Septembre 2003 **Numéro 33**

Cahiers de l'IMA

Designs de testage incomplets et modèle non-paramétrique de la réponse à l'item

Jean-Philippe Antonietti

Institut de Mathématiques Appliquées
Faculté des S.S.P.
Université de Lausanne
BFSH 2
1015 Lausanne

Designs de testage incomplets et modèle non-paramétrique de la réponse à l'item

Jean-Philippe Antonietti

Introduction

Supposons que l'on veuille connaître l'étendue des connaissances mathématiques des élèves romands ayant suivi quatre années d'école primaire. Pour ce faire, il suffit de proposer à un échantillon représentatif d'élèves un échantillon représentatif de tâches mathématiques. L'échantillonnage des élèves ne devrait pas poser de problèmes. Celui des tâches mathématiques est plus délicat et l'élaboration d'un test forcément très long pourrait conduire à une situation irréaliste. Il serait par exemple extravagant que l'épreuve dure toute une journée.
Un moyen de contourner cette difficulté consiste à fragmenter l'épreuve mathématique et ainsi de ne soumettre chaque élève de l'échantillon qu'à une portion de l'épreuve complète. Pour pouvoir néanmoins comparer les élèves les uns aux autres, le plan de testage doit être connexe. Cela signifie que chaque item doit apparaître dans deux cahiers différents au moins.
Montrons comment créer simplement un tel design de testage. Supposons que l'on désire créer K cahiers différents et que le nombre d'items de notre test soit égal à k. Nous commencerons par répartir les k items en K blocs distincts, chaque item n'appartiendra qu'à un et un seul bloc. Le premier cahier sera constitué des blocs 1 et 2, le deuxième des blocs 2 et 3, et ainsi de suite jusqu'au dernier cahier qui lui sera constitué du bloc K et du bloc 1. Les élèves, quant à eux, seront répartis en K groupes disjoints. Lors du testage, l'on proposera aux élèves du groupe 1 le cahier 1, aux élèves du groupe 2 le cahier 2 et finalement aux élèves du groupe K le cahier K (voir Figure 1).
Dans le cadre du modèle de Rasch, l'estimation des compétences mathématiques de chaque élève se fait aisément à partir de tels designs [1, 8]. Malheureusement les conditions qui doivent être remplies pour que l'on puisse véritablement se placer dans un tel cadre théorique ne sont, dans la pratique, quasiment jamais satisfaites.
Dans ce travail, nous montrerons qu'il est possible de traiter convenablement ces designs incomplets en recourant à un modèle de réponse à l'item non-paramétrique moins contraignant. Le plan que nous allons suivre est celui-ci :

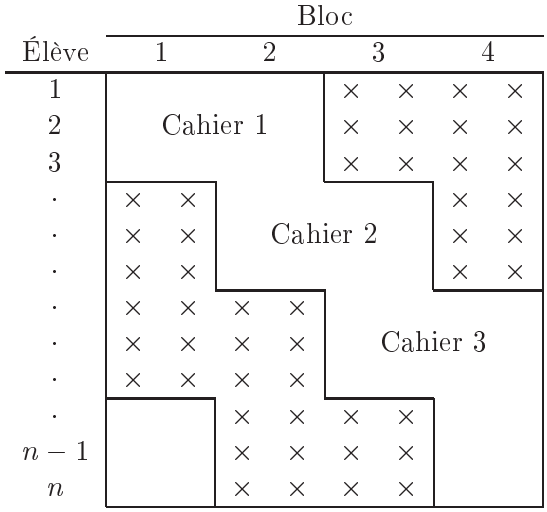

FIGURE 1 – *Schéma d'un design de testage incomplet connexe. Les croix* $\times$ *représentent des données manquantes par design.*

nous commencerons par présenter un modèle non-paramétrique de la réponse à l'item. Puis, dans le cadre de ce modèle, nous montrerons comment traiter les designs de testage incomplets : nous proposerons une méthode d'imputation des données manquantes basée sur les propriétés du modèle. Nous validerons ensuite la nouvelle méthode en comparant les résultats qu'elle fournit à ceux que l'on obtient à l'aide du modèle paramétrique de Rasch. Nous terminerons en montrant comment certaines caractéristiques du design influencent la précision du classement des élèves.

1 Modèle non-paramétrique de la réponse à l'item

Le modèle que nous allons proposer s'appuie sur l'idée générale qu'une réponse correcte à un item est déterminée, d'une part, par les compétences mathématiques de l'élève et, d'autre part, par les caractéristiques de l'item telles que sa difficulté et sa discrimination. Pour caractériser un modèle, il suffit donc de définir la courbe caractéristique de chaque item, c'est-à-dire la fonction qui lie la probabilité de réussite d'un item au trait latent, en l'occurrence aux compétences mathématiques.
Définissons à titre d'exemple le modèle de Rasch, autrement dit le modèle logistique à un paramètre. Soit la variable aléatoire X_j qui représente le score à l'item j. Ce score est égal à 0 si la réponse à l'item j est incorrecte et à 1 si la réponse est correcte. La courbe caractéristique de l'item $P_j(\theta)$ est définie par la fonction :

$$P_j(\theta) = P(X_j = 1|\theta) = \frac{e^{(\theta-\delta_j)}}{1+e^{(\theta-\delta_j)}}. \tag{1}$$

Cette fonction exprime la manière dont la probabilité de réussite dépend de la valeur du trait latent θ. Cette fonction est monotone croissante : plus les compétences mathématiques sont grandes, plus la probabilité de répondre correctement à l'item j est élevée.

$$\text{Si } \theta_1 \leq \theta_2, \text{ alors } P(X_j = 1|\theta_1) \leq P(X_j = 1|\theta_2). \tag{2}$$

Les items se distinguent par un seul paramètre δ_j qui représente leur difficulté. δ_j permet de localiser la fonction caractéristique le long du trait latent; lorsque $\theta = \delta_j$, la probabilité de répondre correctement à l'item j vaut 0.5. Nous avons représenté dans la Figure 2 quelques courbes caractéristiques satisfaisant le modèle de Rasch.

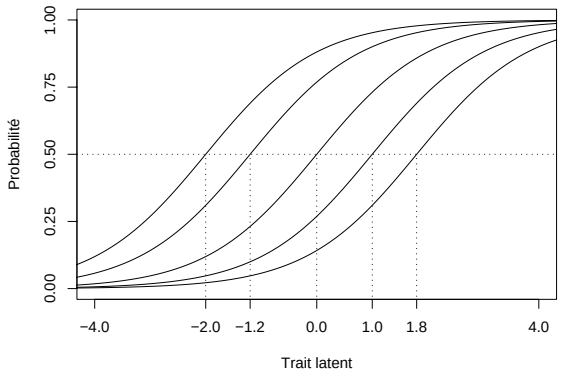

FIGURE 2 – *Quelques courbes caractéristiques de l'item satisfaisant le modèle de Rasch.*

Le modèle que nous allons proposer maintenant est beaucoup moins restrictif que le modèle de Rasch [5, 6, 7, 9]. Il ne se fonde que sur quatre hypothèses que le modèle de Rasch satisfait aussi forcément.

1.1 Contraintes du modèle

1.1.1 Unidimensionnalité

La première hypothèse qui doit être satisfaite est l'unidimentionnalité. Cela signifie que tous les items du test doivent contribuer à un seul trait latent, celui que l'on cherche à mesurer à l'aide du test. Dans notre exemple, cela sous-entend que les compétences mathématiques sont monolithiques et ne se décomposent pas en compétences numériques et spatiales.

1.1.2 Indépendance locale

La deuxième hypothèse impose l'indépendance locale. Ainsi la réponse d'un élève à une première question ne devrait pas influencer ses réponses ultérieures. Cette hypothèse exclut donc qu'un élève puisse apprendre quoi que ce soit lors de la passation du test ! Cette hypothèse est très forte.
Soit $\boldsymbol{X} = (X_1,\ X_2,\ \ldots,\ X_k)$ le vecteur des variables des scores des items et $\boldsymbol{x} = (x_1,\ x_2,\ \ldots,\ x_k)$ l'une des réalisations de $\boldsymbol{X}$. Les items x_j étant dichotomiques, ils prennent comme valeur 0 ou 1. La probabilité qu'un élève ayant un niveau de compétences mathématiques θ obtienne le score x_j à l'item j vaut $P(X_j = x_j|\theta)$. L'indépendance locale se traduit formellement par l'égalité suivante :

$$P(\boldsymbol{X} = \boldsymbol{x}|\theta) = \prod_{j=1}^{k} P(X_j = x_j|\theta). \tag{3}$$

1.1.3 Monotonicité des courbes caractéristiques des items

L'hypothèse suivante stipule que les courbes caractéristiques des items sont monotones croissantes.
Supposons que l'élève 2 soit plus compétent en mathématiques que l'élève 1. Dans ces conditions la probabilité que l'élève 2 réponde correctement à l'item j est supérieure ou égale à la probabilité que l'élève 1 y réponde correctement :

$$\text{si } \theta_1 \leq \theta_2, \text{ alors } P_j(\theta_1) \leq P_j(\theta_2) \tag{4}$$

et ceci quel que soit l'item j.

1.1.4 Non-intersection des courbes caractéristiques des items

Si l'on ne s'intéresse qu'aux élèves, l'hypothèse de non-intersection des courbes caractéristiques est superflue. Par contre, si l'on veut pouvoir aussi ordonner les items en fonction de leur difficulté, cette hypothèse est indispensable. Selon cette hypothèse, aucune des k courbes caractéristiques des items ne se coupent. Plus précisément cela signifie que, quel que soit θ, les courbes caractéristiques des items peuvent être ordonnées et numérotées toujours de la même façon :

$$(\forall \theta \in \mathbb{R})\ P_1(\theta) \leq P_2(\theta) \leq \ldots \leq P_k(\theta). \tag{5}$$

L'item 1 est alors le plus difficile, puis vient l'item 2 et ainsi de suite jusqu'à l'item k, le plus facile. La Figure 3 représente quatre courbes caractéristiques monotones croissantes qui ne se coupent pas.

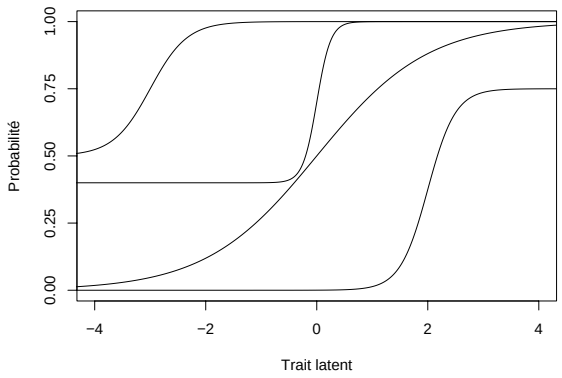

FIGURE 3 – *Quelques courbes caractéristiques monotones croissantes qui ne se coupent pas.*

Dérivons maintenant deux propriétés importantes de notre modèle.

1.2 Propriétés du modèle

1.2.1 Sériation des compétences à partir des performances

Nous allons montrer que si les hypothèses d'unidimensionnalité, d'indépendance locale et de monotonicité sont satisfaites, alors il est possible de classer les élèves le long du trait latent θ à partir de la connaissance de leur score global X^+ au test, où :

$$X^+ = \sum_{j=1}^{k} X_j. \tag{6}$$

Choisissons deux entiers a et b compris entre 0 et k, tel que a soit plus petit que b. Si les conditions d'unidimensionnalité, d'indépendance locale et de monotonicité sont satisfaites, alors la fonction de répartition caractérisant la distribution des compétences des élèves ayant obtenu un score global égal à a est toujours supérieure ou égale à celle qui caractérise la distribution des compétences des élèves ayant obtenu un score global égal à b [3, lemme 4.1]. Autrement dit :

$$(\forall 0 \leq a < b \leq k)\ (\forall c \in \mathbb{R})\ P(\theta \leq c | X^+ = a) \geq P(\theta \leq c | X^+ = b). \qquad (7)$$

Rappelons que l'espérance mathématique d'une variable statistique X est égale à l'aire algébrique délimitée par la courbe cumulative, les axes de coordonnées et la droite d'ordonnée 1 (voir Figure 4).

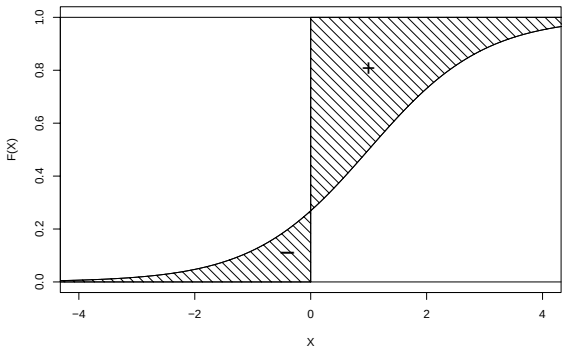

FIGURE 4 – *Évaluation graphique de la moyenne.*

Ce rappel nous permet de calculer la moyenne des compétences mathématiques des élèves ayant obtenu un score global $X^+ = a$:

$$\begin{aligned} E(\theta | X^+ = a) &= \int_0^\infty [1 - P(\theta \leq x | X^+ = a)]\ dx \\ &\quad - \int_{-\infty}^0 P(\theta \leq x | X^+ = a)\ dx. \end{aligned} \qquad (8)$$

La moyenne des compétences mathématiques des élèves ayant obtenu un score global $X^+ = b$ se calcule de la même manière :

$$\begin{aligned} E(\theta | X^+ = b) &= \int_0^\infty [1 - P(\theta \leq x | X^+ = b)]\ dx \\ &\quad - \int_{-\infty}^0 P(\theta \leq x | X^+ = b)\ dx. \end{aligned} \qquad (9)$$

La différence des espérances conditionnelles vaut donc :

$$E(\theta|X^+ = b) - E(\theta|X^+ = a) =$$
$$\int_0^\infty [P(\theta \leq x|X^+ = a) - P(\theta \leq x|X^+ = b)]\ dx$$
$$- \int_{-\infty}^0 [P(\theta \leq x|X^+ = b) - P(\theta \leq x|X^+ = a)]\ dx. \tag{10}$$

Étant donné la relation (7), la première intégrale est positive ou nulle et la seconde négative ou nulle. Leur différence est donc positive ou nulle. Ainsi :

$$(\forall 0 \leq a < b \leq k)\ E(\theta|X^+ = a) \leq E(\theta|X^+ = b). \tag{11}$$

C'est exactement ce que nous voulions démontrer. Il est donc possible d'inférer l'ordre des compétences θ des élèves à partir de l'ordre de leurs performances X^+. Mais contrairement à ce qu'il est possible de faire avec un modèle de Rasch, nous ne pouvons pas estimer numériquement les compétences des élèves. Nous ne pouvons construire, à l'aide d'un modèle non-paramétrique, qu'une échelle ordinale des compétences des élèves et non pas une échelle d'intervalle comme dans un modèle de Rasch.

1.2.2 Ordination des items selon leur difficulté

Remarquons que l'espérance mathématique de la variable X_j conditionnellement à θ est égale à la valeur de la fonction caractéristique de l'item j en θ :

$$\begin{aligned} E(X_j|\theta) &= 0 \times P(X_j = 0|\theta) + 1 \times P(X_j = 1|\theta) \\ &= P(X_j = 1|\theta) \\ &= P_j(\theta). \end{aligned} \tag{12}$$

En remplaçant les probabilités conditionnelles $P_j(\theta)$ par les espérances conditionnelles $E(X_j|\theta)$ dans l'équation (5) qui définit la non-intersection des courbes caractéristiques des items, nous obtenons :

$$(\forall \theta \in \mathbb{R})\ E(X_1|\theta) \leq E(X_2|\theta) \leq \ldots \leq E(X_k|\theta). \tag{13}$$

Ainsi, si les hypothèses de notre modèle sont satisfaites, quelle que soit la valeur de θ, les scores moyens des items s'ordonnent de la même manière. Malencontreusement, θ étant une variable latente, les espérances conditionnelles ne peuvent pas être observées. Comment donc estimer l'ordre des items ? Supposons que dans la population le trait latent se distribue selon la densité de probabilité $f(\theta)$. La proportion des réponses correctes à l'item j dans la population vaut donc :

$$P_j = \int_\theta E(X_j|\theta)\ f(\theta)\ d\theta = \int_\theta P_j(\theta)\ f(\theta)\ d\theta. \tag{14}$$

Calculons la différence entre P_i et P_j :

$$\begin{aligned} P_i - P_j &= \int_\theta P_i(\theta)\ f(\theta)\ d\theta - \int_\theta P_j(\theta)\ f(\theta)\ d\theta \\ &= \int_\theta [P_i(\theta) - P_j(\theta)]\ f(\theta)\ d\theta. \end{aligned} \tag{15}$$

Admettons que la courbe caractéristique de l'item i se trouve en dessous de celle de l'item j :

$$(\forall\theta)\ P_i(\theta) \leq P_j(\theta). \tag{16}$$

Il en découle que :

$$P_i - P_j \leq 0. \tag{17}$$

Ainsi l'ordre des proportions de réussite est le même que celui des espérances conditionnelles :

$$\text{si } (\forall\theta \in \mathbb{R})\ E(X_i|\theta) \leq E(X_j|\theta), \text{ alors } P_i \leq P_j. \tag{18}$$

Si les courbes caractéristiques des items ne se croisent pas, la réciproque est aussi vraie :

$$\text{si } P_i \leq P_j, \text{ alors } (\forall\theta \in \mathbb{R})\ E(X_i|\theta) \leq E(X_j|\theta). \tag{19}$$

Finalement donc, si les hypothèses du modèle sont satisfaites, l'ordre des espérances conditionnelles des items peut être approximé par l'ordre des taux de réussite globaux observés :

$$P_1 \leq P_2 \leq \ldots \leq P_k. \tag{20}$$

2 Traitement des designs de testage incomplets

2.1 Procédure

Nous allons maintenant décrire comment traiter des designs de testage incomplets dans le cadre de notre modèle. Nous proposerons une méthode d'imputation des données manquantes basées sur les deux propriétés du modèle que nous venons de présenter [2].
Notons les scores obtenus par un élève aux différents items du test $x_1, x_2, \ldots, x_k$. Le score x_j prend comme valeur 0 si la réponse à l'item j est fausse, 1 si la réponse est juste, et 9 si la question n'a pas été posée. L'imputation des valeurs manquantes se fait selon la procédure suivante :

1. Ordonner de manière décroissante les items selon la proportion de réponses correctes en ne tenant compte naturellement que des questions qui ont été posées. Les items sont donc classés du plus facile au plus difficile. Renumérotons les items selon leur position dans ce classement $[1], [2], \ldots, [k]$.
2. Pour donner une valeur à $x_{[j]}$, il suffit d'appliquer la première des règles figurant dans la liste qui suit dont la condition est satisfaite :
   (a) Si $x_{[j+1]} = 1$, alors $x_{[j]} = 1$.
   Si l'on n'a pas posé la question $[j]$ à un élève, mais que cet élève a résolu correctement la question $[j+1]$ qui est à peine plus difficile que la question $[j]$, alors nous supposerons que si la question $[j]$ lui avait été posée, il y aurait répondu correctement.
   (b) Si $x_{[j-1]} = 0$, alors $x_{[j]} = 0$.
   Si l'on a pas posé la question $[j]$ à un élève, mais que cet élève a répondu de manière erronée à la question $[j-1]$ qui est à peine plus facile que la question $[j]$, alors nous supposerons que si la question $[j]$ lui avait été posée, il n'y aurait pas répondu correctement.
   (c) Si parmi les scores $x_{[1]}, x_{[2]}, \ldots, x_{[j-1]}$ le nombre de 0 est supérieur ou égal au nombre de 1, alors $x_{[j]} = 0$.
   Si, parmi les questions plus faciles que la question $[j]$, un élève ne répond pas correctement à une majorité d'entre elles, alors nous supposerons que l'élève n'aurait pas répondu correctement à la question $[j]$.
   (d) Si parmi les scores $x_{[j+1]}, \ldots, x_{[k]}$ le nombre de 1 est supérieur ou égal au nombre de 0, alors $x_{[j]} = 1$.
   Si, parmi les questions plus difficiles que la question $[j]$, un élève répond correctement à la moitié ou même davantage, alors nous supposerons qu'il aurait répondu correctement à la question $[j]$.
   (e) Tirer au sort la valeur de $x_{[j]}$ en se conformant à la distribution de $X_{[j]}$.
   Si l'on ne peut tirer aucune information à partir du motif des réponses d'un élève, alors n'utiliser que l'information concernant l'item.

Cette procédure permet de construire un jeu de données complet aussi proche que possible du modèle déterministe de Guttman.

2.2 Application

Complétons à titre d'exemple le jeu de données suivant :

	Item							
Élève	1	2	3	4	5	6	7	8
1	9	1	1	9	0	1	0	9
2	9	0	1	9	0	0	1	9
3	9	1	1	9	0	0	0	9
4	9	0	1	9	1	0	0	9
5	0	9	1	1	0	1	9	1
6	1	9	0	0	0	0	9	1
7	1	9	1	1	0	0	9	0
8	1	9	1	1	0	1	9	0
9	1	1	9	1	9	9	0	1
10	1	0	9	1	9	9	1	1
11	0	0	9	0	9	9	0	0
12	0	1	9	1	9	9	0	1
Difficulté	3/8	4/8	1/8	2/8	7/8	5/8	6/8	3/8

Ordonnons les items selon leur difficulté et appliquons à chaque donnée manquante la première règle dont la condition est satisfaite. Dans le tableau ci-dessous les règles appliquées sont indiquées entre parenthèse.

	Item							
Élève	3	4	8	1	2	6	7	5
1	1	(d)→ 1	(d)→ 1	(a)→ 1	1	1	0	0
2	1	(e)→	(e)→	(e)→	0	0	1	0
3	1	(e)→	(e)→	(a)→ 1	1	0	0	0
4	1	(e)→	(e)→	(e)→	0	0	0	1
5	1	1	1	0	(a)→ 1	1	(e)→	0
6	0	0	1	1	(c)→ 0	0	(b)→ 0	0
7	1	1	0	1	(e)→	0	(b)→ 0	0
8	1	1	0	1	(a)→ 1	1	(e)→	0
9	(a)→ 1	1	1	1	1	(e)→	0	(b)→ 0
10	(a)→ 1	1	1	1	0	(a)→ 1	1	(e)→
11	(e)→	0	0	0	0	(b)→ 0	0	(b)→ 0
12	(a)→ 1	1	1	0	1	(e)→	0	(b)→ 0
Difficulté	1/8	2/8	3/8	3/8	4/8	5/8	6/8	7/8

Effectuons les tirages au sort et réordonnons les items du test. L'on obtient ainsi le jeu de données complet suivant :

	Item							
Élève	1	2	3	4	5	6	7	8
1	1	1	1	1	0	1	0	1
2	1	0	1	1	0	0	1	0
3	1	1	1	0	0	0	0	1
4	0	0	1	1	1	0	0	1
5	0	1	1	1	0	1	1	1
6	1	0	0	0	0	0	0	1
7	1	0	1	1	0	0	0	0
8	1	1	1	1	0	1	0	0
9	1	1	1	1	0	0	0	1
10	1	0	1	1	0	1	1	1
11	0	0	1	0	0	0	0	0
12	0	1	1	1	0	0	0	1

3 Validation de la méthode d'imputation

Afin de nous faire une idée de l'efficacité de la méthode de traitement des designs de testage incomplets que nous avons proposée, nous allons l'appliquer à une situation artificielle mais plausible et confronter les résultats que l'on obtient grâce à cette méthode à ceux que l'on obtient à l'aide d'un modèle de Rasch.

3.1 Données

D'une population théorique infinie d'élèves, nous avons tiré aléatoirement un échantillon de taille 2016. Nous avons supposé que les compétences mathématiques de ces élèves se distribuaient normalement le long d'un seul trait latent θ. De manière arbitraire, nous avons fixé la moyenne théorique des compétences mathématiques à 0 et l'écart-type à 1. Afin d'évaluer les compétences de ces élèves nous leur avons proposé un test constitué de 24 items. Nous voulions que ce test soit relativement facile. Ainsi nous avons tiré les questions de l'épreuve d'une banque d'items dont la difficulté moyenne valait -1 et l'écart-type 1. Nous avons décidé que les données se conformeraient au modèle de Rasch. Ainsi la probabilité qu'un élève ayant un niveau de compétences mathématiques égal à θ_i réponde correctement à une question de difficulté δ_j est définie par :

$$P(X_j = 1|\theta = \theta_i) = \frac{e^{(\theta_i - \delta_j)}}{1 + e^{(\theta_i - \delta_j)}}. \tag{21}$$

Le tableau de données complet $\boldsymbol{X}$ est de dimension 2016×24, il ne contient que des 0 et des 1.
Pour valider notre méthode de traitement des designs incomplets, nous avons créé 4 cahiers selon la procédure décrite en introduction. Les items ont été répartis de manière à ce que les cahiers soient tous *grosso modo* de la même difficulté. L'échantillon, quant à lui, a été scindé aléatoirement en 4 groupes de même taille – sans rien connaître des élèves, cette procédure est la plus raisonnable. Au groupe 1 nous avons proposé le cahier 1, au groupe 2 le cahier 2, au groupe 3 le 3 et au groupe 4 le 4.
Ce design nous a permis de créer un masque que l'on a appliqué au tableau de données complet. Une moitié des données a donc été estompée. Nous allons maintenant traiter ce nouveau tableau de données partiellement masqué $\boldsymbol{X}^\circ$.

3.2 Traitements

3.2.1 Modèle non-paramétrique de la réponse à l'item

Appliquons aux données manquantes du tableau lacunaire $\boldsymbol{X}^\circ$ la méthode d'imputation décrite au paragraphe 2. Nous obtenons ainsi un tableau de

données complet $\boldsymbol{X}^{\bullet}$ à partir duquel il est possible d'estimer les scores globaux au test $\hat{X}_1^+ = \sum_{j=1}^{k} X_j^{\bullet}$ et le classement des élèves selon leurs compétences mathématiques.

3.2.2 Modèle de Rasch

À partir des données $\boldsymbol{X}^{\circ}$ nous avons estimé, d'une part, les compétences mathématiques des élèves $\hat{\theta}$ et, d'autre part, la difficulté des items $\hat{\delta}$. Nous avons utilisé comme méthode d'estimation la méthode du maximum de vraisemblance inconditionnelle [8]. Nous avons également estimé le score global au test $\hat{X}_2^+$. Pour prédire le score global d'un élève ayant un niveau de compétences égal à $\hat{\theta}_i$, nous avons simplement additionné à ses scores observés l'espérance mathématique des items auxquels il n'a pas été soumis :

$$\hat{X}_{2i}^+ = \sum_{j \in C_i} x_{ij} + \sum_{j \in \overline{C_i}} E(X_{ij}) \tag{22}$$

avec :

$$E(X_{ij}) = P(X_{ij} = 1 | \theta = \hat{\theta}_i,\, \delta = \hat{\delta}_j) = \frac{e^{(\hat{\theta}_i - \hat{\delta}_j)}}{1 + e^{(\hat{\theta}_i - \hat{\delta}_j)}} \tag{23}$$

où C_i est l'ensemble des items appartenant au cahier que reçut l'élève i.

3.3 Résultats

3.3.1 Comparaison des estimations des scores globaux

Calculons les erreurs d'estimation faites, d'une part, dans le cadre du modèle non-paramétrique de la réponse à l'item $\Delta X_1^+ = \hat{X}_1^+ - X^+$ et, d'autre part, dans le cadre du modèle de Rasch $\Delta X_2^+ = \hat{X}_2^+ - X^+$ (voir Figure 5 et Tableau 1).

	Modèle non-paramétrique	Modèle de Rasch
Moyenne	0.51	-0.02
Écart-type	2.40	2.06

TABLEAU 1 – *Erreurs d'estimation faites sur les scores globaux.*

Dans le cadre du modèle non-paramétrique, les scores globaux sont légèrement surestimés ($t = 9.52$, $ddl = 2015$, $p < 0.05$). Ceci est vraisemblablement dû à d'ordre dans lequel les règles d'imputation sont appliquées. Dans le cadre du modèle de Rasch, les estimations sont non biaisées ($t = -0.50$, $ddl = 2015$, $p > 0.05$) et légèrement plus précises ($F = 1.36$, $ddl_1 = 2015$, $ddl_2 = 2015$, $p < 0.05$).

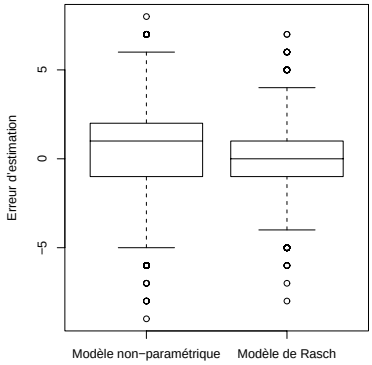

FIGURE 5 – *Distribution des erreurs d'estimation faites, d'une part, dans le cadre du modèle non-paramétrique de la réponse à l'item et, d'autre part, dans le cadre du modèle de Rasch.*

Ainsi les estimations faites dans le cadre du modèle non-paramétrique de la réponse à l'item sont légèrement moins bonnes que celles faites dans le cadre du modèle de Rasch. Ceci n'est pas surprenant car, rappelons-le, les données d'origine se conforment au modèle de Rasch. Soulignons néanmoins qu'à toute fin utile, le modèle non-paramétrique est largement suffisant. Représentons pour mieux s'en convaincre les scores estimés $\hat{X}_1^+$ et $\hat{X}_2^+$ en fonction des scores vrais (voir Figure 6). Le coefficient de corrélation entre scores vrais et scores estimés vaut 0.883 pour le modèle non-paramétrique et 0.909 pour le modèle de Rasch, ce qui est tout à fait remarquable.

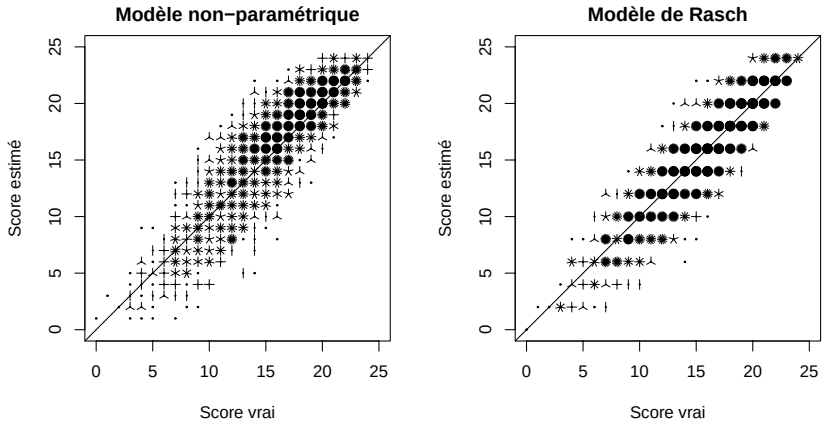

FIGURE 6 – *Diagrammes en tournesols représentant les scores estimés en fonction des scores vrais.*

3.3.2 Comparaison des classements selon les compétences

Calculons la valeur du coefficient de Spearman entre les compétences mathématiques vraies θ et les scores globaux estimés dans le cadre du modèle non-paramétrique $\hat{X}_1^+$:

$$r_S(\theta,\, \hat{X}_1^+) = 0.784. \tag{24}$$

Afin d'apprécier la qualité du classement obtenu par notre méthode non-paramétrique, calculons également la valeur du coefficient de Spearman entre les compétences mathématiques vraies θ et celles estimées dans le cadre du modèle de Rasch $\hat{\theta}$:

$$r_S(\theta,\, \hat{\theta}) = 0.799. \tag{25}$$

Nous constatons que les classements sont quasiment aussi bons l'un que l'autre. Ce résultat prouve que notre méthode est fiable !
Nous allons maintenant examiner comment certaines caractéristiques du design influencent la qualité des imputations.

4 Facteurs influençant la qualité des imputations

Afin de mieux cerner les limites de notre procédure d'imputation, nous avons effectué quelques simulations.

4.1 Facteurs

Les échantillons que nous avons étudiés – tous d'une même taille égale à 2016 – ont toujours été extraits de la même population d'élèves. Rappelons que dans cette population théorique les compétences mathématiques se distribuent normalement le long d'une seule échelle et que la moyenne de la distribution des compétences vaut 0 et l'écart-type 1. Notre étude ne porte pas sur les caractéristiques des échantillons mais exclusivement sur celles du test.

4.1.1 Difficulté du test

Nous avons construit un ensemble de tests faciles et un ensemble de tests à peine plus difficiles. Pour créer les tests faciles nous avons extrait les questions d'une banque d'items de difficulté moyenne égale à -1 et d'écart-type égal à 1. Ces tests s'apparentent aux tests d'acquisition où souvent l'on constate un entassement des scores au-dessus de la moyenne signalant ainsi qu'une minorité seulement des élèves éprouvent des difficultés.
Pour créer les tests moyennement difficiles nous avons utilisé une banque d'items ayant une difficulté moyenne de 0 et un écart-type de 1. Dans ce cas, comme la difficulté des items se distribuent de la même manière que les compétences des élèves, les tests sont plus discriminants.

4.1.2 Longueur du test

Nous avons construit des tests contenant 24 items et d'autres 48.

4.1.3 Nombre de cahiers

Les items ont été répartis successivement en 3, 4, 6 et 8 cahiers selon la procédure décrite dans l'introduction. Comme l'illustre le Tableau 2, le nombre K de cahiers formés influence directement la proportion P des données manquantes, en effet :

$$P = \frac{K-2}{K}. \tag{26}$$

4.1.4 Répartition des items dans les cahiers

Nous avons opté pour deux modes de répartition des items dans les cahiers. Le premier mode permet de créer des cahiers de difficultés similaires, le

Test court $k = 24$ items			Test long $k = 48$ items		
K	k/K	P [%]	K	k/K	P [%]
3	8	33	3	16	33
4	6	50	4	12	50
6	4	67	6	8	67
8	3	75	8	6	75

TABLEAU 2 – *Proportion P des données manquantes dans la matrice des données selon le nombre de cahiers K.*

second des cahiers de difficultés très différentes. L'attribution des items dans les différents blocs se fait selon le schéma suivant :

Composition de blocs de même difficulté							
Rang	$[1]$	$[2]$	$\ldots$	$[K]$	$[K+1]$	$\ldots$	$[2K]$
Bloc	$\pi_1(1)$	$\pi_1(2)$	$\ldots$	$\pi_1(K)$	$\pi_2(1)$	$\ldots$	$\pi_2(K)$
Rang	$[2K+1]$	$\ldots$	$[3K]$	$\ldots$	$[k-K+1]$	$\ldots$	$[k]$
Bloc	$\pi_3(1)$	$\ldots$	$\pi_3(K)$	$\ldots$	$\pi_{k/K}(1)$	$\ldots$	$\pi_{k/K}(K)$

Composition de blocs de difficultés très différentes							
Rang	$[1]$	$[2]$	$\ldots$	$[\frac{k}{K}]$	$[\frac{k}{K}+1]$	$\ldots$	$[2\frac{k}{K}]$
Bloc	1	1	$\ldots$	1	2	$\ldots$	2
Rang	$[2\frac{k}{K}+1]$	$\ldots$	$[3\frac{k}{K}]$	$\ldots$	$[k-\frac{k}{K}+1]$	$\ldots$	$[k]$
Bloc	3	$\ldots$	3	$\ldots$	K	$\ldots$	K

π_1, π_2, $\ldots$, $\pi_{k/K}$ sont des permutations aléatoires des K premiers entiers.

4.2 Plan des simulations

Dans les quatre situations définies par la difficulté du test (facile *versus* moyennement difficile) et sa longueur (court *versus* long) nous avons créé 20 jeux de données complets se pliant aux contraintes du modèle de Rasch. Chaque jeu de données a ensuite été troué de 8 façons différentes, puis reconstruit à l'aide de la procédure d'imputation décrite au paragraphe 2. Un tableau de données représentant les résultats de 2016 élèves ayant répondu à un test facile de 24 items, par exemple, a permis la création et le traitement de 8 tableaux de données incomplets différents. La structure du premier tableau est celle d'un design incomplet où les items sont répartis en 3 cahiers homogènes, la structure du deuxième est celle d'un design incomplet où les items sont répartis en 4 cahiers homogènes et ainsi de suite jusqu'au huitième dont la structure est celle d'un design incomplet où les items sont répartis en 8 cahiers hétérogènes (voir Tableau 3). Techniquement, le plan expérimental utilisé est un plan factoriel à parcelles partagées [4, p. 562].

Longueur : 24 items												
Difficulté : Faible						Difficulté : Moyenne					
Cahiers : Homogènes			Cahiers : Hétérogènes			Cahiers : Homogènes			Cahiers : Hétérogènes		
Nb. cahiers			Nb. cahiers			Nb. cahiers			Nb. cahiers		
3	...	8	3	...	8	3	...	8	3	...	8
e_{01}	...	e_{01}	e_{01}	...	e_{01}	e_{21}	...	e_{21}	e_{21}	...	e_{21}
⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋱	⋮
e_{20}	...	e_{20}	e_{20}	...	e_{20}	e_{40}	...	e_{40}	e_{40}	...	e_{40}

Longueur : 48 items												
Difficulté : Faible						Difficulté : Moyenne					
Cahiers : Homogènes			Cahiers : Hétérogènes			Cahiers : Homogènes			Cahiers : Hétérogènes		
Nb. cahiers			Nb. cahiers			Nb. cahiers			Nb. cahiers		
3	...	8	3	...	8	3	...	8	3	...	8
e_{41}	...	e_{41}	e_{41}	...	e_{41}	e_{61}	...	e_{61}	e_{61}	...	e_{61}
⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋱	⋮	⋮	⋱	⋮
e_{60}	...	e_{60}	e_{60}	...	e_{60}	e_{80}	...	e_{80}	e_{80}	...	e_{80}

TABLEAU 3 – *Plan expérimental des simulations. Les e_{01}, e_{02}, ..., e_{80} représentent les 80 échantillons différents sur lesquels nous avons travaillé.*

4.3 Résultats

Pour évaluer la qualité des imputations, nous avons simplement calculé, après avoir reconstitué la matrice des données $\boldsymbol{X}^{\bullet}$ et calculé l'estimation des scores globaux $\hat{X}^{+}$, le coefficient de corrélation de Spearman entre X^{+} et $\hat{X}^{+}$. Où, rappelons-le :

$$X^{+} = \sum_{j=1}^{k} X_j \text{ et } \hat{X}^{+} = \sum_{j=1}^{k} X_j^{\bullet}. \tag{27}$$

Reportons dans le Tableau 4 les effets principaux de chacun des facteurs selon la longueur du test (voir également la Figure 7).

	Longueur	
	24 items	48 items
Sur l'ensemble	0.805	0.858
Difficulté faible	0.789	0.844
Difficulté moyenne	0.821	0.871
3 cahiers	0.916	0.943
4 cahiers	0.850	0.895
6 cahiers	0.758	0.823
8 cahiers	0.696	0.769
Cahiers homogènes	0.842	0.898
Cahiers hétérogènes	0.768	0.818

TABLEAU 4 – *Effets principaux de certaines caractéristiques du design selon la longueur du test.*

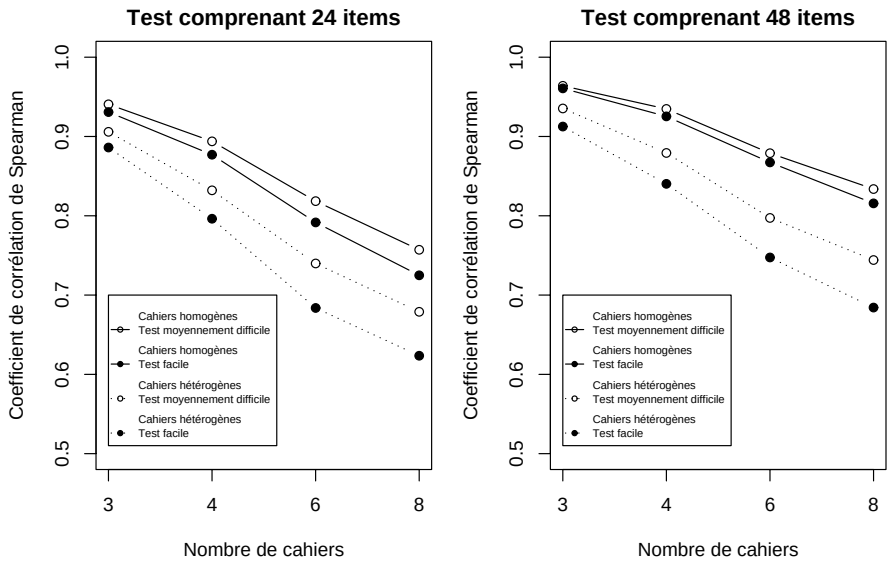

FIGURE 7 – *Effet de quelques facteurs sur la qualité des imputations.*

Les résultats des simulations sont cohérents et compréhensibles. Les quatre variables indépendantes de notre plan expérimental ont un effet significatif sur la similarité entre le classement établi à partir d'un tableau de données complet $\boldsymbol{X}$ et celui établi à partir d'un tableau tronqué $\boldsymbol{X}^{\circ}$.
Si le test est long, la précision des imputations est meilleure. Les imputations sont également de meilleure qualité si la difficulté du test coïncide avec les compétences des élèves; pour obtenir de bonne mesure, il ne faut donc ni un test trop facile, ni un test trop difficile.
Plus le nombre de cahiers est élevé, moins les élèves résolvent de problèmes et plus la qualité des mesures se détériore. Il est néanmoins raisonnable de distribuer les items dans 3 ou 4 cahiers même lorsque le test est court.
La répartition des items qui conduit à composer une série de cahiers homogènes est la plus avantageuse.
Aux effets principaux s'ajoutent également des effets d'interactions. Ainsi, par exemple, quand les cahiers sont *grosso modo* tous de la même difficulté, la difficulté moyenne du test n'a quasiment aucune influence sur la précision des classements. En revanche, quand les cahiers sont de difficultés très différentes, la précision d'un classement établi à partir d'un test facile est très inférieure à celle d'un classement établi à partir d'un test de difficulté moyenne (voir Tableau 5).

	ddl	*SC*	*MC*	*F*	*p*
Erreur : ECH					
LON	1	0.44113	0.44113	666.9182	< 0.0001
DIF	1	0.13602	0.13602	205.6436	< 0.0001
LON×DIF	1	0.00093	0.00093	1.4115	0.2385
Résidu	76	0.05027	0.00066		
Erreur : ECH×CAH					
CAH	3	3.6305	1.2102	12384.2934	< 0.0001
LON×CAH	3	0.0507	0.0169	172.8137	< 0.0001
DIF×CAH	3	0.0178	0.0059	60.8306	< 0.0001
LON×DIF×CAH	3	0.0005	0.0002	1.7613	0.1554
Résidu	228	0.0223	0.0001		
Erreur : ECH×REP					
REP	1	0.94209	0.94209	3835.1752	< 0.0001
LON×REP	1	0.00165	0.00165	6.7369	0.0113
DIF×REP	1	0.02778	0.02778	113.0765	< 0.0001
LON×DIF×REP	1	0.00143	0.00143	5.8185	0.0183
Résidu	76	0.01867	0.00025		
Erreur : intra					
CAH×REP	3	0.096636	0.032212	551.3834	< 0.0001
LON×CAH×REP	3	0.003223	0.001074	18.3925	< 0.0001
DIF×CAH×REP	3	0.002322	0.000774	13.2469	< 0.0001
LON×DIF×CAH×REP	3	0.000148	0.000049	0.8454	0.4703
Résidu	228	0.013320	0.000058		

TABLEAU 5 – *Résumé de l'analyse de variance. La variable ECH permet d'identifier les échantillons. Les variables indépendantes inter-parcelles sont LON (Longueur du test) et DIF (Difficulté du test); les variables indépendantes intra-parcelles sont CAH (Nombre de cahiers) et REP (Répartition des items).*

Conclusion

Les résultats que nous venons de présenter montrent clairement que les designs de testage incomplets peuvent être traités avec confiance à l'aide de la méthode décrite au paragraphe 2, méthode d'imputation fondée sur un modèle non-paramétrique de réponse à l'item. Le seul hiatus est que nous ne possédons pas encore d'outils qui nous permettent de savoir si une matrice de données incomplète satisfait ou non les conditions du modèle !

Bibliographie

[1] H. Hoijtink & A. Boomsma, *On person parameter estimation in the dichotomous Rasch model*, in Rasch models : Foundations, recent developments, and applications, G. H. Fischer & I. Molenaar, eds., Springer, New York, 1995, pp. 53–68.

[2] M. Huisman & I. W. Molenaar, *Imputation of missing scale data with item response models*, in Essays on item response theory, A. Boomsma, M. A. J. Van Duijn, & T. A. B. Snijders, eds., Springer, New York, 2001, pp. 221–244.

[3] B. W. Junker, *Conditional association, essential independence and monotone unidimensional item response models*, The Annals of Statistics, 21 (1993), pp. 1359–1378.

[4] R. E. Kirk, *Experimental design : Procedures for the behavioral sciences*, Books Cole, Pacific Grove, California, 1995.

[5] R. J. Mokken, *A theory and procedure of scale analysis with applications in political research*, De Gruyter, Mouton The Hague Paris, 1971.

[6] ——, *Nonparametric models for dichotomous responses*, in Handbook of modern item response theory, W. J. Van der Linden & R. K. Hambleton, eds., Springer, New York, 1997, pp. 351–367.

[7] R. J. Mokken & C. Lewis, *A nonparametric approach to the analysis of dichotomous item responses*, Applied Psychological Measurement, 6 (1982), pp. 417–430.

[8] I. W. Molenaar, *Estimation of item parameters*, in Rasch models : Foundations, recent developments, and applications, G. H. Fischer & I. Molenaar, eds., Springer, New York, 1995, pp. 39–51.

[9] K. Sijtsma & I. W. Molenaar, *Introduction to nonparametric item response theory*, Sage, Thousand Oaks, California, 2002.

Table des matières