Les variables latentes

Qu’est ce que c’est?
Définitions informelles:

- Variable hypothétique (dont on fait l’hypothèse)
- Variable non-mesurable
- Variable qui synthétise plusieurs variables observées

Qu’est ce que c’est? Définitions formelles:

- Variable qui rend les variables observées indépendantes si l’on en tient compte. (→ local independence)
- Variable dont les valeurs correspondent à la valeur attendue si l’on avait répétée la mesure un nombre infini de fois. (→ expected value)
- Variable qu’on ne peut calculer à l’aide de variables observées.
- Variable que l’on a pas observée dans notre échantillon

Propriétés des variables latentes

- exploratoire versus confirmatoire
- Continue, catégorielle,…
- Identification: moyenne, variance si continue

Types de modèles

Observées / Latentes	Catégorielle	Continue
Catégorielle	Modèles de profils latents	Modèles de classes latentes
Continue	Modèles de réponse à l'item	Modèles d'équations structurales

A quoi ça sert?

- Synthétiser les données
- S’approcher au plus près des concepts théoriques sous-jacents à ce qui a été mesuré
- Comparer la théorie à l’échantillon

Exemple: échelle de suppression

- There are things I prefer not to think about.
- Sometimes I wonder why I have the thoughts I do.
- I always try to put problems out of mind.
- Sometimes I stay busy just to keep thoughts from intruding on my mind.
- There are things that I try not to think about.
- Sometimes I really wish I could stop thinking.
- I often do things to distract myself from my thoughts.
- I have thoughts that I try to avoid.
- There are many thoughts that I have that I don't tell anyone.

Exemple: échelle d’intrusion

- I have thoughts that I cannot stop.
- There are images that come to mind that I cannot erase.
- My thoughts frequently return to one idea.
- I wish I could stop thinking of certain things.
- Sometimes my mind races so fast I wish I could stop it.
- There are thoughts that keep jumping into my head.

Analyse factorielle

$$Y_i = b_0 + b_1\xi_{i1} + b_2\xi_{i2} + \ldots + b_K\xi_{iK} + u_i$$

- Y est la valeur d'une variable observée pour le *i*ème sujet,
- b_0 est l'intercept,
- b_k est la saturation qui donne l'impact du *k*ième facteur sur Y,

  ξ_{ik} est le score factoriel du *k*ième facteur
- u_i est l'«uniqueness» ou l'erreur du sujet *i*

Exemple: analyse factorielle

Question	suppression	intrusion
1	**.76**	-.21
2	**.37**	.18
3	-.24	**.93**
4	-.11	**.81**
5	.03	**.78**
6	**.41**	**.56**
7	.32	**.52**
8	**.37**	-.01
9	-.03	**.69**
10	**.55**	.20
11	**.94**	-.17
12	**.41**	.34
13	**.48**	.04
14	**.86**	-.00
15	.25	.22

Et maintenant que faire?

- Que représente le facteur?
  - Une combinaison linéaire de toutes les questions qui maximise le pourcentage de variance expliqué par le facteur.
  - Oui, mais… et au niveau théorique
    - Intrusion
    - Suppression
- Que faire des questions qui avaient des saturations basses sur les deux facteurs?
- Peut-on se fier à cette analyse: conditions d'applications

Rasch: structure des données

		Item i				r_υ
		1	2	3	4	
Personne υ	1	0	1	0	1	2
	2	1	1	1	1	4
	3	0	1	1	0	2
	4	1	0	1	0	2
	5	0	1	0	1	2
	6	0	1	0	0	1
	7	0	1	0	0	1
	8	0	1	1	1	3
	n_j	2	7	4	4	

Fréquence des vecteurs de réponse

x				$n(x)$
0	0	0	0	2
0	0	0	1	4
0	0	1	0	2
0	0	1	1	2
0	1	0	0	2
0	1	0	1	1
0	1	1	0	1
…				
1	1	1	1	3

Nb modalités puissance nb d'items vecteurs possibles =

2^4 vecteurs possibles

Définition Item Response Theory

- Il existe une ou plusieurs variables latentes qui sous-tendent (causent) les résultats aux items (catégoriels) d'un test.
- On examine la probabilité d'une réponse à une variable manifeste sachant la variable latente.
- Les composantes du comportement dans un test
  - Caractéristique du sujet: la capacité θ_υ
  - Caractéristique de l'item: la difficulté σ_ι
- La probabilité d'obtenir une certaine catégorie de réponse est une fonction des caractéristiques du sujet et de l'item:

$$\text{Probabilité d'une certaine réponse} = F(\theta_\upsilon, \sigma_\iota)$$

Calcul de probabilité

$P(X_{\upsilon i} = 1) = e^{(\theta\upsilon - \sigma\iota)} / 1 + e^{(\theta\upsilon - \sigma\iota)}$

Probabilité de réponse pour une personne avec $\theta_\upsilon = -2.5$ et un item avec $\sigma_\iota = -1.15$

$$P(X_{\upsilon i} = 1) \quad = 2.72^{(-2.5-(-1.15))} / 1 + 2.72^{(-2.5-(-1.15))}$$

$$= 2.72^{(-1.35)} / 1 + 2.72^{(-1.35)}$$

$$= 0.26 / 1 + 0.26 = .21$$

Probabilités de réponse pour deux personnes

	Personne	
σ_i	θ_v = -2.5	θ_v = 2.4
-1.15002	.21	.97
0.49362	.05	.87
0.52554	.05	.87
0.76415	.04	.84
-1.37806	.25	.98
0.57902	.04	.86
0.12744	.07	.91
1.08483	.03	.79
-1.01154	.18	.97
-0.47209	.12	.94
-0.60380	.13	.95
0.31469	.06	.89
-0.04901	.08	.92
0.77523	.04	.84

Estimation des paramètres

Probabilité de réponse à deux items :

$$P\left(X_{vi}=x_{vi},X_{vj}=x_{vj}\right)=P\left(X_{vi}=x_{vi}\right)\cdot P\left(X_{vj}=x_{vj}\right)$$

$$=\frac{\mathrm{e}^{x_{vi}(\theta_v-\sigma_i)}}{1+\mathrm{e}^{(\theta_v-\sigma_i)}}\cdot\frac{\mathrm{e}^{x_{vj}(\theta_v-\sigma_j)}}{1+\mathrm{e}^{(\theta_v-\sigma_j)}}$$

Attention: exigence d'indépendance entre les items ou indépendance locale

Estimation des paramètres

- Probabilité d'un vecteur de réponse :

$$P(\mathbf{x}) = \prod_{i=1}^{k} P(X_{vi} = x_{vi}) = \prod_{i=1}^{k} \frac{e^{x_{vi}(\theta_v - \sigma_i)}}{1 + e^{(\theta_v - \sigma_i)}}$$

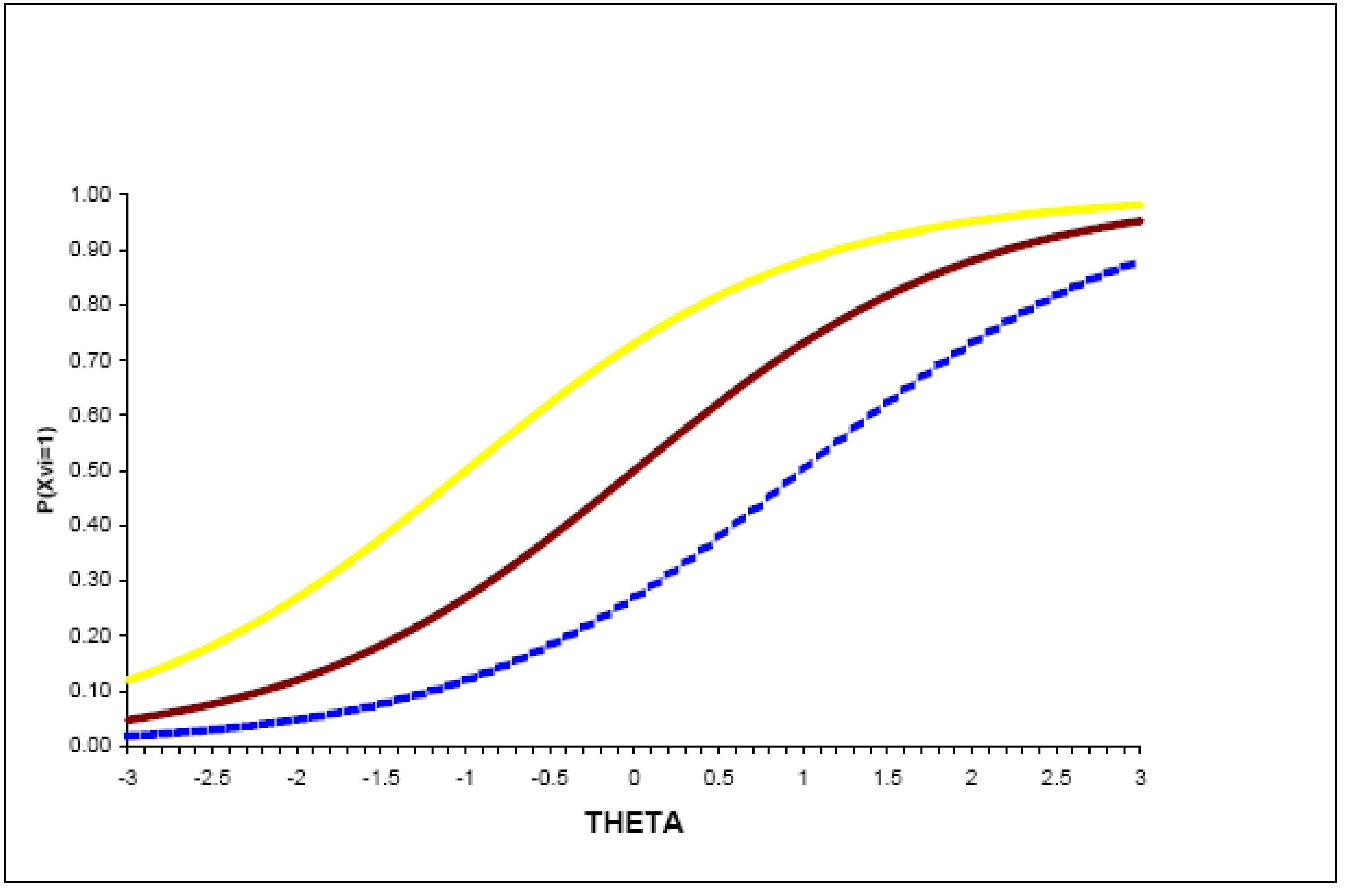
- Probabilité d'une matrice de réponse :

$$P(\mathbf{X}) = \prod_{v=1}^{N} \prod_{i=1}^{k} P(X_{vi} = x_{vi}) = \prod_{v=1}^{N} \prod_{i=1}^{k} \frac{e^{x_{vi}(\theta_v - \sigma_i)}}{1 + e^{(\theta_v - \sigma_i)}}$$

= fonction de vraisemblance

Graphiquement

Exemple d'estimation d'un paramètre de personne si on connaît les paramètres des item $\sigma_1 = -1$, $\sigma_2 = 0$, $\sigma_3 = 1$:



Fonction d’information

- Pour le modèle de Rasch, c’est la probabilité de répondre non fois la probabilité de répondre oui.
- Elle est la plus élevée au point d’inflexion de la courbe caractéristique (point où la probabilité vaut .5 et donc où la réponse est la plus incertaine)

Test d'adéquation du modèle

- Le plus connu: test de Pearson

$$PE = \Sigma\ (o_x - e_x)^2 / e_x$$

- Compare les effectifs attendus ayant un certain vecteur de réponse aux effectifs observés

6	6.00082	0 0 0 0 0 0 0 0 0 0 0 0 0 0	0.0000
2	0.04870	0 0 0 0 0 0 0 0 0 0 0 1 0 1	78.1866
1	0.19338	0 0 0 0 0 0 0 0 0 0 1 1 0 0	3.3644
1	1.26620	0 0 0 0 0 0 0 0 0 1 0 0 0 0	0.0560
1	0.24388	0 0 0 0 0 0 0 0 1 0 0 1 0	2.3443
3	2.17161	0 0 0 0 0 0 0 0 1 0 0 0 0 0	0.3160
1	0.18344	0 0 0 0 0 0 0 1 0 0 0 0 0 1	3.6348
2	0.41826	0 0 0 0 0 0 0 0 1 0 0 0 1 0	5.9816
1	0.63855	0 0 0 0 0 0 0 0 1 1 0 0 0 0	0.2046
1	0.28103	0 0 0 0 0 0 0 0 1 1 1 0 0 0	1.8394
2	0.69523	0 0 0 0 0 0 1 0 0 0 0 0 0 0	2.4487
1	0.02585	0 0 0 0 0 0 1 0 0 0 1 0 0 1	36.7166
2	0.35061	0 0 0 0 0 0 1 0 1 0 0 0 0 0	7.7594
1	0.02474	0 0 0 0 0 1 0 0 1 0 0 0 0 1	38.4495
1	0.00629	0 0 0 0 0 1 1 0 0 0 0 1 1 0	157.0542
...			
1	0.41288	1 1 1 1 1 1 1 1 1 1 1 1 0 0	0.8349
2	1.11387	1 1 1 1 1 1 1 1 1 1 1 1 0 1	0.7049
15	15.00054	1 1 1 1 1 1 1 1 1 1 1 1 1 1	0.0000

Fréquence observée

Fréquence attendue

Vecteur de réponse

$$\frac{(o_x - e_x)^2}{e_x}$$

Problème: test d’adéquation du modèle

- Il y a énormément de vecteurs de réponses possibles
- Les effectifs attendus sont très faibles donc la p-valeur du test de Pearson est fausse
- Solution:
  - Faire un bootstrap
  - Supprimer les valeurs de distance quand les valeurs attendues sont trop faibles
  - ??? À trouver

Mauvais ajustement des items: indice Q

$$Q_i = \frac{\ln \dfrac{p(\mathbf{x}_{\text{obs}})}{p(\mathbf{x}_{\text{max}})}}{\ln \dfrac{p(\mathbf{x}_{\text{min}})}{p(\mathbf{x}_{\text{max}})}}, \qquad 0 \le Q_i \le 1.$$

Exemple: 7 sujets, n_i = 4

$\hat{\theta}_v$	-2.5	-1.4	-1.0	0	0.9	1.4	2.4
x_{obs}	0	1	0	1	0	1	1
X_{min}	0	0	0	1	1	1	1
X_{max}	1	1	1	1	0	0	0

$Q_i = 0$, si $p(x_{obs)} = p(x_{max})$

$Q_i = 1$, si $p(x_{obs)} = p(x_{min})$

« Graded response model »

$$P(X_{vi} = x) = \frac{\exp(x\theta_v - \sigma_{ix})}{\sum_{s=0}^{m} \exp(s\theta_v - \sigma_{is})}$$

$$\sigma_{ix} = \sum_{s=0}^{x} \tau_{is},\ \sigma_{i0} = 0 \quad \text{et} \quad \sum_{i=1}^{k} \sum_{x=1}^{m} \tau_{ix} = 0$$

« Graded response model »

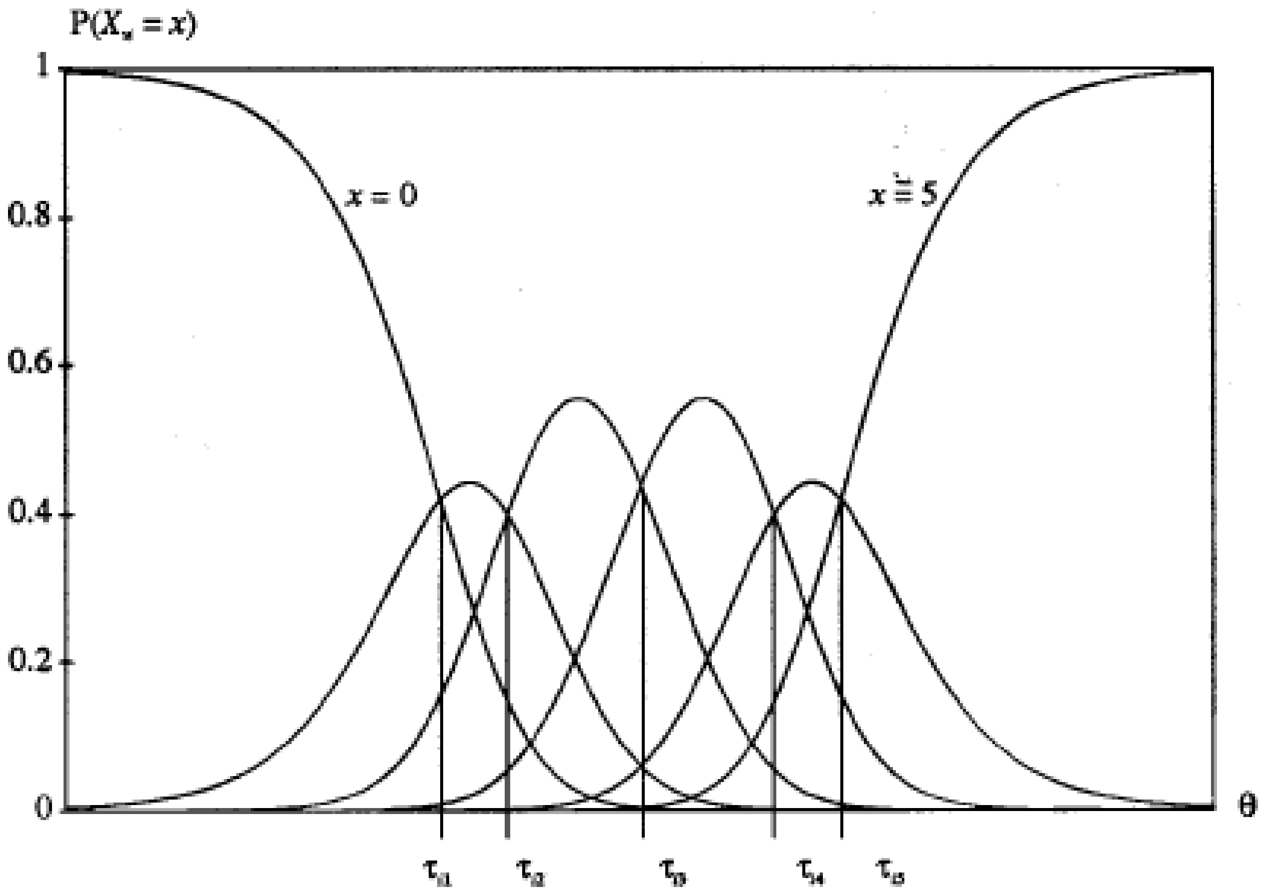


Figure 1. Category characteristic curves for an item with six categories and the threshold parameters $\tau_{i1} = -3$, $\tau_{i2} = -2$, $\tau_{i3} = 0$, $\tau_{i4} = 2$, and $\tau_{i5} = 3$.

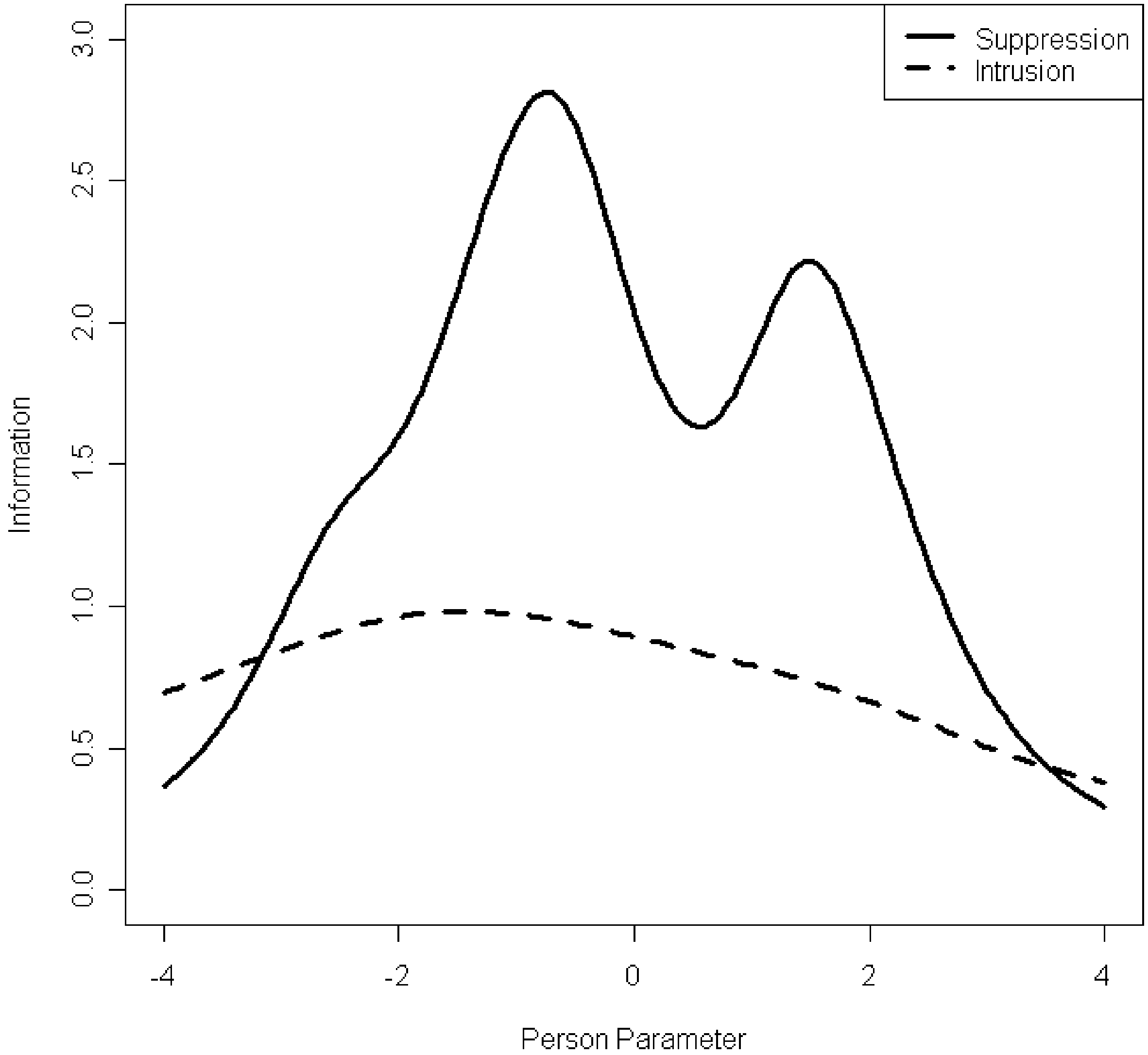
Exemple: réponse à l'item (graded response model)

Dimension	Modèle	Items	Pearson's χ^2	*df*	Boot *p*
Suppression	1A	1, 2, 8, 10-15	6512235.94	1953053	.03
	1B	1, 2, 10-15	788249.50	390561	.03
	1C	1, 2, 10-14	113986.74	78069	.06
	1D	1, 10-14	14758.42	15577	.27
Intrusion	2	3-7, 9	26545.11	15577	.10

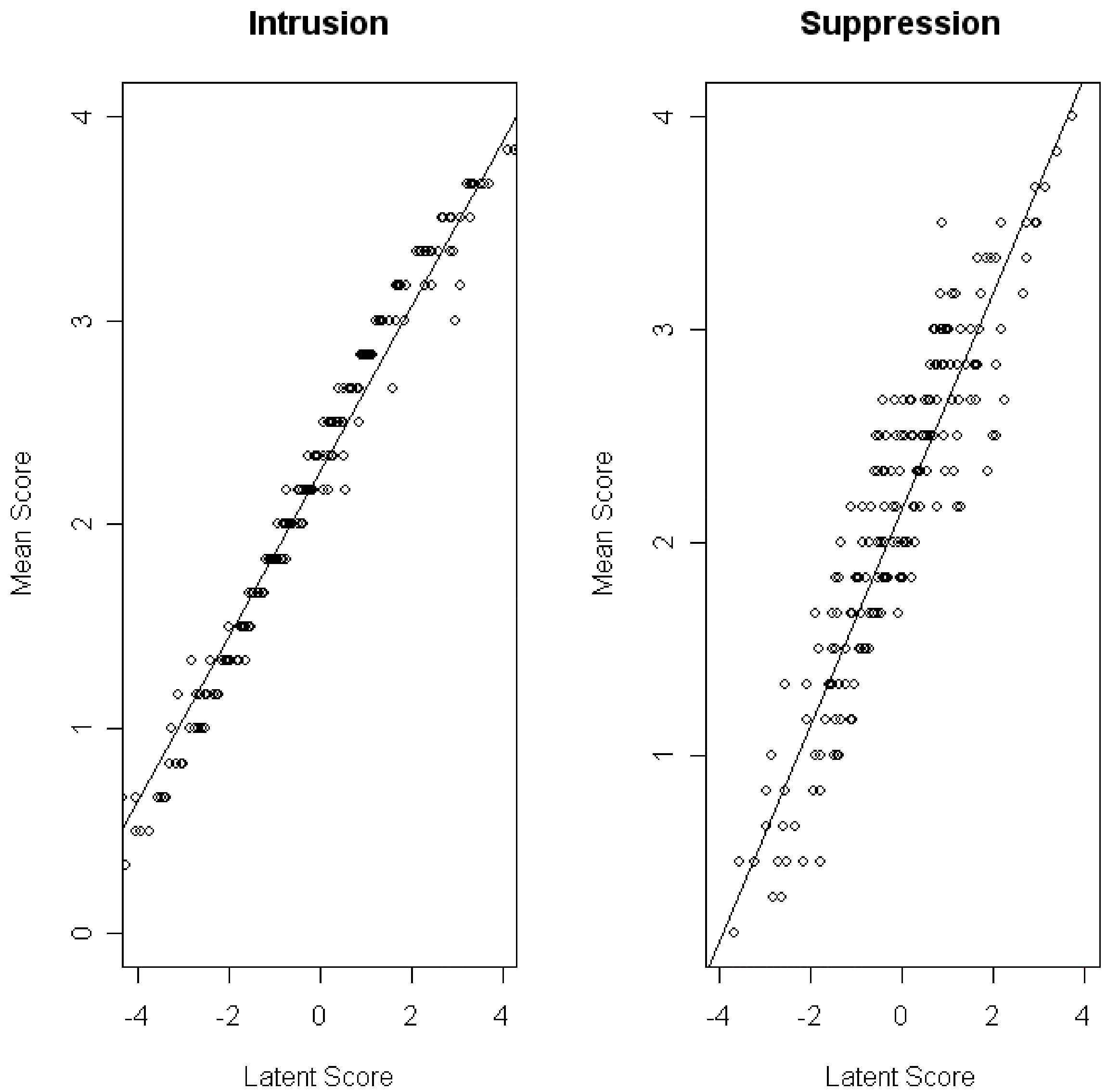
Exemple: indices Q

Question	suppression		intrusion	
1	.14	.36		
2				
3			.07	.62
4			.09	.46
5			.09	.49
6			.09	.46
7			.09	.40
8				
9			.08	.59
10	.14	.31		
11	.08	.83		
12	.11	.39		
13	.18	.18		
14	.08	.82		
15				

Exemple: courbe d'information



Exemple: lien latent-observé



Conclusions

- Pour chaque groupe d’items (suppression et intrusion), plusieurs items ne s’ajustent pas et causent un non-ajustement de l’ensemble du modèle.
- Une fois ces items enlevés, l’échelle de suppression est plus informative que l’échelle d’intrusion.
- L’échelle de suppression correspond moins à un simple score total (r=.93) que l’échelle d’intrusion.