Le testing adaptatif
Intérêts et applications

septembre 2008

Introduction

L'avènement des modèles de réponse à l'item (MRI) dans le domaine des sciences de l'éducation remonte aux années 60, avec les travaux précurseurs de G. Rasch (1960/1980) ou Birnbaum (1968). Leur intérêt est avant tout de fournir un modèle mathématique pour la réponse d'un individu soumis à un test d'évaluation composé d'items dont le format est strictement codifié. En cela, cette approche contraste avec la théorie plus « classique » des tests dans laquelle des principes tels que l'invariance de la mesure (ou objectivité spécifique) ou l'absence de fonctionnement différentiel des items ne sont pas nécessairement à la base de la réflexion. L'idée commune reste toutefois d'opérationnaliser un construct (ou concept hypothétique) et de le *mesurer* de manière quantitative. Lorsqu'un seul construct rentre en jeu, par exemple la compétence linguistique en français langue étrangère, on parle d'un modèle unidimensionnel. Les MRI permettent d'obtenir des paramètres d'items invariants des candidats sur lesquels ils ont été estimés, ce qui présente le double avantage de pouvoir composer des tests d'un niveau de difficulté donné et de délivrer en retour des scores individuels consistants et comparables entre eux. Ces scores doivent être directement interprétables sur une échelle de mesure standardisée et critériée.

Les tests d'évaluation proposés aux candidats sont la plupart du temps organisés de manière séquentielle, c'est-à-dire qu'il comporte un certain nombre de questions associant différentes modalités de réponse, avec généralement une seule réponse correcte, et qui sont présentées successivement. Le candidat effectue le test dans l'ordre prescrit ; les absences de réponse ou les réponses incorrectes ne sont pas pénalisées pour la plupart des épreuves certifiantes françaises et anglo-saxonnes. En fonction du nombre de réponses correctes fournies au test, le candidat se voit attribuer un score lui permettant de se situer par rapport aux autres candidats ou en relation avec des objectifs spécifiques (degré de maîtrise des connaissances ou aptitudes acquises dans un cycle de formation ou d'apprentissage). Dans ce cadre, le test est dit *linéaire* puisque le nombre et l'agencement des items dans le test est déterminé à l'avance et est identique pour tous les candidats.

Dans un test adaptatif au contraire, les items ne sont pas nécessairement présentés de manière linéaire et leur occurence dépend alors du niveau estimé de l'habileté du candidat. Ces tests sont communément administrés sur ordinateur, et on parlera de

test adaptatif sur ordinateur (TAO). L'estimation du paramètre d'habileté est réalisée dynamiquement, entre chaque essai, et elle doit être « optimisée » afin de ne pas perturber le déroulement du programme. La section suivante détaille les modalités de mise en œuvre ainsi que les principes généraux de l'organisation de ce type de test. On trouvera deux ouvrages complets sur ce sujet (cf. bibliographie, **page 12**). Les tests adaptatifs sont utilisés par de nombreux organismes certificateurs de par le monde, par exemple ETS (Educational Testing Service). Leur usage est généralement réservé à des tests d'entraînement, de sélection ou d'auto-évaluation. Dans le contexte d'un test de placement, le test adaptatif présente évidemment de nombreux avantages parmi lesquels on peut citer : la souplesse d'administration, l'obtention immédiate des résultats, une précision de mesure accrue, la possibilité d'utiliser les informations obtenues pour alimenter une banque d'items. En revanche, comme on le verra dans les sections suivantes, la réalisation d'un test adaptatif pose de nombreux problèmes liés aux contraintes associées à son déroulement (longueur du test, maximisation de l'information et choix des items, niveau d'exposition des items, mise à disposition d'une banque d'items, etc.).

Les principaux défis techniques posés par les tests adaptatifs sont donc : l'estimation du niveau du candidat et la procédure de sélection des items. Le présent document se propose de donner quelques repères bibliographiques et technologiques concernant ces deux aspects.

Principe général du TAO

Au concept même de test adaptatif est étroitement associé celui de banque d'items, puisqu'il est nécessaire de tenir à disposition du système de test un grand nombre d'items. On trouvera dans Devouche (2003) les principes généraux ayant guidé l'élaboration de la banque d'items du Test de connaissance du français (TCF). Le système doit pouvoir sélectionner des items en fonction de différents critères, en particulier le niveau estimé du candidat ainsi que les contraintes afférentes au test lui-même (contenu, nature et format des items, etc.). Le test est organisé sur un mode séquentiel, avec comme principales étapes : la sélection d'un item, l'estimation du niveau du candidat, l'évaluation des conditions d'arrêt (**Figure 1**).

En règle générale, plusieurs types de contraintes doivent être prises en compte. Cela se révèle d'autant plus important selon que le test est délivré à titre d'entraînement, dans une optique formative, voire avec une visée certificative. Parmi ces contraintes, mentionnons seulement les contraintes d'ordre psychométrique et celles relevant des aspects plus en lien avec l'évaluation elle-même, les contraintes de contenu. Les contraintes psychométriques sont assez simples : il s'agit de sélectionner le prochain item de sorte qu'il apporte le maximum d'information sur le niveau du candidat lorsque celui-ci aura fourni sa réponse. En d'autres termes, on sélectionnera l'item qui

présente la plus faible erreur de mesure et dont le niveau de difficulté est adapté au niveau d'habileté estimé préalablement. En ce qui concerne les contraintes de contenu, elles sont évidemment assez variables puisqu'elles dépendent en grande partie des objectifs spécifiques du test, des contraintes liées à l'administration et au contenu de la banque d'items, et de nombreux autres paramètres. Quelques exemples de contraintes de contenu sont fournies dans Swanson & Stocking (1993), par exemple. Les solutions informatiques proposées par ces auteurs sont exposées très brièvement dans la section suivante, ainsi que les méthodes plus générales reposant sur la maximisation de l'information.

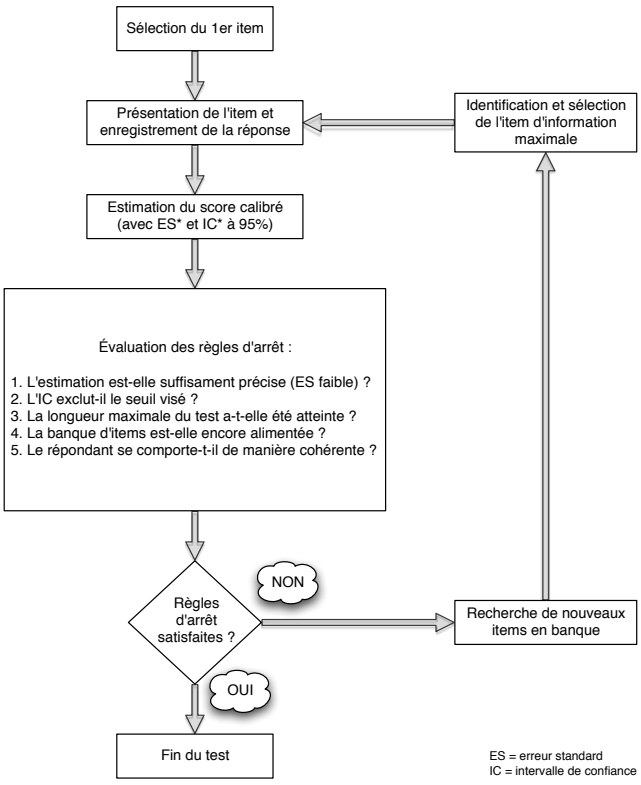


Figure 1 Organisation général d'un TAO.

Les performances, en termes d'estimation de l'habileté du candidat, sont généralement assez bonnes. L'exemple reproduit dans la **Figure 2** illustre à quel point les estimations obtenues à partir de 5 items ou moins sont cohérentes par rapport à celles obtenues sur un test incluant 53 items (Ware et al., 2000).

Formalisation mathématique

On trouvera dans n'importe quel ouvrage traitant des MRI le cadre théorique nécessaire à l'estimation des paramètres du modèle de Rasch ou d'un modèle à plusieurs paramètres. De manière très générale, si l'on se donne Z_i une combinaison

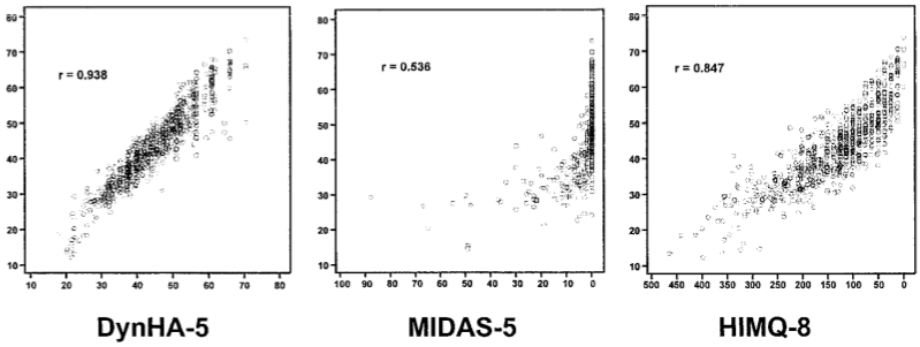


Figure 2 Comparaison entre les estimations de θ sur la base d'un test complet ou après réduction (≤ 5 items). *Tiré de Ware et al. (2000).*

linéaire d'un ou plusieurs paramètres (difficulté, discrimination, « guessing »), et si l'on associe à la difficulté de l'item et l'habileté du candidat les paramètres β et θ, respectivement, alors la probabilité d'observer une réponse correcte à un item est donnée par

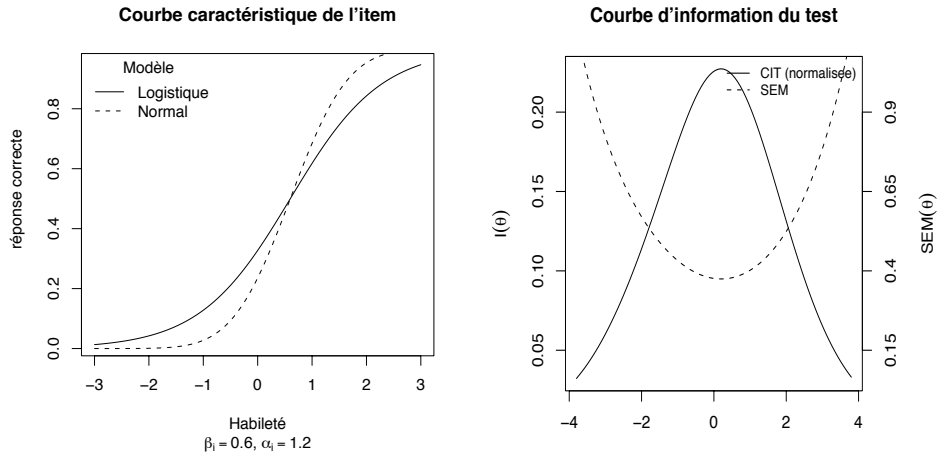
$$P_i(\theta) = P(\alpha_i^\star, \beta_i, \theta) \equiv \eta(Z_i) = \frac{e^{Z_i}}{1 + e^{Z_i}} \, . \tag{1}$$

Si $Z_i = \beta_i - \theta_j$, pour un item i et un individu j, cela revient à considérer que la réponse dépend de l'écart entre le niveau du candidat et la difficulté de l'item (modèle à seuil). Plus un individu aura un niveau élevé, plus la probabilité qu'il réponde correctement à un item de faible difficulté sera élevée. La fonction η dans le modèle **1** correspond simplement à une fonction de lien, dite logistique, classiquement employée dans les MRI. N'importe quelle fonction appartenant à la famille exponentielle pourrait être utilisée. En guise d'exemple, la **Figure 3** (a) illustre une courbe caractéristique de réponse pour un item dont la difficulté est de niveau intermédiaire.

L'estimation de ces paramètres (θ et β) dans un modèle de Rasch peut être réalisée de différentes manières, en maximisant la log-vraisemblance puisque le score total constitue une statistique suffisante. Parmi les méthodes les plus utilisées, citons les méthodes de maximisation marginale et conditionnelle.

Dans l'approche marginale, on considère les effets liés aux individus comme des tirages aléatoires effectués dans une densité de probabilité définie sur la population des individus. Cette densité, notée $g(\theta_p \mid \psi)$, est caractérisée par un vecteur de paramètres de population inconnus, ψ, qui doit être estimé, de même que les paramètres des effets fixes β_i. La vraisemblance à maximiser s'exprime sous la forme :

$$\ell(\beta, \psi) = \prod_{p=1}^{P} \int_{-\infty}^{+\infty} \prod_{i=1}^{I} \Pr(Y_{pi} = y_{pi} | \theta_p) g(\theta_p \mid \psi) d\theta_p \, . \tag{2}$$



a b

Figure 3 (a) Courbe de réponse à un item de difficulté $\beta = 0.6$. (b) Courbe d'information d'un test de 55 items et erreur standard de mesure associée.

Si la densité est discrète, l'intégrale doit être remplacée par une somme. En fonction des hypothèses que l'on s'autorise à faire sur la densité de probabilité théorique des effets aléatoires, on distingue trois approches d'estimation : l'approche non-paramétrique, l'approche semi-paramétrique et l'approche paramétrique. Dans le cas le plus général, l'EMV *non-paramétrique* ou entièrement semi-paramétrique ne suppose aucune hypothèse sur $g(\theta_p \mid \psi)$ – celle-ci n'est tout simplement pas spécifiée. Dans ce cas, il a été montré que l'estimée de la fonction de distribution $G(\theta_p \mid \psi)$ est une fonction en escalier avec un nombre fini de pas. Dans la méthode d'*estimation semi-paramétrique*, la position des pas est supposée connue mais les masses de probabilité en ces points définis doivent être estimés. Dans la méthode d'*estimation paramétrique*, la densité de probabilité $g(\theta_p \mid \psi)$ est choisie comme étant une densité paramétrique dont les paramètres sont à estimer. Dans la plupart des modèles, on supposera une gaussienne centrée, i.e. $g(\theta_p \mid \psi) \sim \mathcal{N}(0; \sigma^2)$ (σ inconnu).

Dans l'approche conditionnelle (applicable au modèle de Rasch), on dérive comme statistique suffisante pour l'effet spécifique de l'individu (θ_p) le score total $s_p = \sum_{i=1}^{I} y_{pi}$. Après conditionnement, la probabilité d'observer un certain profil de réponse ne dépend pas de l'effet lié à l'individu, mais seulement de cette statisque suffisante. Par conséquent, l'effet spécifique lié à l'individu disparaît de la vraisemblance dite *conditionnelle* :

$$\ell(\beta) = \prod_{p=1}^{P} \Pr(Y_{p1} = y_{p1}, \ldots, Y_{pI} = y_{pI} \mid s_p) \; .$$

La vraisemblance conditionnelle est maximisée par rapport à β.

En raison de contraintes d'identifiabilité dans ce type de modèle dans le cadre de l'estimation par marginalisation, la distribution *a priori* des habiletés est supposée être une gaussienne centrée-réduite $\mathcal{N}(0;1)$. La mise à jour de l'algorithme itératif pour l'estimation fournit en sortie une estimation a posteriori de cette densité. Les procédures d'estimation sont décrites dans Baker & Kim (2004), mais de manière générale on retiendra que c'est l'algorithme EM (*Expectation-Maximization*) qui est le plus souvent utilisé (e.g. Johnson, 2007). Il s'agit d'une *méthode indirecte* de maximisation de la vraisemblance (telle qu'exprimée dans l'**équation 2**, par exemple) : l'optimisation de la (log)vraisemblance est transférée à une autre fonction pour laquelle on peut montrer que sa maximisation conduit à une augmentation de la vraisemblance marginale initiale.

Dans l'algorithme EM, l'ensemble des effets aléatoires de tous les individus $\theta = (\theta_1, \ldots, \theta_P)$ sont considérés comme des données manquantes et, avec les données observées $\mathbf{y} = (y_1', | \ldots, y_P')'$, ils forment les données complètes. Les effets aléatoirses sont manquants et donc ne sont pas observés, de sorte qu'à chaque étape de l'algorithme, on commence par calculer la valeur attendue de la vraisemblance des données complètes, étant données les valeurs observées et les estimations des effets fixes $\beta^{\,old}$ et $\sigma_\theta^{2\;old}$ obtenues à l'étape précédente, et des données observées. Il s'agit de l'*étape E*. Ensuite, la log-vraisemblance attendue est maximisée : c'est l'*étape M*. Chaque itération de l'algortihme EM consiste donc en une étape E, suivie d'une étape M, et ce cyle se poursuit jusqu'à la convergence.

L'espérance de la log-vraisemblance des données complètes, $\ell_C(\beta, \sigma_\theta^2)$, que l'on notera $\gamma = E\left(\ell_C(\beta, \sigma_\theta^2) \mid y, \sigma_\theta^{2\;old}\right)$, est définie comme :

$$
\begin{aligned}
\gamma &= E\left(\log \prod_{p=1}^{P} \big(\Pr(y_p \mid \beta, \theta_p)\phi(\theta_p \mid 0, \sigma_\theta^2)\big) \mid y, \sigma_\theta^{2\;old}, \beta^{\;old}\right) \\
&= \sum_{p=1}^{P} E\left(\log\big(\Pr(y_p \mid \beta, \theta_p)\phi(\theta_p \mid 0, \sigma_\theta^2)\big) \mid y, \sigma_\theta^{2\;old}, \beta^{\;old}\right) \\
&= \sum_{p=1}^{P} \int \Big(\log\big(\Pr(y_p \mid \beta, \theta_p)\big) + \log\big(\phi(\theta_p \mid 0, \sigma_\theta^2)\big) h(\theta_p \mid y, \sigma_\theta^{2\;old}, \beta^{\;old})\Big)\, d\theta_p \;,
\end{aligned}
\tag{3}
$$

où $h(\theta_p \mid y, \sigma_\theta^{2\;old}, \beta^{\;old})$ est la densité conditionnelle des effets aléatoires connaissant les données observées, les estimations actualisées des paramètres fixes et la variance de la distribution des effets aléatoires. Après avoir calculé la log-vraisemblance attendue avec les données complètes (étape E), celle-ci est maximisée par rapport à β et σ_θ^2 (étape E).

Notons que l'intégrale impropre n'a pas disparu de la log-vraisemblance attendue des données complètes (**Equation 3**). Ainsi, l'intégrale doit-elle être approchée par une technique d'intégration gaussienne ou de type Monte Carlo.

Pourquoi utiliser l'algorithme EM dans ce cas ? Celui-ci offre trois avantages. Premièrement, cet algorithme garantit qu'à chaque itération la log-vraisemblance marginale augmente, bien que l'algorithme ne la maximise pas directement. Cela rend l'algorithme numériquement très stable. Ce n'est pas garanti lorsque l'intégrale n'est qu'une approximation. Deuxièmement, la log-vraisemblance attendue dans l'**équation 3** est écrite sous la forme d'une somme d'une composante décrivant les paramètres des effets fixes et d'une composante décrivant le paramètre de variance. Cela signifie que l'estimation de ces deux ensembles de paramètres peut être effectuée séparemment durant l'étape M, ce qui réduit la dimension du problème d'optimisation. En dernier lieu, l'étape M dans l'algorithme EM donne des solutions admissibles pour certains paramètres. Pour les composantes de variance sous une hypothèse de normalité, une telle solution existe. Pour les paramètres qui ne possèdent pas de solutions admissibles, il est nécessaire de recourir lors de l'étape M à une méthode d'optimisation itérative, par exemple la méthode de Newton-Raphson.

Un désavantage de l'algorithme EM est que la convergence vers le maximum n'est généralement pas très rapide, en particulier au voisinage du maximum de la vraisemblance marginale. Il existe des variantes de l'algorithme EM qui permettent d'accélérer la convergence ou de faciliter le calcul de l'étape de maximisation.

Dans le contexte de la modélisation traditionnelle de réponse à l'item en consiérant seulement les indicatrices des items comme variables prédictrices, l'application de l'algorithme EM présente un autre avantage. Si l'on considère le modèle de Rasch (mais cela reste vrai avec un modèle à deux paramètres), le vecteur des paramètres des items β peut être subdivisé en I sous-ensembles disjoints de paramètres (dans ce cas, des paramètres individuels), $\beta_1, \ldots, \beta_I$, chacun étant associé à un item. Étant donné l'effet aléatoire θ_p, il y a indépendance conditionnelle, et par conséquent, la log-vraisemblance attendue peut s'écrire comme une somme de termes indépendants — un pour chaque item — et chacun peut être maximisé séparemment. Cette propriété permet d'analyser des jeux de données avec un grand nombre d'items (e.g. 50 ou plus), ce qui serait autrement impossible. La même propriété s'applique pour les paramètres des individus dans le modèle. La composante liée aux individus dans le modèle de régression peut être vue comme la moyenne non nulle d'une distribution normale, et de ce fait les coefficients de régression peuvent être estimés séparemment de la difficulté des items. Ceci explique la popularité de l'estimation MML avec EM dans le domaine de la psychométrie.

L'information (statistique) d'un test représente la somme des quantités d'information apportées par chaque item (la propriété d'additivité résulte de l'hypothèse d'indépendance locale). Elle se calcule comme

$$I(\theta) = \sum_{i=1}^{k} I_i(\theta; b_i) \ , \tag{4}$$

la difficulté de l'item i étant présumée fixée à une certaine valeur b_i.

On montre que la précision de la mesure (estimation de θ) est inversement proportionnelle à la quantité d'information apportée par l'item. Pour un item donné, on a $I_i = \frac{\{P_i'(\theta)\}^2}{P_i(\theta)(1-P_i(\theta))}$ et l'erreur standard de l'estimé de $\hat{\theta}$ est donnée par $\left(\sqrt{\sum_i I_i(\theta)}\right)^{-1}$. On peut en voir une illustration dans la **Figure 3** (b). L'approche informationnelle est centrale au testing adaptatif dans la mesure où c'est le seul critère qui permet d'optimiser à la fois la difficulté globale du test, au travers des items administrés, tout en contrôlant le degré d'erreur de mesure. Connaissant la difficulté des items contenus dans une banque, il est possible de sélectionner un certain nombre d'entre eux et d'avoir une estimation précise des candidats pour qui ils seront le plus appropriés ; en effet, le maximum d'information correspond au cas $\beta_i - \theta_j = 0$ dans le **modèle 1** (par analogie avec la théorie de l'information, on maximise l'incertitude associée à l'observation d'une réponse lorsque $\beta_i = \theta_j$, puisqu'alors la probabilité d'observer une réponse correcte vaut 0.5).

Élaboration d'un test contraint

L'idée général du TAO est (1) d'évaluer dynamiquement le niveau du candidat (son habileté, notée θ dans la section précédente) en sélectionnant les items les plus appropriés à son niveau (i.e. ceux maximisant l'information), et (2) optimiser à la fois le contenu (couverture du domaine) et la mesure (précision de l'estimation). On part donc d'emblée avec l'idée d'un système de contraintes à appliquer au système de testing, certaines contraintes, en particulier les contraintes psychométriques, pouvant être relaxées selon les objectifs du test. Il est clair que si le test adaptatif vise à une auto-évaluation de la part d'un apprenant, il est préférable d'assurer la couverture du domaine que la précision de la mesure.

L'approche la plus élémentaire dans l'élaboration d'un TAO consiste donc à proposer, de manière séquentielle, des items de difficulté croissante à un candidat. En cas d'échec de la part de ce dernier, on propose un item moins difficile tout en maximisant l'information que l'on peut espérer obtenir par sa nouvelle réponse. Une condition d'arrêt simple dans ce cas est l'atteinte d'une erreur de mesure inférieure à un seuil défini a priori (cf. **Figure 1**). Comme on souhaite généralement assurer un principe d'équité et de standardisation dans l'administration du test, le test ne devra pas excéder un nombre maximal d'essais. On se retrouve donc avec un système de double contraintes (erreur minimale + nombre d'items maximal), d'où un problème fréquemment rencontré avec des candidats « extrêmes » : les candidats de

très fort ou très faible niveau se verront administrer le nombre maximal d'items autorisés alors que l'estimation de leur habileté sera acceptable dès les premiers essais (effets plancher/plafond). Inversement, des candidats pour lesquels on trouve peu d'items correspondant à leur niveau dans la banque auront une erreur d'estimation sur leur score plus importante puisqu'on ne pourra jamais maximiser totalement l'information. En tout état de cause, les principaux problèmes soulevés par ce type d'approche sont

- le choix de l'item de départ : généralement, on utilise les informations dont on dispose sur le niveau du candidat, avant le test, ou alors on choisit un item de niveau faible à intermédiaire pour initialiser la séquence d'items ;
- la longueur du test, si elle n'est pas fixe, peut engendrer des problèmes de comparabilité entre les tests (e.g. un candidat répond à 20 items tandis qu'un autre candidat répond à 60 items) ;
- l'erreur de mesure n'est pas nécessairement le critère le plus pertinent pour décider de l'arrêt du test ;
- la nécessité d'une banque d'items, dont la taille va dépendre des facteurs précédents mais aussi du public auquel s'adresse le test (s'agit-il d'un public tout-venant, représentatif de tous les niveaux d'habileté ?), et pour laquelle il faudra développer un système de contrôle et de suivi (exposition des items, sécurisation, etc.).

On voit d'emblée que ce type d'approche présente le désavantage de reposer sur une banque d'items bien alimentée. Par exemple, pour le TCF, qui est organisé autour de trois compétences, on peut imaginer partir sur le base de 300 à 500 items, recouvrant les différents niveaux du Cadre européen. D'autre part, il est difficile d'assurer une sélection optimale d'items répondant à différentes thématiques (contrainte de contenu) ou ayant des formats différents.

Les solutions à cette approche « gloutonne » consistent par exemple à partitionner la banque d'items selon différents attributs et à sélectionner un ensemble d'items représentatifs. Des algorithmes de type `minimax` permettent généralement de répondre à ce type d'approche. Une autre solution possible consiste à opter pour une approche de programmation linéaire pure, par exemple la méthode de déviation pondérée (e.g. Stocking & Swanson, 1993), dans laquelle les spécifications du test sont exprimées sous forme de contraintes numériques (sélection séquentielle des items par minimisation d'une fonction d'utilité). Ces solutions sont coûteuses en développement et généralement la phase d'optimisation est délicate et conduit souvent à travailler sur un espace de solutions admissibles trop limité.

Si l'on se détourne de ces approches de sélection item par item, on trouve des solutions tout à fait acceptables qui consistent à proposer aux candidats des sous-ensembles d'items. Parmi ces solutions figurent

- les tests reposant sur des blocs d'items, ou *testlets* : à la place d'unités discrètes, on travaille avec des paquets d'items partageant des spécifications communes (en termes de contenu) ;

- les tests hiérarchisés, ou *multi-stage* : on utilise des « mini-tests » cohérents, ordonnés par dificulté (globale) croissante ;

- les tests virtuels, ou *shadow tests* : il s'agit d'une procédure itérative dans laquelle on élabore un test complet mais on ne propose qu'un seul item au candidat ; en fonction de la réponse obtenue, on élabore un nouveau test dont le niveau global de difficulté est adapté au niveau estimé du candidat, et on réitère jusqu'à satisfaire des règles d'arrêt définies à l'avance (erreur de mesure ou longueur du test, par exemple).

Ces procédures sont largement détaillées dans van der Linden & Glas (2000), et elles font généralement appel à des techniques de programmation linéaire. Dans le cas des *tests virtuels*, par exemple, on utilise des techniques de programmation linéaire binaire, et le schéma général de l'organisation du test est alors le suivant :

1. Initialisation de l'estimateur
2. Construction d'un test virtuel remplissant les contraintes et possédant l'information maximale
3. Administration de l'item de ce test possédant l'information maximale au voisinage de l'habileté estimée
4. Mise à jour de l'habileté estimée
5. Retour des items dans la banque
6. Ajustement des contraintes pour tenir compte de l'item utilisé
7. *Itération des étapes 2 à 6*

La fonction d'utilité est de la forme

$$\max_i \sum_{i \in O_k} I_i(\hat{\theta}_{k-1}) x_i \ ,$$

où O_k désigne l'ensemble des items support. On reconnaît bien l'idée générale développée jusqu'à présent : maximiser l'information apportée en utilisant l'item i, connaissant la réponse fournie à cette item, x_i, en fonction de l'habileté estimée à l'étape précédente ($\hat{\theta}_{k-1}$).

L'idée de construire des mini-tests, équivalents en terme de difficulté globale et homogénéisés du point de vue de leur contenu, apparaît donc une solution plus économique et moins coûteuse en développement. La construction de ce type de formes parallèles de test présente l'avantage de répondre à un souci d'optimisation dans l'utilisation d'un nombre limité d'items (sélection manuelle des items constituant les tests, contrôle du niveau d'exposition de chaque item, harmonisation du format des items et d'un contenu générique pour le mini-test). Ceci se prête bien à des banques d'items en cours de développement et permet d'insérer des items nouvellement créés afin de les évaluer auprès des candidats testés (« live pretesting »).

Le choix du nombre d'items (longueur du test) et plus généralement des règles d'arrêt apparaît également un élément décisif de la qualité des TAO. Généralement, on les choisira sur la base de simulations réalisées à partir des données contenues dans la banque d'items – on parle alors de simulations sur données réelles – ou à partir de données fictives. On simule par exemple le fonctionnement du système de test avec les réponses collectées sur des candidats à qui les items ont été préalablement administrés (en ignorant les autres réponses fournies par ces candidats). Dans le cas où l'on travaille avec plusieurs versions de test, cette approche n'est plus possible et il faut utiliser directement les paramètres d'items obtenus à l'aide d'un MRI : on simule alors les réponses qui auraient été fournies par un échantillon aléatoire représentatif de la population testée. Par exemple, on peut utiliser un ensemble d'individus virtuels, dont l'habileté est distribuée de manière uniforme sur l'échelle de mesure (e.g. de -3.5 à $+3.5$ logit, par pas de 0.25). En choisissant $n = 100$ individu par intervalle, on a au total 2900 répondants virtuels. À partir de là, il est possible d'estimer le nombre d'items nécessaires afin d'obtenir une erreur de mesure bornée sur un certain intervalle (analogue à un calcul de puissance statistique). C'est la technique utilisée par Fliege et al. (2005) pour valider un test adaptatif dans le domaine biomédical. Dans le domaine de l'éducation, on pourra consulter les rapports élaborés par ETS pour valider leurs tests GRE ou TOEFL, e.g. Schaeffer et al. (1998).

Implémentation informatique

Sur le plan informatique, les solutions évoquées dans les sections précédentes présupposent :

- l'existence d'une banque d'items et un moyen d'accès à celle-ci, de manière locale ou distante ;
- la possibilité de pré-assembler des versions de test ;
- un moteur pour l'estimation des paramètres statistiques (information, habileté, erreur de mesure, etc.) ;

- un gestionnaire de test (sélection des items, règles d'arrêt, etc.) ;
- une interface pour la présentation des items (affichage écran, saisie utilisateur, édition des résultats).

On trouvera sur le site de Lawrence M. Rudner un tutoriel ainsi qu'une petite **démonstration** en Javascript. Bien que « simpliste », cette application permet de se faire une idée des attendus en termes calculatoires. Les algorithmes plus évolués sont détaillés dans van der Linden & Glas (2000) ou Wainer et al. (1990). D'autres solutions informatiques sont bien entendu disponibles sur le marché. Par exemple, FastTest est un logiciel (payant) distribué par Assessment System Corporation qui offre une solution de testing intégrale incluant une banque d'item et un moteur de développement de test. Il est disponible sur le **site de** ASC.

Conclusion

On a vu que les tests adaptatifs présentent de nombreux avantages, tant du point de vue du concepteur de test car il peut bénéficier en retour des réponses des candidats pour améliorer sa banque d'items, que du point de vue de l'utilisateur final qui bénéficie de conditions de test optimisées (en un certain sens). Ils souffrent également de contraintes méthodologiques (programmation sous contraintes, approche dynamique, calcul numérique) et théoriques (par exemple, la difficulté pour garantir l'absence de fonctionnement différentiel des items) particulièrement délicates.

Les approches actuelles du TAO tendent à privilégier

- la maximisation de l'information apportée par une mesure ponctuelle (réponse à un item), en fonction des réponses antérieures (*contrainte de mesure*),
- une procédure itérative de sélection des items ou de blocs d'items, en respectant la couverture du domaine à évaluer (*contrainte de contenu*),
- la validité du test et l'équité des scores délivrés aux candidats.

Références

1. Baker, F.B. and Kim, S.-H. (2004). *Item Response Theory : Parameter Estimation Techniques* (2nd Ed.). Dekker.
2. Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In Lord, F.M. & Novick, M.R. (Eds.), *Statistical Theories of Mental Test Scores*. Reading, MA : Addison-Wesley.
3. Devouche, E. (2003). Les banques d'items. Construction d'une banque pour le Test de Connaissance du Français. In P. Dickes and A. Flieller (Eds.), Mesure et Education, numéro spécial de *Psychologie et Psychométrie*, 24, 57–88. **On-line** PDF **version**
4. Fliege, H., Becker, J., Walter, O.B., Bjorner, J.B. et *al.* (2005). Development of a computer-adaptive test for depression (DCAT). *Quality of Life Research*, 14, 2277–

2291.

5. Johnson, M.S. (2007). Marginal Maximum Likelihood estimation of item response models in R. *Journal of Statistical Software*, 20(10), 24 pp. **On-line PDF version**
6. van der Linden, W.J. and Glas, C.A.W. (2000). *Computerized Adapative Testing. Theory and Practice*. Kulwer.
7. van der Linden, W.J. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2), 201–216.
8. Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. Chicago : The University of Chicago Press.
9. Schaeffer, G.A., Bridgeman, B., Golub-Smith, M.L., Lewis, C., Potenza, M.T., and Steffen, M. (1998). *Comparability of Paper-and-Pencil and Computer Adaptive Test Scores on the GRE General Test*. ETS Report RR-98-38. **On-line PDF version**
10. Stocking, M.L. and Swanson, L. (1993). A method for severely constrained item selection in adaptive testing. *Applied Psychological Measurement*, 17(3), 277–292.
11. Swanson, L. and Stocking, M.L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement*, 17(2), 151–166.
12. Wainer, H., Dorans, N.J., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., and Thissen, D. (1990). *Computerized Adaptive Testing : A Primer*. Hillsdale NJ : Erlbaum.
13. Ware, J.E. Jr., Bjorner, J.B., and Kosinski, M. (2000). Practical Implications of Item Response Theory and Computerized Adaptive Testing. *Medical Care*, 38(9), Supp. II, II-73–II-82.

Christophe Lalanne, christophe.lalanne@gmx.net

$Id : intro.tex,v 1.0 2008/09/11 chl$