$$L(\theta|x_i) = \sum_j \left\{ x_{ij} \log P_j(\theta) + (1 - x_{ij}) \log\left[1 - P_j(\theta)\right] \right\} + \log g(\theta),$$

où g(θ) désigne la densité de la distribution a priori de θ.

Il faut alors résoudre l'équation de vraisemblance suivante

$$\sum_j \frac{x_{ij} - P_j(\theta)}{P_j(\theta)\left[1 - P_j(\theta)\right]} \frac{\partial P_j(\theta)}{\partial \theta} + \frac{\partial \log g(\theta)}{\partial \theta} = 0.$$

Les tests sur mesure administrés par ordinateur

Les tests classiques

- Les tests à pic présentent des items à difficulté homogène, généralement concentrés autour d'un niveau moyen de difficulté. Ces items sont seulement adaptés à des individus à compétence moyenne. Cela suffit certes pour la majorité de la population, mais un tel test ne permet pas de réaliser des discriminations précises pour des individus à compétence faible ou élevée.

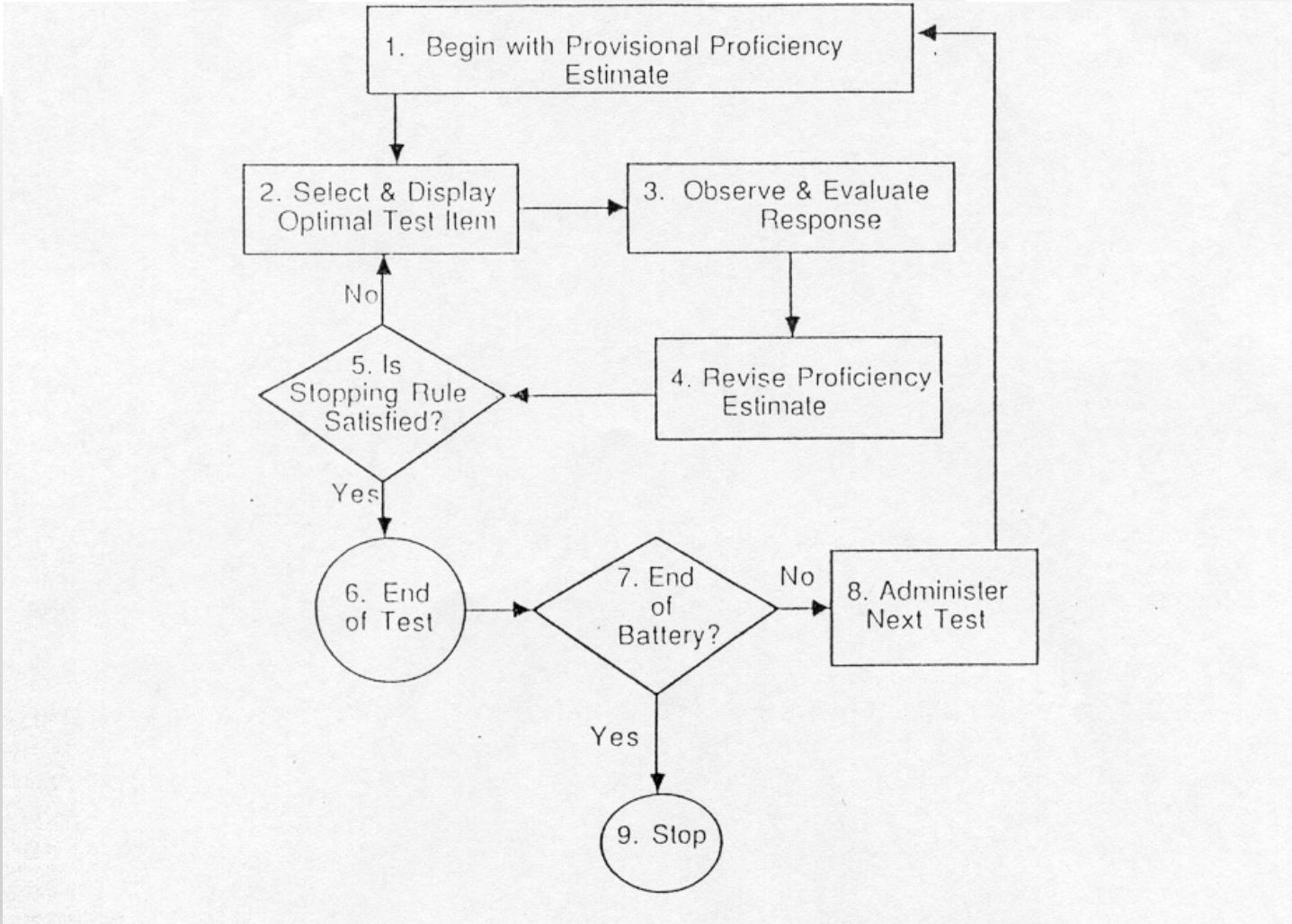
- Les tests rectangulaires sont construits de façon à avoir un nombre égal d'items à chaque niveau de difficulté. Ils permettent de mesurer toutes les aptitudes avec la même précision, mais cette précision est assez faible, puisque, pour chaque sujet, la plupart des items sont non adaptés.

Les tests sur mesure

- En posant des items différents aux différents sujets, car cela permet d'adapter le niveau de difficulté des items à la compétence de l'individu qui est testé. Un tel test procure des mesures d'une précision élevée, égale à tous les niveaux du trait latent, à condition de disposer d'une banque d'item assez grande.

- En plus de ces avantages psychométriques évidents, cela implique également que la probabilité de répondre correctement reste toujours dans un voisinage de 50%.

Les tests sur mesure par ordinateur

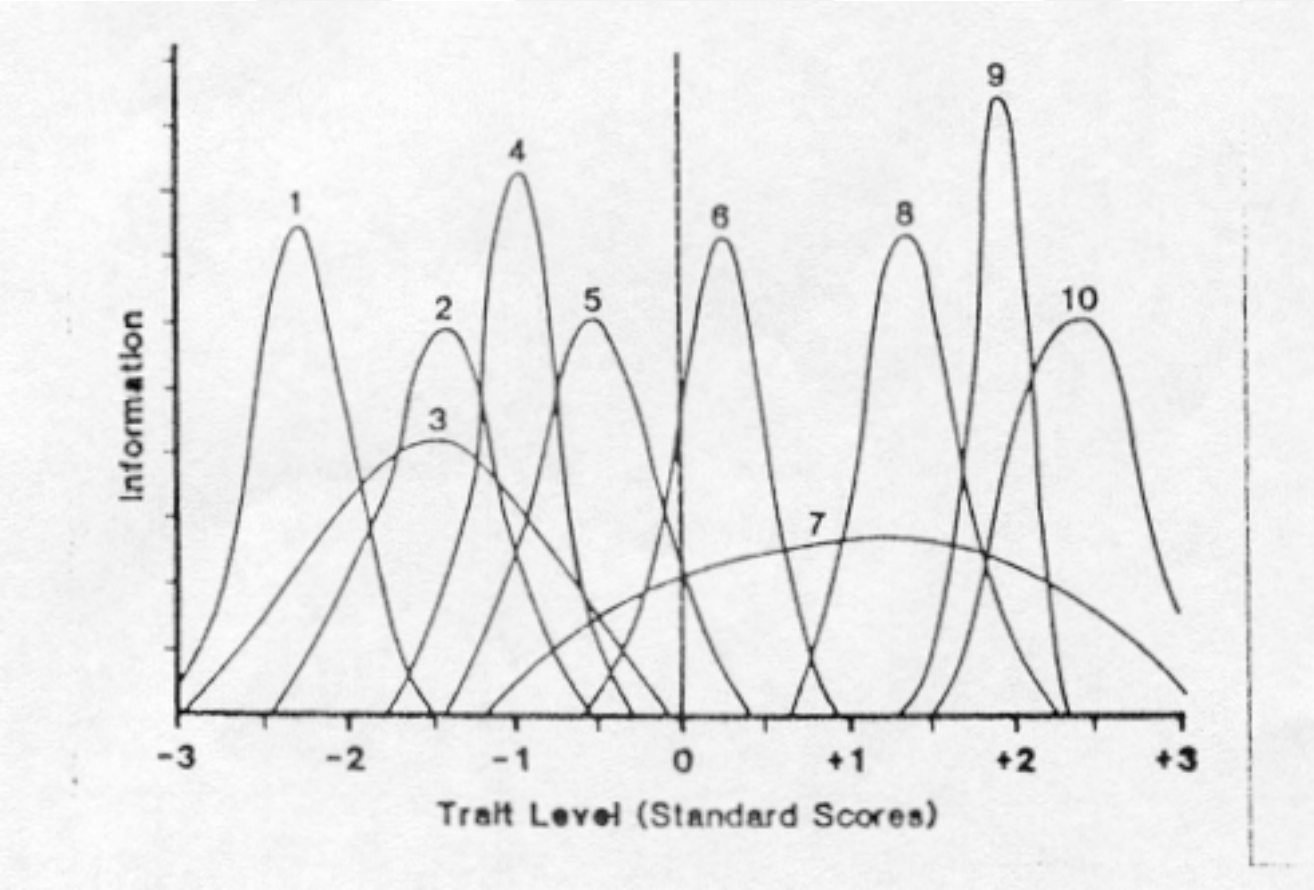


Le premier item administré

- On choisit généralement la moyenne de la population comme estimation initiale de θ.

- Les propriétés psychométriques de l'IRT impliquent que, de toute façon, un mauvais choix du premier item n'aura que peu d'incidences sur le résultat final, à moins que le test soit vraiment très court.

La stratégie d'information maximale



La table d'information

θ	Items in order of preference from left to right, at each value of θ	Information
−2.25	955 50 948 915 136 57 64 331 296 362 938 35 366 106 349	5.2
−2.12	948 955 50 915 136 57 296 331 64 35 362 106 938 366 349	5.6
−2.00	948 955 50 136 915 35 296 331 106 57 64 366 362 349 938	5.8
−1.88	948 955 50 35 136 296 106 331 915 349 64 57 366 362 938	5.9
−1.75	948 955 35 50 106 296 136 349 331 915 366 64 57 377 362	5.8
−1.62	948 35 106 349 955 296 50 331 136 366 915 377 64 57 321	5.6
−1.50	948 35 349 106 296 955 331 136 50 377 321 366 64 915 68	5.5
−1.38	35 349 948 106 296 321 377 331 955 280 136 68 366 50 103	5.5
−1.25	349 35 948 106 321 377 280 296 68 331 103 366 136 955 50	5.4
−1.12	349 321 35 280 106 377 103 948 68 296 341 146 331 366 144	5.5
−1.00	321 349 280 35 103 377 341 68 146 106 144 296 948 331 42	5.6
−0.88	321 280 103 341 349 144 146 35 377 68 106 42 296 948 67	5.9
−0.75	144 321 341 280 103 146 349 42 377 35 68 106 67 1 296	6.3
−0.62	144 341 321 146 103 280 42 349 377 68 35 1 67 8 106	6.5
−0.50	144 341 146 42 321 103 280 1 377 68 349 67 8 128 35	6.8
−0.38	144 42 341 146 321 103 280 1 128 8 328 377 67 107 68	7.1
−0.25	144 42 341 146 321 103 128 280 1 88 328 8 49 107 67	7.5
−0.12	144 42 341 146 88 128 321 328 49 103 1 376 8 107 280	7.8
0.00	144 42 88 146 128 376 341 49 328 8 1 107 321 103 383	8.0
0.12	88 42 376 144 128 49 328 146 383 341 365 8 107 1 85	8.3
0.25	88 376 42 365 383 49 128 144 328 85 335 146 107 8 341	8.9
0.38	376 88 365 383 85 49 42 128 126 335 328 144 107 146 241	9.4
0.50	126 365 376 85 346 383 215 88 335 49 241 128 362 328 42	10.1
0.62	126 365 376 85 346 383 215 88 335 49 241 128 262 328 42	11.1
0.75	126 365 346 215 376 85 383 88 241 335 49 262 355 399 128	11.6
0.88	346 126 215 365 85 376 383 241 355 335 88 262 228 399 49	11.8
1.00	346 215 126 365 85 355 376 228 347 241 262 209 383 399 335	11.7
1.12	346 215 126 347 228 355 365 85 209 376 241 399 262 160 251	11.3
1.25	346 347 215 228 355 126 160 209 365 251 85 399 262 376 241	10.9
1.38	347 160 346 228 355 215 237 209 251 126 365 399 263 378 262	10.8
1.50	160 347 237 228 346 355 251 209 215 378 263 399 126 262 365	10.7
1.62	160 237 347 228 251 355 209 346 215 378 263 186 399 262 126	10.4
1.75	160 237 347 228 251 355 209 186 346 378 263 399 215 262 241	10.0
1.88	160 237 347 251 228 186 355 209 378 263 346 399 215 262 241	8.9
2.00	160 237 186 251 347 228 355 378 209 263 399 346 262 215 115	7.7
2.12	160 237 186 251 347 228 355 378 263 209 399 346 262 115 215	6.3
2.25	160 186 237 251 347 228 378 355 263 209 399 262 115 346 241	5.0

La règle de terminaison

- Un test sur mesure peut se terminer quand une précision déterminée est atteinte, mais aussi, après un nombre fixé d'items, ou lorsqu'un temps déterminé s'est écoulé.
- En pratique, lorsqu'on veut atteindre un degré de précision déterminé, on impose souvent simultanément un nombre maximal d'items administrés car, dans certains cas, la banque d'items peut être épuisée avant que la précision voulue ne soit atteinte.

L’analyse des items

Le contexte

- Nouvelle procédure d'orientation pour la transiiton entre l'école primaire et l'école secondaire.
- Test mathématique de novembre 1996
- Il n'y avait pas beaucoup d'items à choix multiple. Par contre certains items demandaient des constructions géométriques
- 81 items
- Population: 3590 élèves, généralement de 12 ans
- 7 items ne vérifiaient pas l'hypothèse d'indépendanc locale. Il restaient 74 items à analyser plus sérieusement.

La calibration des items

- Faite avec BILOG-MG
- J'ai testé le modèle IRT à 1 respectivement 2 paramètres
- Le test du quotient de vraisemblance indiquait que les deux modèles étaient valides
- Mais le modèle à 1 paramètre conservait seulement 26 items, alors que celui à 2 paramètres permettait d'en garder 67.
- Item Response Function:

$$P_j(\theta) = \frac{e^{a_j(\theta - b_j)}}{1 + e^{a_j(\theta - b_j)}}$$

Tester la propriété d'invariance

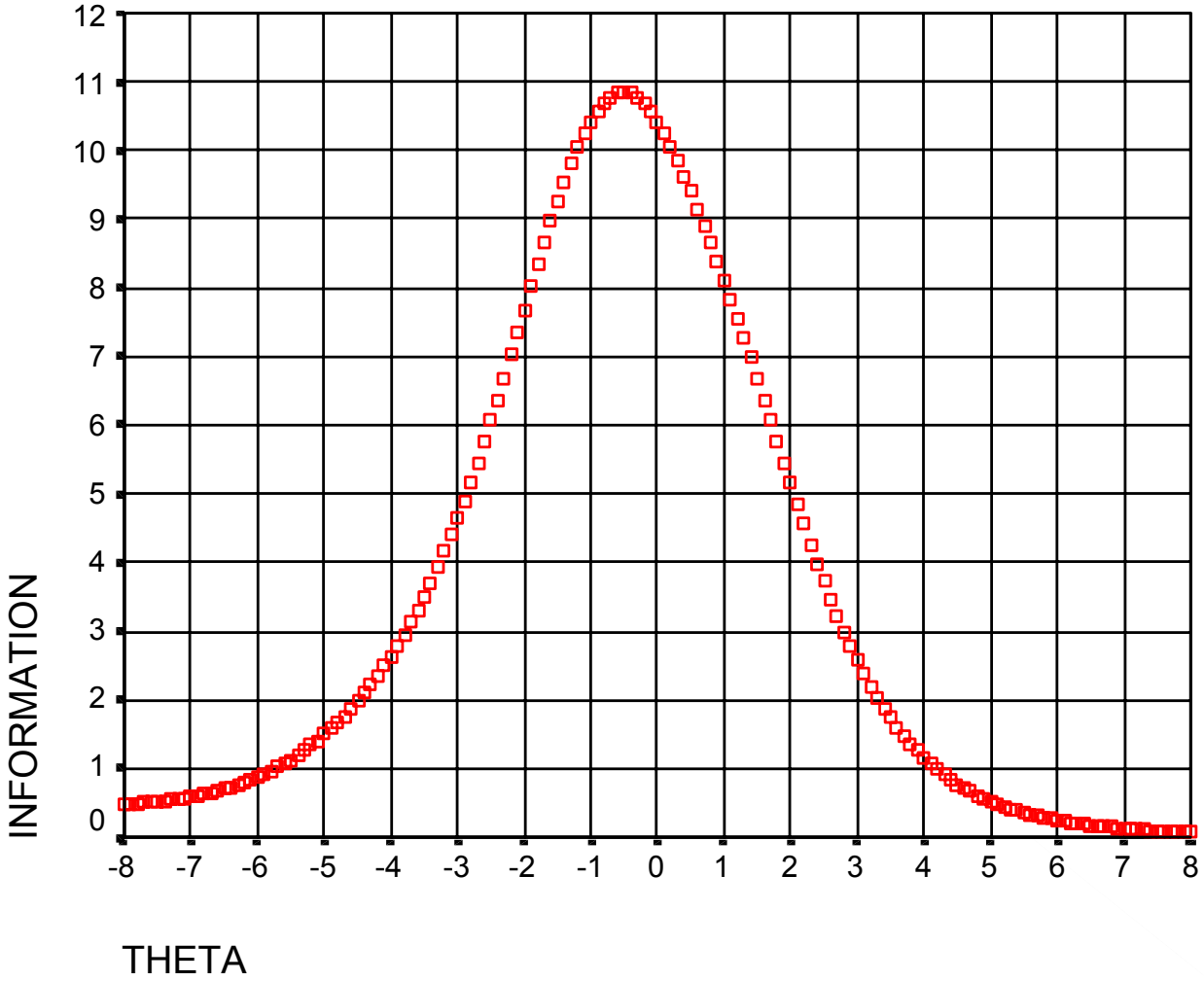
- La prochaine étape consistait en l'élimination des ditems qui ne satisfaisaient pas à la propriété d'invariance.

- J'ai séparé la population en 2 groupes en les divisant à la médiane des notes obtenus dans le test classique.

- J'ai calibré les items séparément pour les 2 groupes en gardant la même distribution et j'ai rejeté les items pour lesquels la différence de l'indice de difficulté entre les deux groupes était en valeur absolue plus grande que l'écart-type multiplié par 1,96.

- J'ai gardé finalement une banque à 63 items.

La fonction d'information du test



Le logiciel

Construction du CAT

- J’ai utilisé la plateforme de programmation multimédia Quest Net+ for Windows™.

- Initialisation: Un item à difficulté moyenne (0,078) et de discrimination assez faible (0,618).

Estimation de la capacité mathématique

- J'ai utilisé la méthode d'estimation bayesienne modale avec une distribution normale de θ.
- Fonction de vraisemblance postérieure:

$$L(\theta) = \frac{\exp\left[\sum_{i=1}^{k-1} a_{j_i}(b_{j_i} - \theta)(1 - u_{j_i})\right]}{\prod_{i=1}^{k-1}\left[1 + \exp\left[-a_{j_i}(\theta - b_{j_i})\right]\right]} g(\theta).$$

Elle atteint son maximum pour θ vérifiant

$$-\sum_{i=1}^{k-1} a_{ji}(1 - u_{ji}) + \sum_{i=1}^{k-1} \frac{a_{ji} e^{a_{ji}(b_{ji} - \theta)}}{1 + e^{a_{ji}(b_{ji} - \theta)}} - 2\theta = 0,$$

Estimation de la capacité mathématique

```
float algo (float x_0)
   {float res, a, b;
    a = x_0-2.0;
    b = x_0+2.0;
    do {
            if (funct(a)*funct(b)<0)
            {if (funct(a)*funct((a+b)/2.0)<0)  b=(a+b)/2.0;
            else a=(a+b)/2.0;}
            else a=a-1.0, b=b+1.0;
        }
    while (b-a>0.005);
    res=a;
  return res;}
```

Choix de l'item suivant

- Stratégie d'information maximale:
- Fonction d'information des items: $I(\theta) = a_j^2 P_j(\theta) \left[1 - \mathrm{P_j}(\theta)\right].$

```
for (i=1 ; i<=63 ; i++)
            {info[i-1]=I(theta, i);}

for (i=1 ; i<=nom ; i++)
      {info[ens[i-1]-1]=0.0;}

 float sup;
            int item=1;
            sup=info[0];
            for (i=2; i<=63 ; i++)
            {if (info[i-1]>sup) {sup=info[i-1];
                                 item=i;}}
```

Étude d’évaluation

Le plan expérimental

- On a soumis à 123 élèves de septième d'une part notre CAT et d'autre part un test papier-crayon comportant la banque d'item du CAT.
- Pour neutraliser l'effet de l'ordre de passation, les élèves furent classés en deux groupes. Le groupe A commença avec le CAT et le groupe B avec les test papier-crayon.
- Deux mois plus tard, les conditions expérimentales furent inversées.

Critère un: Notes obtenues et proportion de succès

- Il n'y a pas de différence significative, ni entre les moyennes des notes du CAT (M = 0.55, SD = 0.47) et celles du test papier-crayon (M = 0.61, SD = 0.65) ($t = 1.578$, $p > 0.05$), ni entre les proportions d'échec entre les deux tests (8,9 % contre 10,1 %) ($\chi^2 = 2.65$, $p > 0.05$).
- Les résultats et le pourcentage de réussite ne sont pas significativement différents dans la situation ordinateur et la situation papier-crayon.
- Les corrélations entre les scores du CAT et du test papier-crayon sont significatifs ($r = 0.593$, $p < 0.01$).

Critère deux : l'effet de l'ordre de passation

- CAT (groupe A: $m = 0.5674$, $\sigma = 0.408$; groupe B: $m = 0.5420$, $\sigma = 0.497$; $t = 0.285$; $p > 0.05$)
- Test papier-crayon (groupe A: $m = 1.0093$, $\sigma = 0.558$; groupe B: $m = 0.4668$, $\sigma = 0.560$; $t = 5.100$; $p < 0.001$).
- Il existe bien une influence significative de l'ordre de passation sur les notes du test papier-crayon, mais pas sur les notes du test sur ordinateur.

Critère trois: Relation avec l'intelligence générale

Correlations

		NOTE	NOTEPAP	GL	QU3_4	QU7_10	QU13_14	QU15
NOTE	Pearson	1	.493**	.362**	.337**	.393**	.035	.187*
	Sig. (2-tailed)	.	.000	.000	.000	.000	.705	.039
NOTEPAP	Pearson	.493**	1	.319**	.280**	.379**	.171	.131
	Sig. (2-tailed)	.000	.	.000	.002	.000	.060	.149
GL	Pearson	.362**	.319**	1	.541**	.722**	.476**	.338**
	Sig. (2-tailed)	.000	.000	.	.000	.000	.000	.000
QU3_4	Pearson	.337**	.280**	.541**	1	.461**	.353**	.197*
	Sig. (2-tailed)	.000	.002	.000	.	.000	.000	.030
QU7_10	Pearson	.393**	.379**	.722**	.461**	1	.406**	.119
	Sig. (2-tailed)	.000	.000	.000	.000	.	.000	.192
QU13_14	Pearson	.035	.171	.476**	.353**	.406**	1	.280**
	Sig. (2-tailed)	.705	.060	.000	.000	.000	.	.002
QU15	Pearson	.187*	.131	.338**	.197*	.119	.280**	1
	Sig. (2-tailed)	.039	.149	.000	.030	.192	.002	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Corrélations entre les notes du CAT et u test papier-crayon et les résultats de tests d'intelligence.

NOTE: CAT score

NOTEPAP: paper-and-pencil score

GL: general intellectual ability

QU3_4: reasoning

QU7_10: spatial representation

QU13_14: perceptual speed

QU15: computational skill