Création d'un score global dans le cadre d'une épreuve adaptative

Fabrice Murat* et Thierry Rocher**

La construction d'un indicateur global de compétence à partir de réponses à des exercices fait appel à des techniques statistiques spécifiques, plus complexes que le simple comptage des bonnes réponses. Le recours à ces techniques est particulièrement utile quand tous les individus évalués n'ont pas passé les mêmes épreuves.

C'est le cas dans l'enquête *Information et Vie Quotidienne*, où les exercices sont proposés dans un cadre adaptatif, pour améliorer la motivation des personnes interrogées, surtout de celles qui sont sorties depuis fort longtemps de l'école. En fonction des résultats au premier exercice, l'enquêté se voyait proposer des questions plus ou moins difficiles. Si cette procédure améliore sensiblement les conditions de collecte et la qualité des données recueillies, elle engendre des difficultés particulières lors de l'estimation d'un indicateur de compétence global, valable pour toute la population. En effet, comment comparer les résultats de populations n'ayant pas passé les mêmes épreuves ?

Cet article propose différentes méthodes pour tenir compte de ce processus d'orientation. Elles sont testées sur des données fictives, pour en étudier la robustesse et les limites. Si la perte d'information induite par le processus d'orientation est assez minime, elle n'est cependant pas nulle, en particulier pour un nombre important de personnes se situant autour des seuils d'orientation vers les exercices difficiles ou vers les exercices faciles. Ces résultats suggèrent quelques aménagements dans la structure des épreuves pour la prochaine édition de l'enquête.

** Au moment de la rédaction de cet article, Fabrice Murat travaillait à la division Emploi de l'Insee.*
*** Thierry Rocher travaille à la Direction de l'évaluation, de la prospective et de la performance, du ministère de l'Éducation nationale*

L'enquête *Information et Vie Quotidienne (IVQ)* cherche à évaluer les compétences face à l'écrit d'une population d'adultes, quels que soient leur rapport à l'écrit, leur maîtrise de la langue française, leur niveau d'éducation, etc. Les compétences visées se situent sur un assez large spectre, du décodage de mots simples à la compréhension de l'implicite d'un texte. Le caractère très hétérogène de la population évaluée a ainsi motivé le recours à un questionnement en deux temps, avec un processus d'orientation. En fonction de ses résultats au premier exercice, l'enquêté se voit proposer des questions plus ou moins difficiles. Cette adaptation est apparue indispensable lors des tests sur le terrain : des questions trop difficiles découragent les individus les moins compétents ; des exercices trop simples démotivent les meilleurs ou les incitent à chercher des pièges (Vallet et al., 2002). Ce choix améliore la qualité de la mesure, mais il complique la construction d'un score global, valable pour tous les individus.

De manière plus générale, ce protocole s'inscrit dans une démarche de *test adaptatif*. Dès la naissance des premiers tests psychologiques, au début du XX^{e} siècle, il est apparu intéressant d'adapter le niveau de difficulté des épreuves au niveau de compétences des individus. Le principe est le suivant : on présente à chaque individu un exercice ; s'il échoue, on lui présente un exercice plus facile ; s'il réussit, on lui présente un exercice plus difficile. Ce processus itératif conduit à une estimation plus précise - et plus rapide - du niveau de compétence de chaque individu. Avec le développement de l'informatique, cette procédure s'est répandue (Wainer, 2000). À chaque item, suivant la réponse de l'individu, son niveau de compétence est réestimé et l'ordinateur propose un nouvel item dont la difficulté correspond à ce niveau. Il est également possible de proposer plusieurs items, notamment en compréhension de l'écrit, où un même texte fait généralement l'objet de plusieurs questions. La contrainte principale de ce type de procédure est qu'il est nécessaire d'avoir estimé au préalable la difficulté d'un grand nombre d'items. Cela suppose que chaque item ait été passé par un échantillon représentatif de la population visée, que sa difficulté ait été estimée et enregistrée dans une *banque d'items,* parmi lesquels il sera possible de choisir le plus approprié lors de la procédure de test adaptatif. La constitution d'une telle banque implique un coût financier très important, qui limite la mise en pratique des tests adaptatifs (1).

Il existe d'autres stratégies d'adaptation, moins exigeantes. C'est le cas par exemple de la procédure en deux temps avec un test d'orientation (*two-stage testing*) adoptée dans *IVQ*. L'adaptation des items n'est pas faite individuellement mais pour des groupes d'individus déterminés en fonction de leurs résultats à un test d'orientation. Cette procédure est moins contraignante en pratique. Le recours à l'ordinateur n'est pas requis. Elle a l'avantage de pouvoir être appliquée pour une passation collective de tests papier-crayon, comme par exemple les tests de la Journée d'Appel de Préparation à la Défense (Rocher, 2004). Elle ne nécessite pas d'estimer au préalable la difficulté des items et donne potentiellement des résultats plus précis que ceux obtenus par un seul test, dans le cas où les niveaux de compétence sont très dispersés (Lord, 1980).

Au-delà des aspects pratiques, cette procédure se justifie également sur le plan théorique. Les dimensions cognitives fines que l'on souhaite évaluer ne sont pas forcément les mêmes selon les niveaux de compétences. Pour les personnes en difficulté face à l'écrit, il convient d'insister sur le décodage des mots par exemple (permettant d'étudier la maîtrise des mécanismes de base de l'écriture), alors que pour les autres personnes, différents aspects de la compréhension pourront être plus finement évalués. Ainsi, ce n'est pas seulement la difficulté du test qui est adaptée, mais la nature même de ce qu'il est censé mesurer.

Mais cette perspective remet en cause la démarche consistant à établir une échelle commune sur laquelle seront placés tous les individus. Quelle est la validité d'un score global si les éléments du test renvoient à des dimensions différentes ? Cela suppose que les individus peuvent être classés sur un *continuum* unidimensionnel, en fonction de leurs réponses aux items, qui contribuent chacun à la mesure d'une même dimension cognitive (2). En l'occurrence,

---

1. Autre difficulté, il faut aussi que la réponse de l'individu soit corrigée immédiatement, ce qui rend difficile le recours à un codage manuel et impose une procédure d'estimation des compétences intégrée à l'outil de collecte, ce qui peut poser problème.

2. Postuler « l'undimensionnalité » d'un ensemble de données revient à supposer qu'elles peuvent être « engendrées » par une seule variable, selon un modèle statistique déterminé. Le caractère unidimensionnel ou multidimensionnel des tests psychologiques cherchant à mesurer l'intelligence est une question centrale de la psychométrie, ayant fait l'objet d'une des premières grandes controverses de ce domaine entre Spearman et Thurstone : les mêmes données analysées par ces deux chercheurs ont pu apparaître grossièrement structurée autour d'une dimension dominante ou au contraire relever de plusieurs facteurs. La question de la « réalité » psychique ou physique de ces facteurs est un autre grand sujet de débat de la psychologie cognitive (Gould, 1987).

l'hypothèse envisagée ici est que tous les items d'*IVQ* portant sur la compréhension de l'écrit mesurent une même dimension, que ce soient ceux destinés aux personnes en difficulté ou ceux qui s'adressent aux « bons lecteurs ». L'unidimensionnalité est envisagée ici comme la présence d'une dimension dominante (Blais et Laurier, 1997). S'il existe un cadre formel pour tester cette hypothèse (Stout, 1990), il est impossible de l'appliquer à *IVQ*, dans la mesure où les individus, selon leur niveau, ne passent pas les mêmes items.

Le score global aux épreuves de compréhension de l'écrit de l'enquête *IVQ* est donc ici perçu comme un indicateur synthétique des compétences des individus face à l'écrit. Ce score présente certainement assez peu d'intérêt pour le psychologue, qui préférera procéder à une analyse plus fine des réponses aux items (Megherbi *et al.*, ce numéro), mais il a l'avantage de pouvoir être plus facilement confronté aux caractéristiques des individus, dans une perspective d'analyse économique ou sociologique. Dans ce cadre, le problème posé par la construction d'un tel score est de nature statistique.

Comment tenir compte alors du fait que tous les individus, selon leurs résultats, ne passent pas les mêmes exercices ? La dépendance entre le processus d'orientation et le niveau de compétence de la personne, estimé approximativement par l'exercice d'orientation, rend assez délicate l'estimation de ce niveau. Différentes techniques sont possibles pour synthétiser l'ensemble des réponses aux exercices. Généralement, cet ensemble peut être représenté comme une matrice de réponses, appelée aussi matrice de Stern, souvent réduite à une distinction entre les « bonnes » réponses, les « mauvaises » réponses et les absences de réponses (3). On peut distinguer très grossièrement trois familles principales de techniques (Bernier, Pietrulewicz, 1997 et Dickes *et al.*, 1994) :

- *l'analyse classique* : elle consiste à simplement considérer le nombre de bonnes réponses comme indicateur de compétence du sujet ou comme indicateur de difficulté d'un item. Pour calculer les scores individuels, on peut éventuellement avoir recours à une pondération des différents items, par exemple en fonction de leur difficulté. Cette analyse est souvent complétée par celle de la corrélation entre chaque item et le score global, comme mesure de sa « qualité ».

- *l'analyse factorielle* : l'analyse factorielle a, on le sait, été développée par Spearman pour analyser les réponses à des tests d'intelligence. Cette technique est encore assez largement utilisée pour explorer la structure d'un ensemble d'items. En revanche, elle sert moins lors de la phase de construction proprement dite des scores.

- *les modèles de réponse à l'item* : de plus en plus diffusés, ces modèles logistiques posent de façon plus claire que dans l'analyse classique, le caractère latent de la compétence. Ils cherchent à paramétrer de façon indépendante la compétence des individus et la difficulté des items. On peut ainsi comparer le fonctionnement de la même épreuve sur deux populations différentes ; il est aussi possible d'ancrer assez facilement deux épreuves différentes l'une sur l'autre, pour peu qu'elles aient un minimum d'items en commun.

L'objectif de cet article est d'appliquer ces techniques sur les données d'*IVQ*, afin d'aboutir à la construction d'un score global en compréhension de l'écrit, valable pour toute la population. La principale difficulté rencontrée tient à l'existence de « trous » dans la matrice des réponses, ces lacunes n'étant pas aléatoires : on ignore les réponses que les personnes orientées vers les exercices simples auraient données sur les exercices complexes. Supposer un échec complet apparaît vite comme une hypothèse trop forte. On proposera donc plusieurs solutions à ce problème, en incluant l'usage des modèles de réponse à l'item. L'analyse des différents scores obtenus portera sur leur distribution ou leur corrélation avec les caractéristiques des individus. Cette confrontation est à la fois une validation (une corrélation forte avec le diplôme est attendue, par exemple) et une illustration de la sensibilité des résultats aux hypothèses retenues lors de la construction des indicateurs. Enfin on présentera une simulation de type Monte Carlo, sur données fictives, proches de celles d'*IVQ*, pour lesquelles l'ensemble des données sera disponible. Des « trous » seront « creusés » dans ces données de façon similaire à la procédure d'*IVQ* et on comparera les résultats obtenus suivant les différentes techniques, avec les scores tenant compte de l'ensemble de l'information initiale. Ces simulations donneront une idée des conséquences du caractère adaptatif des épreuves sur

3. On ne s'intéresse bien sûr ici qu'aux procédures d'évaluation dans des enquêtes statistiques. Dans le cadre scolaire habituel, les évaluations faites par les professeurs, les notes, ne sont pas aussi facilement décomposables en processus élémentaires (pensons à la correction d'une dissertation, notamment). De plus, comme l'a montré Merle (1996), les notes n'évaluent pas toujours uniquement le résultat obtenu, mais parfois aussi les progrès accomplis ou les efforts fournis.

la mesure finale. Elles permettront aussi d'envisager et de tester des variantes pour l'enquête de 2010.

Le processus d'orientation dans *IVQ*

Cet article va se centrer sur l'évaluation des compétences en compréhension de texte, qui fait l'objet du plus grand nombre d'exercices dans l'enquête (cf. encadré 1). L'orientation se fait en deux étapes. Le module d'orientation construit une première image des compétences de la personne, qui conduit à répartir la population en trois groupes de compétences. Les deux groupes extrêmes sont orientés directement vers des exercices d'une difficulté adaptée. Pour le groupe Intermédiaire, une deuxième étape d'orientation est nécessaire pour savoir si ce sont les exercices complexes ou les exercices simples qui sont préférables. Ainsi, quatre parcours différents sont possibles (cf. tableau 1) (4) :

- Groupe « ANLCI direct » : ces personnes obtiennent de faibles performances à l'exercice d'orientation. Elles passent directement le module ANLCI (la partie compréhension se trouve alors à la fin, après les exercices d'écriture et de lecture de mots écrits) ;

4. Afin de travailler sur des données parfaitement fiables, nous n'avons pas retenu dans cette analyse une minorité de sujets qui ne semblent pas avoir répondu avec assez d'implication : il s'agit des personnes qui avaient un score nul à l'orientation (167 personnes) ou un score nul au module ANLCI (37 personnes) : en effet, il s'agit généralement d'une suite de « ne sait pas » qui indiquent sans doute davantage du désintérêt et du découragement qu'une absence complète de compétences. De même, nous avons aussi écarté les individus qui n'ont passé aucun exercice en invoquant de gros problèmes en français ou à l'écrit (171 personnes).

Encadré 1

L'ENQUÊTE *IVQ* ET L'ÉVALUATION DES COMPÉTENCES À L'ÉCRIT

L'enquête *IVQ* (*Information et Vie Quotidienne*) a été réalisée fin 2004 et début 2005, dans 10 284 ménages de France métropolitaine. Dans chacun de ces ménages, une personne de 18 à 65 ans a été tirée au sort pour passer des exercices d'évaluation à l'écrit, en compréhension orale et en calcul et pour répondre à un questionnaire biographique (voir l'article de présentation de ce numéro pour plus de précisions). Les compétences à l'écrit se subdivisent en trois domaines : lecture de mots, compréhension de textes et écriture de mots. L'architecture de l'évaluation dans ces domaines est la suivante :

- La personne interrogée passe d'abord un exercice d'orientation, assez simple, comportant des questions en lecture de mots et en compréhension de textes écrits (sur un texte court). Des scores dans chacun de ces domaines sont calculés : SL en lecture de mots et SC en compréhension. Ces scores sont le nombre de bonnes réponses aux deux exercices. Les questions étant pondérées de 1 à 3 selon leur difficulté, SL a un maximum de 15 points et SC un maximum de 19 points.

- Le processus d'orientation distingue alors trois cas (cf. schéma *infra*) :

  - Les personnes ayant eu de bons résultats en lecture de mots **et** en compréhension à l'exercice d'orientation (soit (SL >11 et SC >16)) passent un « module haut » avec des exercices plus complexes en compréhension de textes.

  - Celles qui ont eu des performances nettement insuffisantes dans l'un de ces deux domaines (soit SL < 11 ou SC < 11) passent le « module ANLCI », qui affine la mesure en lecture de mots, en compréhension de texte écrit (sur un texte court) et en écriture de mots (il s'agit d'écrire une liste de courses).

  - Les personnes aux résultats moyens à l'exercice d'orientation (soit (SL >10 et SC >10) et (SL < 12 ou SC < 17)) passent un « module intermédiaire ». Il s'agit en fait de l'exercice de compréhension du module ANLCI. À partir des 11 questions, en utilisant comme dans le module d'orientation une pondération selon la difficulté, ce module donne lieu au calcul d'un score sur 24 points. Si la personne obtient au moins 19 points, elle passe le module haut. Sinon, elle passe le reste du module ANLCI.

Du fait d'une erreur informatique, le processus d'orientation a été un peu perturbé et quelques individus destinés à passer le module haut ont passé le module ANLCI ou le module intermédiaire.

Trois types de compétences face à l'écrit sont évalués dans *IVQ* : la lecture de mots, l'écriture de mots, la compréhension de textes écrits. Ces compétences peuvent être considérées comme distinctes et nécessitent la construction de trois indicateurs différents. Cependant, tous les individus ne passent pas des exercices d'égales longueurs dans ces trois domaines. L'écriture de mots n'est évaluée que pour les personnes repérées en difficulté lors de l'exercice d'orientation ou lors de l'exercice intermédiaire. Il n'est donc pas possible pour celles orientées vers le module haut, la majorité de la population, d'avoir une idée de leurs compétences sur ce point. Tous les individus passent des questions de lecture de mots dans l'exercice d'orientation, mais s'il s'en trouve d'autres dans le module ANLCI, le module haut, lui, n'en comporte pas. De plus, cette compétence paraît maîtrisée par presque tout le monde, même parmi les personnes orientées vers le module ANLCI. Reste le domaine de la compréhension de textes écrits, qui est mieux représenté dans les différentes épreuves :

➔

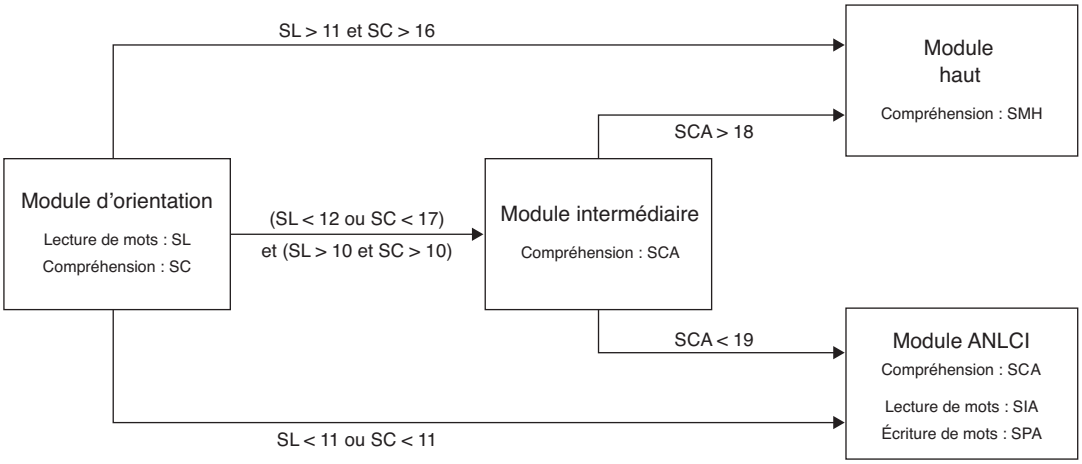
- Groupes « Intermédiaires » : les personnes obtiennent des résultats moyens à l'exercice d'orientation. On leur propose alors tout de suite la partie compréhension du module ANLCI :

- Groupe « Intermédiaire ANLCI » : si les personnes obtiennent des résultats insuffisants, elles passent alors le reste du module ANLCI (elles répondent donc aux mêmes questions que les personnes relevant du cas 1, mais pas dans le même ordre (5) : on fera parfois référence aux « groupes ANLCI », pour désigner les deux premiers groupes pris ensemble) ;

- Groupe « Intermédiaire Haut » : si les personnes obtiennent de bons résultats, elles passent le module Haut ;

5. L'inversion de l'ordre de passation peut poser quelques problèmes. Ainsi, l'une des premières questions de l'exercice de compréhension utilise comme support un CD de musique et porte sur le nom du chanteur. De nombreuses personnes passant cet exercice comme module intermédiaire donnent comme réponse le nom du groupe. Cette réponse est bien moins fréquente pour les personnes orientées directement vers le module ANLCI. En effet, avant l'exercice de compréhension, elles ont passé l'exercice de lecture de mots, au cours duquel, on leur a demandé de lire le nom du groupe. Celui-ci étant alors pour eux bien repéré, elles ne vont pas être tentées de l'attribuer au chanteur, ce qui élimine donc un « distracteur » et rend la question plus facile.

Encadré 1 (suite)



- le module d'orientation comporte 10 items de compréhension de textes écrits ;

- le module ANLCI en comporte 11 ;

- le module Haut en comporte 21, répartis sur 5 exercices (l'épreuve comportait initialement un item supplémentaire, mais il a été éliminé lors de l'analyse fine des données, car il était de mauvaise qualité).

Le module d'orientation et le module ANLCI ont été élaborés par le PsyEF (laboratoire de psychologie de l'université Lyon II) sur les mêmes fondements théoriques. Portant sur des supports de la vie quotidienne (une page de programme de télévision, un CD de musique), les questions se répartissent selon trois processus cognitifs :

- les questions explicites et littérales (la réponse est telle quelle dans le texte) ;

- les questions explicites et sémantiquement proches (la réponse dans le texte utilise le même champ lexical) ;

- les questions implicites (demandant le recoupement de différentes informations données par le texte).

Le module Haut a été construit par une équipe de l'université Paris V. Il porte sur des supports plus nombreux et plus complexes : des graphiques statistiques, une carte routière, un texte long, etc. Les processus évalués sont aussi globalement plus élaborés :

- la compréhension des informations de surface se rapproche des questions explicites dans le modèle précédent : il s'agit de prélever de l'information ponctuelle dans le texte ;

- la compréhension sous forme propositionnelle demande une conceptualisation plus grande du contenu du texte pour répondre aux questions ;

- les questions sur le modèle de situation exigent à la fois la compréhension du texte et sa confrontation avec les connaissances préalables.

Les « questions implicites » du module d'orientation et ANLCI semblent relever à la fois de ces deux derniers processus. De façon assez logique, il semble bien que le module ANLCI cible les processus les plus simples, tandis que le module Haut s'intéresse à des processus plus complexes. Une analyse plus approfondie devra cependant être faite lors de la mise en place de la nouvelle enquête. Par ailleurs, le mode de passation est différent selon les modules : pour le module d'orientation et le module ANLCI, l'enquêté a le texte en main et c'est l'enquêteur qui lit les questions ; pour le module haut, s'adressant à des personnes plus à l'aise à l'écrit, c'est l'enquêté qui lit les questions, posées à côté du texte et l'enquêteur ne les répète pas lorsqu'il enregistre les réponses dans l'ordinateur. Dans les deux cas cependant, les réponses sont données oralement et non écrites.

- Groupe « Haut direct » : les personnes ayant eu de bons résultats à l'exercice d'orientation passent directement le module Haut.

Comment placer sur la même échelle les personnes relevant du groupe « Haut direct » et celles relevant des groupes « ANLCI direct » et « Intermédiaire ANLCI », sachant que le groupe « Intermédiaire Haut » est particulier puisque ces personnes passent toutes les questions de compréhension ? L'enquête *IVQ* est fondée sur une hiérarchisation *a priori* de difficultés entre le module Haut et le module ANLCI. L'examen du contenu des épreuves a conforté cette hypothèse (cf. encadré 1), mais ne permet pas de quantifier l'ampleur de l'écart de difficulté entre le module Haut et le module ANLCI. Pour le faire, on peut essayer de tirer partie des « passerelles » existant entre les épreuves : le module Haut et le module ANLCI peuvent ainsi être reliés chacun au module d'orientation. Le module Haut peut être relié directement au module ANLCI, uniquement par le groupe « Intermédiaire Haut ». Cependant, cela revient à faire l'hypothèse que la hiérarchisation des épreuves ou des items observée sur une sous-population est valable pour une autre sous-population. Il faudra autant que possible tester cette hypothèse.

D'un point de vue théorique, on peut déjà dégager l'importance de plusieurs facteurs influant sur la qualité de la synthèse des résultats :

- *la censure des données* : le lien entre les différents scores ne sera généralement observé que sur une partie de la population. Cette sous-population sera généralement définie par un niveau élevé ou au contraire faible sur l'un des scores. Les paramètres nécessaires à la synthèse seront donc estimés en contraignant parfois fortement les variations de l'un des scores. De plus, la forme de la relation (linéaire ou non) sur l'ensemble peut être différente de celle sur la sous-population sélectionnée. La taille de la sous-population, l'amplitude de ses résultats à l'épreuve d'orientation sont deux facteurs qui auront alors une influence sur la qualité des estimations.

- *la qualité des scores d'orientation* : ce phénomène prend toute son importance quand on tient compte de l'erreur de mesure affectant les épreuves, surtout celle concernant l'épreuve d'orientation. Par construction, les individus orientés vers les exercices difficiles, sont un peu moins bons qu'ils ne le paraissent à l'épreuve d'orientation. Certains, aux compétences modestes, ne doivent cette orientation qu'à la faveur du hasard. S'ils repassaient une épreuve de difficulté équivalente à celle d'orientation, ils obtiendraient ainsi, en moyenne, des résultats un peu moins bons. Inversement, le niveau des individus orientés vers les épreuves simples est un peu sous-estimé par l'épreuve d'orientation. Ce phénomène est bien sûr fortement lié à la taille de l'exercice d'orientation. Avec une dizaine de questions, celle d'*IVQ* ne peut, par exemple, prétendre à une mesure parfaite. La difficulté de l'épreuve est aussi un paramètre important, car il indique quelles catégories seront les plus affectées : une épreuve facile sera précise pour les individus les moins compétents, mais l'erreur de mesure sera importante pour les plus compétents.

- *les conditions de passation* : l'ordre de passation des exercices peut avoir son importance. Ainsi dans le cas d'*IVQ*, le groupe « Intermédiaire ANLCI » passe la partie compréhension juste après le module d'orientation et le module oral, alors que le groupe « ANLCI direct » la passe à la fin, après avoir répondu à l'exercice d'écriture de mots et à celui de lecture de mots. Cette dernière sous-population risque d'être un peu plus fatiguée et moins performante. De même, le groupe « Intermédiaire Haut » passe un exercice de plus que le groupe « Haut direct », ce qui peut aussi dégrader sa motivation. Cependant, dans les deux cas, les exercices intercalaires sont assez courts et n'ont probablement pas perturbé durablement l'investissement des personnes interrogées.

Tableau 1
Les épreuves passées par les différents groupes

Groupe	Effectif	Module d'orientation	Partie Compréhension du Module ANLCI	Module Haut
ANLCI direct	1 015		A la fin	
Intermédiaire ANLCI	1 270		Au début	
Intermédiaire Haut	1 058			
Haut direct	6 666			

Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Les scores moyens à l'épreuve d'orientation s'étagent de 5,7 sur 10 pour les personnes orientées directement vers le module ANLCI à 9,3 pour celles orientées directement vers le module Haut

L'analyse statistique descriptive donne déjà quelques indications intéressantes sur le fonctionnement des épreuves (cf. tableau 2). L'*alpha* de Cronbach (6) mesure la fidélité de l'épreuve, en quantifiifant l'ampleur des corrélations entre items, qui est un gage de qualité. Il est pour chaque épreuve assez satisfaisant, compte tenu du faible nombre d'items : il varie de 0,64 pour le module d'orientation à 0,76 pour le module Haut (qui comporte le double d'items, ce qui explique sa meilleure consistance interne).

Sans surprise, les scores moyens à l'épreuve d'orientation s'étagent de 5,7 sur 10 pour les personnes orientées directement vers le module ANLCI après cette épreuve à 9,3 pour celles orientées directement vers le module Haut (7). Pour ces dernières, une proportion aussi élevée d'items réussis provoque un effet de saturation sur cette épreuve, mesurée par l'écart-type, qui n'est que de 0,7 (contre 1,6 pour l'ensemble de la population). De même, sur le module ANLCI, le groupe « Intermédiaire Haut » (réorienté vers le module Haut) se caractérise logiquement par des résultats très élevés (avec une moyenne de 10 sur 11) et une faible dispersion (seulement 0,8 contre 2,6 par exemple pour la population du module ANLCI). Les personnes du groupe « Intermédiaire Haut » obtiennent en moyenne 1,7 point de moins aux exercices du module Haut que les personnes orientées directement vers ce module.

Le module d'orientation et le module ANLCI semblent d'une difficulté assez proche : le taux de réussite est à peu près de 70 % pour l'ensemble des trois premiers groupes (« ANLCI direct », « Intermédiaire ANCLI », « Intermédiaire Haut »), qui ont passé les deux épreuves. En revanche, le module Haut paraît nettement plus difficile : le groupe « Haut direct » a un taux de réussite de 69 % à cette épreuve contre 93 % au module d'orientation. L'écart de réussite du groupe « Intermédiaire Haut » entre les deux modules est à peine moins grand (61 % de réussite au module Haut contre 82 % au module d'orientation). Ces résultats confortent l'hypothèse d'une hiérarchisation d'ensemble entre les modules d'orientation et ANLCI, d'une part, et le module Haut, d'autre part.

*6. L'*alpha *de Cronbach est un indice statistique variant entre 0 et 1 qui permet d'évaluer l'homogénéité (la consistance ou cohérence interne) d'un instrument d'évaluation ou de mesure composé par un ensemble d'items qui, tous, devraient contribuer à appréhender une même entité (ou dimension) « sous-jacente ». Cet indice traduit un degré d'homogénéité (une consistance interne) d'autant plus élevé(e) que sa valeur est proche de 1.*

7. Dans les articles publiés jusqu'ici à partir de l'enquête IVQ*, les scores utilisés dans le module d'orientation et dans le module ANLCI pondéraient différemment les items selon la nature de la tâche évaluée (de 1 à 3 points). Ainsi, le score sur les 10 items de compréhension du module d'orientation était sur 19 points. Dans cet article, certaines analyses ne permettant pas d'utiliser facilement des pondérations par item, on ne les a pas retenues. Les variantes avec ou sans pondération sont fortement corrélées (le coefficient de corrélation est de 0,97 sur le module d'orientation et 0,98 sur le module ANLCI).*

Tableau 2
Statistiques descriptives sur les trois modules

Groupe	Module d'orientation		Module ANLCI		Module Haut	
	Nombre moyen d'items réussis	Écart-type	Nombre moyen d'items réussis	Écart-type	Nombre moyen d'items réussis	Écart-type
ANLCI direct	5,7	2,1	6,5	2,6		
Intermédiaire ANLCI	7,4	1,1	6,7	1,8		
Intermédiaire Haut	8,2	1,1	10,0	0,8	12,8	3,7
Haut direct	9,3	0,7			14,5	3,7
Ensemble	**8,6**	**1,6**				
Nombre d'items	10		11		21	
Nombre d'individus	10 009		3 343		7 724	
Alpha de Cronbach	0,64		0,69		0,76	

Lecture : la partie compréhension du module d'orientation, passée par l'ensemble de la population (soit 10 009 personnes) compte 10 questions. L'alpha de Cronbach sur ces 10 questions est de 0,64. Les individus orientés directement vers le module ANLCI (« ANLCI direct ») réussissent en moyenne 5,7 questions sur 10 à cette épreuve. La partie Compréhension du module ANLCI est passée par 3 343 personnes, dont 1 058 relèvent du groupe « Intermédiaire Haut ». Ces personnes sont aussi incluses dans le total de 7 724 pour le module Haut.
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Une analyse plus fine montre cependant des irrégularités : ainsi, le groupe « ANLCI direct » réussit un peu mieux le module ANLCI que le module d'orientation, alors que c'est l'inverse pour le groupe « Intermédiaire ANLCI ». Cela s'explique par l'impact de l'erreur de mesure évoqué plus haut : le niveau du groupe des moins performants à une épreuve est sous-estimé par cette épreuve. Ainsi, le niveau du groupe « ANLCI direct » est sous-estimé par le module d'orientation ; le niveau du groupe « Intermédiaire ANLCI » est sous-estimé par le module ANLCI.

Ces résultats tendent à minimiser l'influence de l'ordre des épreuves. Par rapport au groupe « Intermédiaire ANLCI », le groupe « ANLCI direct » a été un peu plus sollicité avant le passage de l'exercice de compréhension par les exercices d'écriture et de lecture de mots du module ANLCI et devrait donc être moins motivé. Pourtant, on vient de le voir, il obtient des résultats comparables, alors qu'il était nettement moins performant au module d'orientation. De même, lors de la passation du module Haut, le groupe « Intermédiaire Haut » devrait être handicapé par rapport au groupe « Haut direct », à cause de l'exercice intermédiaire qu'il a passé avant. L'écart entre les deux sous-populations n'est cependant que de 8 points (61 % de réussite contre 69 %), alors qu'il est de 11 points sur le module d'orientation (82 % contre 93 %).

Les corrélations entre les différents scores ne sont pas très élevées (cf. tableau 3) : celle qui est mesurée entre le module d'orientation et le module ANLCI est satisfaisante, mais la confrontation entre le module Haut et les deux autres modules donne de moins bons résultats. Cela n'a rien d'étonnant, compte tenu du fait que la confrontation entre le module ANLCI et le module Haut par exemple, ne peut être faite que sur le groupe « Intermédiaire Haut », qui se caractérise par une très faible hétérogénéité pour le score au module ANLCI (en fait, la quasi totalité de cette population a entre 9 et 11 sur 11 à cette épreuve). Les valeurs des corrélations ne sont donc finalement pas aussi décevantes que cela, bien qu'elles suggèrent que les ponts entre les épreuves ne vont pas être faciles à construire, en particulier pour le module Haut.

Calcul de score par imputation simple

La première méthode que nous proposons ne cherche justement pas à établir de pont entre les groupes et se fonde sur un respect total de la procédure d'orientation. L'objectif est de calculer un score sur les 42 items de compréhension écrite que comporte l'ensemble des modules, en appliquant des règles rigides, mais simples, quand l'épreuve n'a pas été passée :

- Groupes « ANLCI » : on supposera que ces personnes auraient échoué à tous les items du module Haut ;

- Groupe « Haut direct » : on supposera que ces personnes auraient réussi tous les items du module ANLCI.

Cela revient à faire l'hypothèse suivante : une personne du groupe « Haut direct » ayant eu 0 sur 21 au module Haut aurait eu 11 sur 11 au module ANLCI et, inversement, une personne du groupe « ANLCI » ayant réussi tous les items du module ANLCI aurait échoué à tous ceux du module Haut.

Cette stricte hiérarchisation entre les sous-populations et les modules peut paraître très forte. C'est toutefois l'hypothèse qui a été faite, de façon implicite, dans toutes les publications précédentes où l'on a détaillé les résultats à l'écrit en juxtaposant une typologie en trois groupes pour les personnes en difficulté et une typologie en quatre groupes pour les autres (Micheaux et Murat, 2006 et Djider et Murat, 2006)

On obtient ainsi un score (*SM1V1* ou « score simple brut ») dont les valeurs vont en théorie de 0 à 42. La formule de ce score est :

$$SM1V1 = SMO + SMA + 11 \times 1_{cas4} + SMH$$

Tableau 3
Liens entre les épreuves (coefficients de corrélation)

	Module d'orientation	Module ANLCI	Module Haut
Module d'orientation	1		
Module ANLCI	0,48	1	
Module Haut	0,35	0,25	1

Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

où *SMO*, *SMA* et *SMH* sont respectivement les scores aux modules d'orientation, ANLCI et Haut ; $\mathbf{1}_{cas4}$ est l'indicatrice repérant les personnes relevant du groupe « Haut direct » ; *SMA* vaut 0 si la personne n'a pas passé le module ANLCI ; de même *SMH* vaut 0 si la personne n'a pas passé le module Haut.

Cette méthode opère une distinction très stricte entre les personnes orientées vers le module ANLCI (directement ou après le module Intermédiaire), qui ont au plus 21 points et celles orientées directement vers le module Haut, qui, outre leurs bons résultats au module d'orientation, se voit gratifier des 11 points complets correspondant au module ANLCI. Cela conduit à une forte bimodalité de la distribution, qui a peu de chance de renvoyer à une partition réelle de la population entre ceux qui savent peu et ceux qui savent beaucoup. On a donc « normalisé » le score en le calant sur une distribution normale de moyenne nulle et d'écart-type 1 : il s'agit du score *SM1V2* ou « score simple normalisé » (8). Cette hypothèse de normalité est bien sûr contestable : la distribution des compétences dans la population est inconnue, celle des performances dépendant par ailleurs de l'épreuve utilisée. Cependant, une telle procédure a été souvent employée (voir par exemple pour l'étalonnage des tests psychométriques Huteau et Lautrey, 1999 et Guillevic et Vautier, 1998) et renvoie à la distribution en cloche généralement observée quand on propose à une population une évaluation d'une difficulté adaptée à son niveau. Cependant, étant donné l'aspect arbitraire de la « normalisation », les analyses avec ce score devront généralement utiliser une spécification sous forme de quantiles, qui permettra de diminuer l'importance de ce choix.

Calcul de score par imputations économétriques

Si la méthode précédente a le mérite de la simplicité, ses limites sont évidentes : les hypothèses faites sont très fortes et impliquent une distribution bimodale peu crédible, que l'on est obligé de caler sur une distribution *a priori* plus probable, mais qui reste arbitraire. On a donc cherché à mieux utiliser l'information disponible, en particulier les passerelles entre les épreuves.

Le recours à des techniques économétriques simples peut être envisagé : on va chercher à modéliser les scores aux modules Haut et ANLCI à partir de celui à l'orientation, disponible pour tous et permettant ainsi de relier les scores entre eux (9). Sur les groupes ayant passé la partie compréhension du module ANLCI, on a construit une régression linéaire entre le score au module ANLCI et le score à l'orientation. Le coefficient de détermination (r^2) est de 23,2 %. L'équation obtenue est (cf. graphique I) :

$$SMA = 3{,}072 + 0{,}647 \times SMO$$

Sur les groupes ayant passé le module Haut, on a construit une régression linéaire entre le score au module Haut et le score d'orientation. Le coefficient de détermination est de 12,5 % et l'équation est :

$$SMH = 0{,}444 + 1{,}502 \times SMO$$

On va utiliser ces deux équations pour imputer des scores pour les parties qui n'ont pas été passées (10). On se fonde donc sur les hypothèses suivantes :

- la relation entre le score *SMO* et *SMA* observée sur les groupes ayant passé la partie compréhension du module ANLCI est extrapolée aux personnes orientées directement vers le module Haut, la hiérarchisation des épreuves étant supposée la même ;

- la relation entre le score *SMO* et *SMH* observée sur les groupes ayant passé le module Haut est extrapolée aux personnes qui ne l'ont pas passé, la hiérarchisation des épreuves étant supposée la même.

La deuxième hypothèse est la plus forte, car on est conduit à prolonger l'équation sur un champ où elle n'a pas été établie : très peu de person-

---

8. De façon plus précise, on a d'abord isolé la proportion de personnes ayant la note minimale : X_1 % ont 2 sur 42. On a déterminé le quantile de la loi normale (0,1) correspondant à cette proportion : Y_1. Ensuite, pour chaque individu ayant 2 sur 42, on a effectué une série de tirages aléatoires suivant la loi normale (0,1) en retenant le premier qui était inférieur à Y_1 comme valeur pour SM1V2. *De même, on calcule X_2, la proportion de personnes ayant 3 sur 42, Y_2 le quantile de la loi normale correspondant à X_1+X_2 et on effectue pour ce cas une série de tirages aléatoires jusqu'à trouver une valeur comprise entre Y_1 et Y_2. Les cas suivants se traitent de façon semblable.*

9. Une autre piste a été explorée : chercher à modéliser la relation entre les scores SMO, SMA *et* SMH *sur le groupe « Intermédiaire Haut », qui a passé toutes les épreuves. L'information est plus riche, mais sur une population plus ciblée. Les résultats obtenus ne sont pas très différents de ceux présentés ici.*

10. Des facteurs aléatoires auraient pu être introduits dans l'imputation, à partir des résidus des équations, pour tenir compte de l'imprécision de la relation. Cependant, vu la faiblesse relative des r^2, la part de ces facteurs aléatoires aurait été trop importante dans le score final. D'autre part, notre objectif dans cet article est plutôt de mettre les différents scores sur une échelle commune (ce qui apparaîtra plus clairement dans la partie consacrée aux modèles de réponse à l'item) que de chercher à faire une imputation proprement dite.

nes appartenant aux groupes « Haut direct » et « Intermédiaire Haut », ont obtenu moins de 7 points sur 10 à l'exercice d'orientation, ce qui rend l'estimation du lien entre le score *SMO* et le score *SMH* fragile dans cette zone. Or, beaucoup de personnes des groupes « ANLCI » se concentrent justement en dessous de 7 points de réussite au module d'orientation. En revanche, comme un nombre non négligeable d'individus ayant bien réussi l'orientation passent le module ANLCI, la première hypothèse est moins problématique. Par ailleurs, la conjonction de la censure due à l'orientation et de l'erreur de mesure sur les scores provoque d'assez graves difficultés d'estimation. L'analyse théorique du problème (cf. encadré 2) et l'étude des simulations présentées plus loin montrent en effet que l'estimation des coefficients de la régression risque d'être biaisée (11). Des modèles plus complexes devront sans doute être développés pour mieux traiter la question.

Si l'on passe cependant outre ces difficultés et les limites indiquées, on obtient le score *SM2* ou « score économétrique », par la formule suivante :

$$SM2 = SMO + SMA + (0{,}444+1{,}502\times SMO) \times \mathbf{1}_{\text{cas 1 et 2}} + SMH + (3{,}072+0{,}647\times SMO)\times \mathbf{1}_{\text{cas 4}}$$

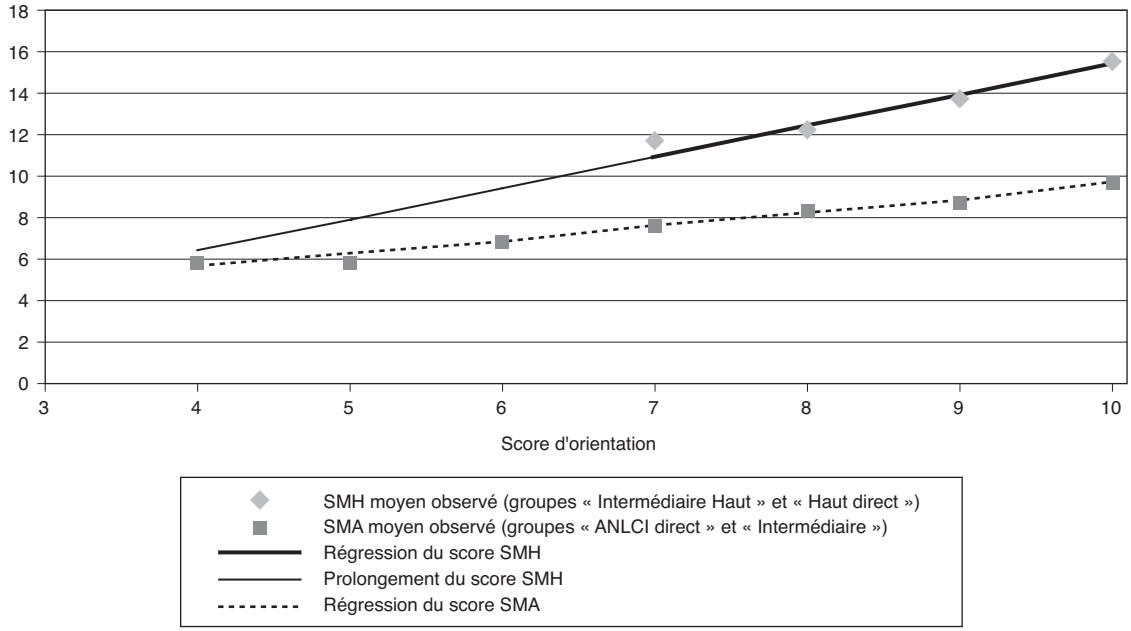
Calcul de score par l'analyse factorielle

L'analyse factorielle des items est une technique très utilisée lorsqu'il s'agit d'identifier des dimensions sous-jacentes (Megherbi *et al.*, ce numéro). L'objectif est de dégager la dimension principale, celle qui résume le mieux l'ensemble des réponses aux items. De ce point de vue, le premier facteur d'une analyse en composantes principales (ACP) offre une solution adaptée à cet objectif, puisqu'il doit correspondre à la variable numérique rendant le mieux compte de l'information contenue dans les données.

Malheureusement, les données d'*IVQ* comportent de nombreuses valeurs manquantes, puisque les individus ne passent pas tous les mêmes items. Dans ce cas, le calcul des scores facto-

11. Ce problème peut se comprendre en considérant sur le graphique le point représentant le score moyen au module haut des individus ayant eu 7 à l'exercice d'orientation. La quasi-totalité des individus ayant eu 7 à l'orientation relève des groupes intermédiaires. Le point est donc actuellement calculé sur le groupe « Intermédiaire haut ». En l'absence de censure, il l'aurait été en ajoutant le groupe « Intermédiaire ANLCI ». Or on peut naturellement supposer que ces individus ayant eu 7 à l'orientation, mais ayant échoué au module Intermédiaire sont moins performants que ceux qui ont eu 7 à l'orientation et ont réussi le module Intermédiaire. Le score moyen au module haut pour SMO = 7 diminuerait donc sensiblement s'ils étaient pris en compte. Le point est donc probablement situé trop haut.

Graphique I
Régressions des scores aux modules ANLCI et Haut sur le score au module d'orientation

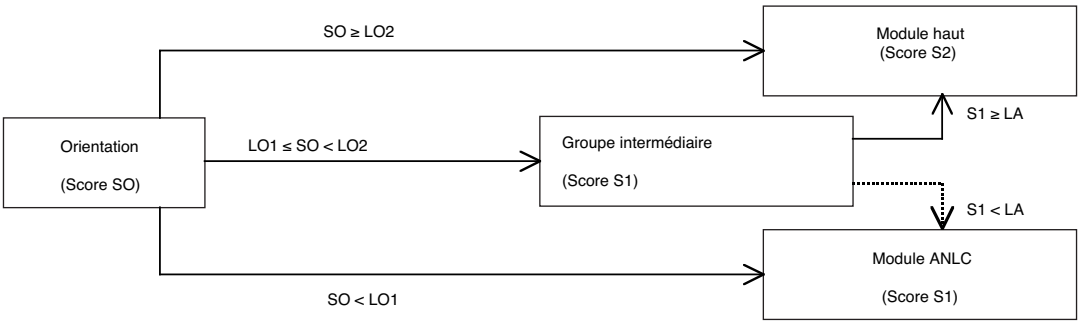


Lecture : la droite de régression de SMH est prolongée par un trait est plus fin à cause de la faiblesse des effectifs en dessous de 7 points. Quelques individus ayant eu 10 sur 10 à la partie compréhension du module d'orientation ont été orientées vers le module ANLCI, car ils ont eu des performances insuffisantes en identification de mots dans le même module. Par commodité, on a présenté les deux régressions sur le même graphique, bien que SMA et SMH soient sur des échelles distinctes, que l'enjeu de cet article est justement de relier.
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Encadré 2

MODÉLISATION ÉCONOMÉTRIQUE ENTRE LES SCORES DE L'ENQUÊTE *IVQ*

Pour poser dans une modélisation économétrique les relations entre les scores aux modules d'orientation, ANLCI et Haut, on va simplifier le processus d'orientation, en le résumant selon le schéma ci-dessous. On laisse ainsi de côté le fait que l'orientation fait aussi intervenir les résultats en lecture de mots et pas seulement en compréhension. Le trait partant du groupe intermédiaire vers le module ANLCI est en pointillé, car le passage du reste du module ANCLI ne nous apporte pas d'information dans le cadre de cette étude, sur le groupe concerné par cette orientation.



L'objectif est d'imputer une valeur pour S1 et S2 quand ces variables sont à blanc (c'est-à-dire le score S1 pour les individus orientés directement vers le module haut ; le score S2 pour ceux orientés vers le module ANLCI, directement ou en passant par le module intermédiaire). On peut poser que les trois scores dépendent d'une même compétence (celle que l'on cherche à résumer en un score global). On va supposer que cette dépendance est linéaire :

$SO_i = a + b \times C_i + e_i$ (1)

$S1_i = a' + b' \times C_i + e_i'$ (2)

$S2_i = a'' + b'' \times C_i + e_i''$ (3)

Les termes d'erreurs e_i, e_i' et e_i'' sont supposés suivre des lois normales de moyenne nulle et, respectivement, de variances σ^2, σ''^2, σ''^2 (la normalité des résidus, sans doute discutable, ne sera utilisée que dans certains cas d'illustration). Ces erreurs sont indépendantes deux à deux, ainsi qu'avec C_i.

Pour simplifier les calculs, on posera $a = 0$ et $b = 1$. Il suffit de remplacer C_i par $C_i' = a+b\times C_i$ pour obtenir un système de cette forme, les coefficients des nouvelles équations (2) et (3) s'obtenant alors facilement : b' est par exemple remplacé par $\frac{b'}{b}$ et a' par $a' - \frac{a \times b'}{b}$.

Les difficultés de cette analyse tiennent au cumul des problèmes causés par l'existence d'erreurs de mesure sur la dimension mesurée et de ceux causés par la censure des données.

En combinant, (1) et (3), on obtient :

$S2_i = a'' + b'' \times SO_i + e_i'' - b'' \times e_i$

L'estimation de cette équation par la méthode des moindres carrés ordinaires (MCO) sur l'ensemble de la population pose un premier problème. En effet, $\hat{b}''_{MCO1}$, l'estimateur correspondant, est asymptotiquement biaisé :

$$p\lim(\hat{b}''_{MCO1}) = b'' + \frac{\mathrm{cov}(e'' - b'' \times e, SO)}{V(SO)} = b'' + \frac{\mathrm{cov}(e'' - b'' \times e, C + e)}{V(SO)} = b'' - b'' \times \frac{V(e)}{V(SO)}$$

car, d'après les hypothèses initiales : cov(e'',C) = 0, cov(e,C) = 0 et cov(e'',e) = 0.

En fait, dans le cadre de notre étude, ce biais n'est toutefois pas très gênant. En effet, on ne cherche pas la « vraie » relation entre SO et S2, mais un modèle descriptif donnant la meilleure prédiction de S2 sachant SO, pour procéder à une imputation.

Le problème vient ici de ce que l'on ne peut procéder à cette estimation sur l'ensemble de la population, mais seulement pour les individus pour lesquels S2 est observé. Le biais est alors différent de ce que l'on observe sur l'ensemble de la population.

➔

Encadré 2 (suite)

En effet, si $\hat{b}''_{MCO2}$ est l'estimateur des MCO de l'équation (3) pour la population caractérisée par la relation SO ≤ LO2, on a alors :

$$p\lim(\hat{b}''_{MCO2}) = b'' + \frac{\mathrm{cov}(e'' - b'' \times e, C + e|SO \geq LO2)}{V(SO|SO \geq LO2)} = b'' - b'' \times \frac{\mathrm{cov}(C, e|SO \geq LO2) + V(e|SO \geq LO2)}{V(SO|SO \geq LO2)}$$

En effet, on montre facilement que cov(*e"*, C|*SO* ≥ *LO*2) = 0 et cov(*e"*, e|*SO* ≥ *LO*2) = 0.

Donc, $p\lim(\hat{b}''_{MCO2}) = b'' - b'' \times \frac{V(e)}{V(SO)} + b'' \times (\frac{V(e)}{V(SO)} - \frac{V(e|SO \geq LO2)}{V(SO|SO \geq LO2)} - \frac{\mathrm{cov}(C, e|SO \geq LO2)}{V(SO|SO \geq LO2)})$

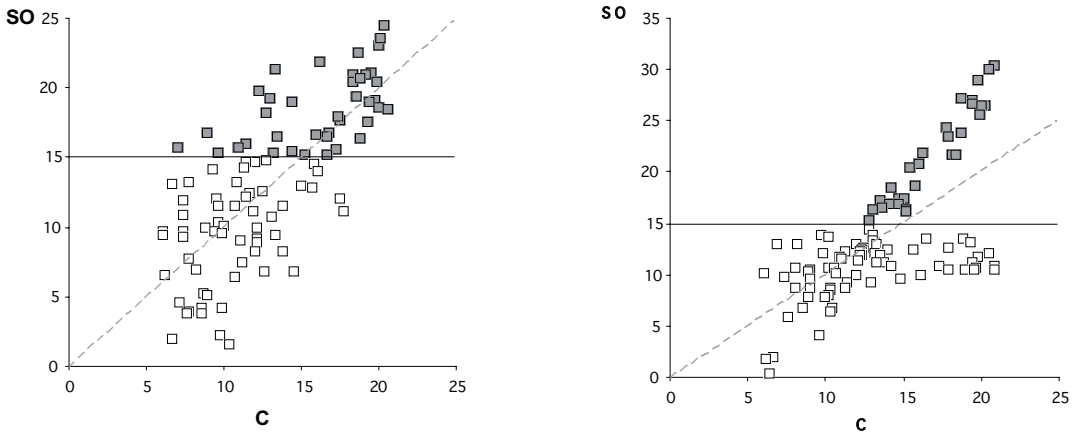
Il apparaît donc que cet estimateur est lui aussi biaisé et que le biais risque d'être différent de celui affectant l'estimateur sur l'ensemble de la population. En d'autres termes, les estimations sur les sous-populations seront biaisées par rapport aux coefficients obtenus sur l'ensemble de la population. Le biais va dépendre de la valeur du dernier terme de l'expression, celui-ci ayant d'ailleurs deux composantes. La valeur et le signe même de ces composantes ne sont toutefois pas faciles à déterminer sans hypothèses supplémentaires sur la distribution des scores et des résidus. Les graphiques ci-dessous présentent deux situations pouvant se produire.

La première situation part d'une distribution uniforme des compétences et d'une distribution normale des résidus. Dans ce cas, $(\frac{V(e)}{V(SO)} - \frac{V(e|SO \geq 15)}{V(SO|SO \geq 15)}) < 0$. En effet, les variations de SO sous la contrainte SO >15 diminuent sensiblement par rapport à celles sur l'ensemble de la population, alors que la diminution est moins forte pour le résidu (l'écart des points à la droite). Donc, la part de la variance des résidus par rapport à celle de SO augmente et $\frac{V(e)}{V(SO)} < \frac{V(e|SO \geq 15)}{V(SO|SO \geq 15)})$. Quant au terme $\frac{\mathrm{cov}(C, e|SO \geq LO2)}{V(SO|SO \geq LO2)})$, il apparaît clairement négatif : pour les faibles valeur de C, sous la contrainte SO >15, les points sont tous nettement au-dessus de la droite et les résidus sont positifs. Pour les plus fortes valeurs de C, les points sont beaucoup plus proches de la droite et le résidu se rapproche de 0, voire pourrait être négatif.

Le même type d'analyse menée sur l'autre situation (qui est une construction *ad hoc*) conduit à des résultats exactement inverses : la variance des résidus sous la contrainte SO >15 paraît alors plus faible que sur l'ensemble de la population et la première composante du biais sera alors positive. L'écart entre les points et la droite tend à croître avec C et le deuxième terme sera alors positif.

Ces deux exemples montrent la sensibilité du résultat à la distribution des résidus. Il sera donc besoin d'une analyse plus poussée pour mieux cerner ce phénomène et chercher à corriger les biais induits.

Mise en évidence des biais d'estimation sur deux exemples fictifs



Lecture : ces graphiques représentent deux formes possibles de relation entre la compétence inobservée C et le score SO. On a distingué dans les deux cas, deux populations selon que les individus se trouvaient au-dessus ou en dessous d'un certain seuil pour SO (fixé à 15).

riels n'est pas possible avec une ACP classique. Néanmoins, un algorithme spécifique a été développé pour conduire des analyses factorielles en présence de valeurs manquantes (cf. encadré 3). Cet algorithme a déjà été employé pour établir une échelle de scores, dans le cadre d'une comparaison internationale des compétences des élèves à partir de supports non traduits (Bonnet *et al.*, 2001).

Le logiciel SIMCA (Umetri, 1998) a été utilisé pour conduire cette analyse. La première valeur propre est de 7,99 et la seconde de 2,36. Le rapport de 3,38 entre les deux valeurs, montre l'importance du premier axe et conforte l'hypothèse d'une structure approximativement unidimensionnelle. Néanmoins, la configuration particulière des épreuves d'*IVQ*, où les données manquantes ne sont pas aléatoires et les exercices sont assez peu nombreux, est susceptible de fragiliser cette analyse. Le premier plan factoriel des items (cf. graphique II), c'est-à-dire la représentation des items selon leur projection sur le premier et le deuxième facteurs illustre cette difficulté. Il apparaît clairement un regroupement des items par module. Les items du module Haut sont ainsi surtout en relation avec le premier facteur, tandis que les deux autres modules s'opposent sur le deuxième facteur, à l'exception des deux items les plus difficiles du module d'orientation. En outre, si les items du module d'orientation restent fortement liés au premier facteur, ce n'est pas le cas des items du module ANLCI qui sont représentés essentiellement par le deuxième facteur. Il en ressort que le score factoriel issu de cette analyse (*SM3* ou « score PLS ») est certainement trop influencé par les items du module Haut, module qui rassemble le plus grand nombre d'individus. Par conséquent, ce score risque d'être biaisé pour les individus ayant passé uniquement le module ANLCI.

Les modèles de réponse à l'item

Les modèles de réponse à l'item (MRI) sont des modèles logistiques qui expliquent la probabilité de réussite d'un individu à un item par le niveau de compétence de l'individu et par des caractéristiques propres à l'item (cf. encadré 4). L'avantage de ces modèles est de séparer les concepts : l'aptitude d'un individu est définie indépendamment de la difficulté de l'épreuve

Encadré 3

ANALYSE FACTORIELLE AVEC L'ALGORITHME *NIPALS*

L'algorithme *NIPALS* (*Nonlinear estimation by Iterative Partial Least Squares*) sous-tend les méthodes de régression *PLS* (*Partial Least Squares*) et permet de réaliser une analyse en composantes principales avec données manquantes. Pour plus de détails, le lecteur pourra consulter l'ouvrage de référence de Tenenhaus (1998). Cet algorithme consiste en une succession de régressions locales effectuées sur les valeurs « disponibles », ce qui permet de conduire l'analyse factorielle en présence de valeurs manquantes.

Plus précisément, en reprenant les notations de Tenenhaus (1998), on désigne par $X = [x_{ij}]$ la matrice des données avec n individus en lignes et p variables (items) en colonnes. La décomposition en composantes principales sur a facteurs peut s'écrire :

$$X = \sum_{h=1}^{a} t_h p_h'$$

où $t_h = (t_{h_1},...,t_{h_n})'$ et $p_h = (p_{h_1},...,p_{h_p})'$ sont respectivement les composantes principales et les vecteurs directeurs des axes principaux.

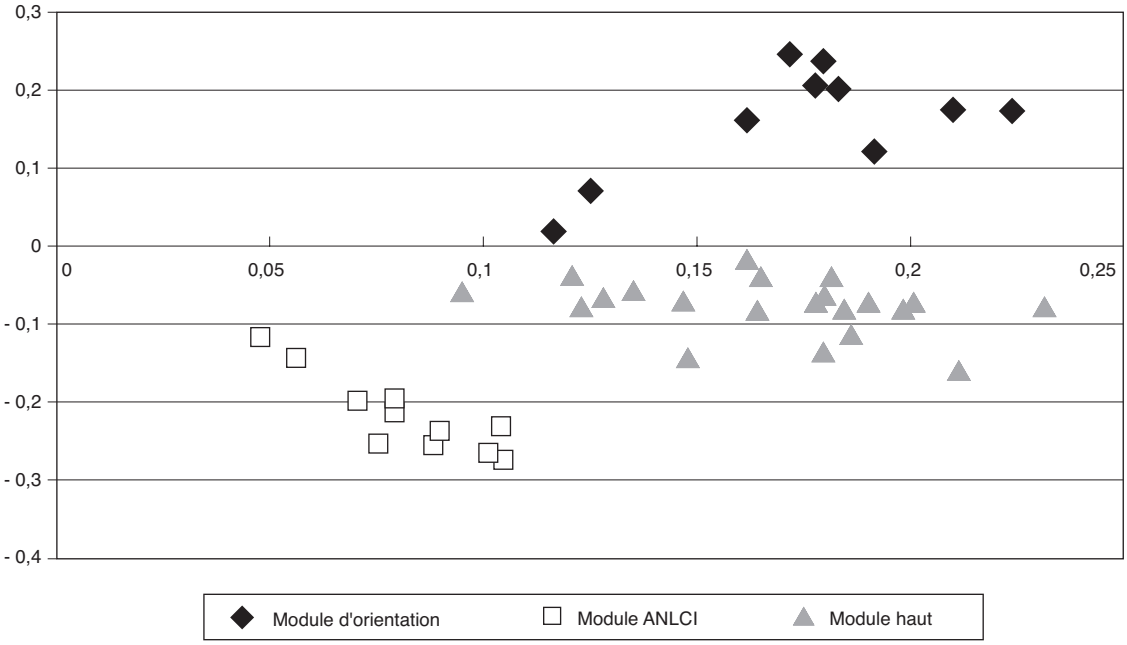
Nous nous limiterons dans la description qui va suivre au seul premier facteur.

Dans l'équation donnée ci-dessus, les variables t_h et p_h sont en fait considérées comme des paramètres à estimer. Ainsi, la « saturation » p_{1j} est la pente de la droite de régression de l'item x_j sur le premier facteur t_1. Il s'agit en fait de la discrimination de l'item (sur la discrimination cf. encadré 4 et Meguerbi *et al.*, ce numéro). Le « score factoriel » t_{1i} est, quant à lui, la pente de la droite de régression du « pattern » de réponses $x_i$de l'individu i sur la première composante p_1. Il s'agit en fait du score de l'individu, pondéré par les discriminations des items (théoriquement proche d'un score issu d'un modèle de réponse à l'item, à deux paramètres, présenté plus loin).

Pour trouver ces valeurs, le premier facteur t_1 est initialisé par le premier item puis l'algorithme consiste en une succession de régressions : p_{1j} est calculé comme la pente de la droite de régression de x_j sur t_1 ; la régression de x_i sur p_1 – avec p_{1j} obtenu précédemment – permet d'estimer t_{1i} ; la régression de x_j sur le « nouveau » facteur t_1 donne une nouvelle valeur de p_{1j} ; etc.

Notons que cette procédure conduit exactement aux mêmes résultats que l'analyse en composantes principales dans le cas où il n'y a pas de données manquantes. Dans le cas de données manquantes, les régressions sont effectuées sur les seules valeurs « disponibles ».

Graphique II
Résultats de l'analyse PLS



Lecture : il s'agit des projections des indicatrices de réussite des items sur les deux premiers facteurs de l'analyse PLS.
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Encadré 4

LES MODÈLES DE RÉPONSE À L'ITEM (MRI)

Historiquement, le premier MRI est le modèle dit « à un paramètre », ou modèle de Rasch (1960). Il cherche à exprimer la probabilité d'un individu *i* de réussir un item *j*, sous la forme suivante :

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

où Y_i^j est la réponse à l'item *j* de l'individu *i*, b_j le paramètre de difficulté de l'item *j* et θ_i le niveau de compétence de l'individu *i*.

Les individus sont définis par leur niveau de compétence, qui correspond à leur position sur la dimension latente (ou trait latent) θ. Les items, quant à eux, sont uniquement caractérisés par leur niveau de difficulté, c'est-à-dire leur position sur la dimension latente. L'avantage de ce type de modèle est de placer les niveaux de compétence des individus et les paramètres de difficulté des items sur la même échelle, ce qui facilite l'interprétation des résultats. Par exemple, un individu a 50 % de chances de réussir un item de difficulté égale à son niveau de compétence.

Dans le modèle à deux paramètres, une caractéristique d'item est ajoutée : la discrimination, c'est-à-dire la sensibilité de l'item à une variation du niveau de compétence. Ce modèle s'écrit :

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

où *D* est un facteur d'ajustement qui vaut 1,7 et a_j le paramètre de discrimination de l'item *j*.

Le paramètre de discrimination, aussi appelé « pente » de l'item, est positif. Lorsque la discrimination est élevée, une variation du niveau de compétence entraîne une variation importante de la probabilité de répondre correctement.

Il existe aussi un modèle à trois paramètres, qui introduit une « asymptote », pour tenir compte du fait qu'un individu de compétence extrêmement faible peut avoir une probabilité non nulle de réussir une question, par exemple en répondant au hasard dans une question à choix multiples.

Les paramètres a_j, b_j et θ_i ne sont pas uniques : en appliquant une transformation linéaire donnée à ces paramètres, on peut obtenir une autre solution acceptable. Il faut donc imposer des contraintes identifiantes (par exemple, une moyenne nulle des θ_i sur l'ensemble de la population). Cette propriété va être utilisée lors des ancrages entre épreuves.

➔

et, inversement, la difficulté des items n'est pas fonction du niveau de compétence des individus. On fait donc l'hypothèse qu'un item du module Haut présente la même difficulté pour les individus orientés vers ce module que pour les individus ayant des difficultés avec l'écrit. Les deux groupes diffèrent simplement par leur niveau de compétence. Une fois ce cadre théorique accepté (il est possible de le tester, au moins partiellement), ces modèles semblent donc particulièrement adaptés à notre étude.

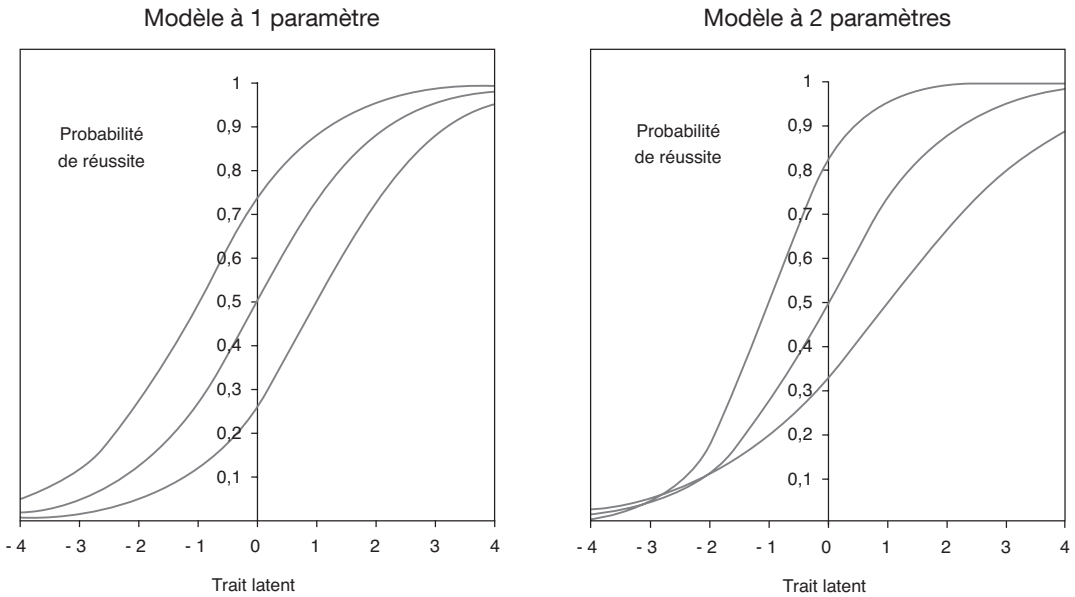
En effet, ils sont très souvent utilisés quand les individus ne passent pas la même épreuve. Le cas des « cahiers tournants » est un exemple classique : pour pouvoir tester un grand nombre d'items sans trop rallonger la durée de l'épreuve, on constitue différents cahiers, comportant une partie commune et une partie variable. Sans l'usage des modèles de réponse à l'item, la constitution d'un score est délicate : le nombre de bonnes réponses ne permet pas de comparer équitablement deux individus ayant passé des cahiers différents, si la difficulté de ceux-ci n'est pas la même. L'une des méthodes d'ancrage souvent utilisée va consister à fixer les paramètres de la partie commune (par exemple, aux valeurs obtenues sur l'ensemble des individus) et à estimer ensuite les paramètres des items des parties variables sur chaque population, avec cette contrainte sur la partie commune. La comparabilité se trouve ainsi assurée et on peut estimer les scores avec ces paramètres pour l'ensemble des items passés, de la partie commune ou non.

Cependant, *IVQ* n'entre pas dans ce cadre idéal, où les « trous » sont aléatoires : dans le cas présent, si les individus ne passent pas le module Haut, c'est parce qu'ils ont eu des résultats insuffisants au module d'orientation. Ce processus risque de provoquer des biais dans l'estimation qui va tenter l'ancrage entre le module Haut et le module d'orientation par exemple : les items du module d'orientation vont apparaître très faciles pour les personnes orientées vers le module Haut, non seulement parce que ces individus sont plus compétents que la moyenne,

Encadré 4 (suite)

Ces modèles sont très largement employés dans le domaine de l'évaluation des compétences, tant au niveau national qu'international (à ce sujet, lire d'Haultefoeuille *et al.*, 2002 et Rocher, 2003). Ils sont parfois utilisés, dans d'autres domaines, comme celui de la santé ou des valeurs. Ainsi, les MRI ont récemment servi à l'analyse des représentations de la pauvreté en France (Accardo et de Saint-Pol, 2009). Pour plus de précisions sur ces modèles, le lecteur pourra consulter, en français, Dickes *et al.* (1994), Juhel (1999).

Probabilité estimée de réussite en fonction du niveau de compétence (θ) pour les modèles de réponses à 1 et 2 paramètres, pour différents items



Lecture : chaque courbe représente pour un item donné la relation entre la compétence et la réussite à l'item. Ainsi, dans le modèle à 1 paramètre, un individu de compétence 0 a un peu plus de 25 % de chances de réussir l'item le plus difficile (courbe à droite). Dans le modèle à 2 paramètres, les courbes n'ont pas la même pente et peuvent éventuellement se croiser.

mais parce que s'ils passent le module Haut, c'est parce qu'ils ont réussi les items d'orientation (12). Inversement, pour les personnes passant le module ANLCI, parce qu'elles ont échoué au module d'orientation, les items de ce module vont paraître plus difficiles qu'ils ne sont vraiment (13). Le calcul du score des individus avec ces paramètres mal estimés risque donc d'être inexact.

Il existe cependant des techniques spécifiques pour estimer les paramètres dans le cadre d'un test adaptatif (cf. encadré 5). Sur les données issues des évaluations nationales américaines (*National Assessment Educational Programme, NAEP*), Bock et Zimowski (1998) ont étudié le cas d'une épreuve en deux temps (« *two-stage design* ») avec module d'orientation vers trois groupes de niveaux de compétences différents. Ces auteurs ont utilisé les MRI à trois paramètres avec une estimation concourante sur un ensemble de questions à choix multiples (QCM). La comparaison avec les paramètres des items tels qu'estimés sur l'ensemble des niveaux de compétences est très positive. Au-delà du fait qu'il s'agit de QCM et d'un modèle à trois paramètres, la principale différence avec *IVQ* tient plus au fait que les items communs aux trois groupes sont moins faciles que ceux du module d'orientation d'*IVQ*. Plus précisément, ces items communs sont mieux répartis le long du *continuum* de difficulté, à la différence d'*IVQ* où ces items sont adaptés pour la détection d'individus ayant des difficultés en lecture. Cette différence rend probablement plus fragile dans notre cas le placement des groupes sur la même échelle.

Plus récemment, Hanson et Béguin (2002) ont fait des simulations permettant de comparer différentes méthodes d'estimation dans le cas d'un ajustement entre groupes. Avec un échantillon de 3 000 individus, 10 items communs et une différence d'un écart-type entre les deux groupes, la méthode concourante employée avec le logiciel BILOG (du Toit, 2003) est supérieure à toutes les méthodes séparées (parmi ces dernières, celle de Stocking et Lord est la meilleure).

Les deux méthodes d'estimations à partir des modèles de réponse à l'item ont été appliquées sur les données (14) : l'une (*SM4V1*) sera appelée « estimation séparée », l'autre (*SM4V2*) « estimation concourante » (15).

Nous allons détailler la procédure pour l'estimation séparée en ce qui concerne l'ajustement des paramètres de difficulté. Dans un premier temps, on effectue trois modélisations (cf.tableau 4) :

- sur les groupes « ANLCI », on construit un modèle avec les items du module d'orientation et du module ANLCI ;

- sur le groupe « Intermédiaire Haut », on construit un modèle avec l'ensemble des items ;

- sur le groupe « Haut direct », on construit un modèle avec les items du module d'orientation et du module haut.

L'objectif est ensuite de placer les coefficients sur la même échelle, en calant les deux séries de coefficients sur la même difficulté moyenne. C'est le groupe « Intermédiaire Haut » qui a été choisi comme référence, car il est en position médiane et a passé l'ensemble des items. Les paramètres associés aux items du module d'orientation et du module ANLCI pour les groupes « ANLCI » sont systématiquement plus élevés que pour le groupe « Intermédiaire Haut » (cf. graphique III) : les coefficients pour le groupe « Intermédiaire Haut » sont presque tous inférieurs à - 3, alors qu'ils s'échelonnent entre - 2,5 et 3 pour les groupes « ANLCI ». En effet, les groupes « ANLCI » étant moins compétents, les items sont moins bien réussis, donc

12. Prenons un exemple caricatural : supposons que le module d'orientation et le module Haut se réduisent chacun à un item (si la personne réussit l'item d'orientation, elle passe l'item « haut »), qui, testés auparavant sur l'ensemble de la population, sont de difficulté identique, réussis par 50 % des individus. Pour les personnes orientées vers le module Haut, le taux de réussite de l'item d'orientation sera de 100 %, tandis que celui sur l'item « haut », du fait de la corrélation nécessairement imparfaite entre les items, sera sans doute inférieur (même s'il reste certainement supérieur à 50 %) et apparaîtra donc comme plus difficile, en contradiction avec les résultats sur l'ensemble de la population.

13. L'existence du module Intermédiaire rend par ailleurs l'analyse plus complexe : si le module d'orientation va paraître « trop » facile par rapport au module ANLCI sur le groupe « ANLCI direct », le biais risque d'être inverse sur le groupe « Intermédiaire ANLCI », qui se caractérise par des résultats moyens au module d'orientation et insatisfaisants au module ANLCI. Ce contraste a déjà été relevé lors de l'analyse des taux de réussite des sous-populations par épreuve (cf. tableau 2).

14. Nous ne présenterons ici que les résultats pour le modèle à 2 paramètres, mais des estimations avec le modèle à 1 paramètre ont aussi été effectuées, conduisant à des résultats légèrement moins satisfaisants.

15. L'estimation séparée est une version « souple » de l'ancrage présenté dans le texte. On estime sur chacune des sous-populations un modèle avec l'ensemble des items passés. Ensuite, on va chercher non à fixer les paramètres de chaque item commun à la même valeur pour les différentes populations, mais plutôt à relier l'épreuve commune dans son ensemble entre les populations. On autorise tel item donné du module haut à être plus difficile pour le groupe 3 que pour le groupe 4 ; en revanche, la difficulté moyenne du module haut doit être quasi identique pour les deux population. L'estimation concourante, sous certaines hypothèses sur la distribution des compétences, permet d'estimer en une seule étape les paramètres des items et ceux des individus (cf. encadré 5 pour plus de détails).

apparemment plus difficiles (16). L'ancrage consiste à recaler les séries l'une sur l'autre. Cette procédure est d'autant plus fiable que la hiérarchie des items est à peu près identique sur les deux groupes, ce qui ne semble pas vraiment le cas ici : il serait sans doute souhaitable de supprimer les items dont le comportement est le moins fiable. Cependant, pour travailler sur un nombre d'items suffisamment important, nous avons calculé les coefficients de passage par la méthode de Stocking et Lord, sur les 21 items des modules d'orientation et ANLCI. La mauvaise correspondance entre les deux séries de paramètres doit toutefois inciter

16. Le décalage est plus grand pour les items du module ANLCI, du fait du biais causé par le processus d'orientation : par construction, les individus du groupe « Intermédiaire Haut » ont bien réussi ces items (sinon, ils se trouveraient dans le groupe « Intermédiaire ANLCI ») et ceux-ci semblent donc pour eux particulièrement faciles.

Encadré 5

AJUSTEMENT DES MÉTRIQUES (*EQUATING*)

Quelle que soit la méthode d'estimation retenue, pour des raisons liées à l'identification du modèle, la variable θ n'est définie qu'à une transformation linéaire près. En général, on fixe la moyenne à 0 et l'écart-type à 1. Dans le cas d'une estimation pour différents groupes d'individus, un ajustement est donc nécessaire pour placer les compétences des individus des différents groupes sur la même échelle.

Estimations séparées

Dans le cas d'une estimation sur plusieurs groupes d'individus non équivalents avec des items communs, une première méthode consiste à estimer séparément les niveaux de compétences θ dans chacun des groupes.

Les paramètres estimés indépendamment sur deux groupes différents doivent alors être reliés par une relation linéaire. Dans le cas du modèle à deux paramètres, si on effectue le changement d'échelle $\theta^* = A\theta + B$, alors les nouveaux paramètres d'items seront tels que $b^* = Ab + B$ et $a^* = a/A$. Ainsi la probabilité estimée pour un individu de réussir un item est inchangée.

Stocking et Lord (1983) ont proposé une procédure pour estimer A et B. Elle consiste à minimiser l'écart entre les deux scores « vrais » d'une estimation à l'autre (*i.e.* d'un jeu de paramètres à l'autre). Plus précisément, soient (a_{j1}, b_{j1}) les paramètres de l'item *j* estimés sur les individus du premier groupe, (a_{j2}, b_{j2}) les paramètres de l'item *j* estimés sur les individus du second groupe et (a^*_{j2}, b^*_{j2}) la valeurs des paramètres du second groupe transformés sur l'échelle du premier. On cherche les « meilleurs » A et B tels que $b^*_{j2} = Ab_{j2} + B$ et $a^*_{j2} = a_{j2} / A$.

Pour trouver A et B, Stocking et Lord (1983) proposent de minimiser F :

$$F = \frac{1}{n}\sum_{i=1}^{n}(\xi_i - \xi_i^*)^2$$

où ξ_i et ξ_i^* sont les scores « vrais » des individus *i* du premier groupe : $\xi_i = \sum_{j=1}^{K} P_{j1}(\theta_i)$ et $\xi_i^* = \sum_{j=1}^{K} P^*_{j2}(\theta_i)$ avec P_{j1} et P^*_{j2} les probabilités estimées respectivement avec les paramètres (a_{j1}, b_{j1}) et (a^*_{j2}, b^*_{j2}).

Un programme de résolution d'équations non-linéaires doit être mis en œuvre. Le programme informatique ST de l'Université de l'Iowa a été utilisé ici (Hanson *et al.*, 2004).

Estimation « concourante »

Le logiciel BILOG (du Toit, 2003) permet d'estimer dans un premier temps les paramètres des items, et dans un second temps, les compétences des individus.

Grâce à l'hypothèse d'indépendance locale (indépendance des items à θ fixé), on peut écrire la fonction de vraisemblance marginale :

$$L(\theta, a, b) = \prod_{i=1}^{n} \int_{\theta} \prod_{j=1}^{K} P_i^j(\theta)^{y_{ij}} [1 - P_i^j(\theta)]^{1-y_{ij}} \, g(\theta) d\theta$$

pour *n* individus, *K* items et la distribution des θ ayant pour fonction de densité *g*.

Si *g* est connue, il est possible d'estimer les paramètres des items sans avoir besoin d'estimer les θ. En pratique, la distribution continue des θ est approchée empiriquement grâce à une distribution finie reposant sur des « points de quadrature » choisis le long du continuum des θ (Mislevy, 1987). L'algorithme EM - maximisation de vraisemblance sur données incomplètes - est utilisé.

Dans le cas d'une épreuve sur plusieurs groupes non-équivalents avec des items communs, cette procédure permet d'estimer directement les paramètres des items pour l'ensemble des groupes. Cette estimation « concourante » suppose que la distribution des niveaux de compétences soit correctement estimée, ce qui en pratique n'est pas évident si les groupes sont très différents.

Une fois les paramètres d'items connus, plusieurs méthodes permettent d'estimer les θ. Dans le cas d'une modélisation portant sur plusieurs groupes non-équivalents, il est préférable d'entreprendre une méthode bayésienne utilisant les « points de quadrature » issus de l'estimation des paramètres d'items (Bock et Zimowski,1998). En général, la moyenne et l'écart-type des θ sont fixés sur un groupe pris pour référence et ne sont pas contraints sur les autres groupes.

Tableau 4
Difficulté et discrimination des items pour le modèle à estimation séparée (MRI)

		Groupes « ANLCI »		Groupe « Intermédiaire Haut »		Groupe « Haut Direct »	
Mod.	Item	Discrimination	Difficulté	Discrimination	Difficulté	Discrimination	Difficulté
MO	I01	0,53	- 1,57	0,33	- 4,92	0,56	- 4,41
	I02	0,47	- 0,08	0,16	- 3,73	0,33	- 5,45
	I03	0,29	- 1,44	0,14	- 5,35	0,20	- 13,35
	I04	0,35	- 2,43	0,24	- 5,32	0,19	- 9,96
	I05	0,18	- 2,28	0,26	- 2,94	0,36	- 3,42
	I06	0,19	2,75	0,27	0,45	0,39	- 0,96
	I07	0,33	- 1,79	0,31	- 3,84	0,24	- 11,12
	I08	0,31	- 1,17	0,22	- 3,90	0,19	- 12,82
	I09	0,80	- 1,22	0,50	- 4,13	0,62	- 4,90
	I10	0,59	- 1,72	0,43	- 4,17	0,40	- 6,94
MA	I11	0,41	- 0,28	0,31	- 3,47		
	I12	0,68	- 1,44	0,39	- 6,40		
	I13	0,39	- 1,60	0,39	- 5,46		
	I14	0,59	- 1,40	0,31	- 7,97		
	I15	0,07	- 8,58	0,20	- 6,39		
	I16	0,64	0,28	0,37	- 4,17		
	I17	0,31	0,16	0,24	- 6,74		
	I18	0,30	- 1,32	0,22	- 5,25		
	I19	0,77	- 0,39	0,49	- 3,95		
	I20	0,51	0,25	0,23	- 3,55		
	I21	0,16	1,31	0,17	- 5,85		
MH	I22			0,71	1,10	0,97	0,24
	I23			0,36	- 1,68	0,37	- 2,16
	I24			0,58	- 1,41	0,61	- 1,94
	I25			0,65	- 1,27	0,67	- 1,71
	I26			0,81	- 1,58	0,76	- 2,00
	I27			0,27	- 1,57	0,39	- 1,60
	I28			0,48	- 1,49	0,60	- 1,89
	I29			0,56	- 1,56	0,74	- 1,79
	I30			0,32	1,79	0,41	0,79
	I31			0,43	1,88	0,48	1,16
	I32			0,54	- 0,57	0,72	- 1,07
	I33			0,44	0,13	0,64	- 0,49
	I34			0,44	- 1,33	0,68	- 1,31
	I35			0,23	- 0,70	0,28	- 0,90
	I36			0,30	0,56	0,36	- 0,03
	I37			0,49	- 1,17	0,55	- 1,73
	I38			0,44	- 0,29	0,43	- 0,59
	I39			1,18	- 1,51	1,11	- 1,94
	I40			0,57	- 0,13	0,56	- 0,57
	I41			0,66	- 0,84	0,67	- 1,44
	I42			0,63	- 0,45	0,76	- 0,81

Lecture : Le coefficient de difficulté varie entre -∞ et +∞ (un peu comme un coefficient de régression logistique). Plus il est négatif, plus l'item est considéré comme facile. La discrimination est grossièrement équivalente à la corrélation entre la réussite à l'item et le score global.
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

à la prudence. L'équation, qui intègre aussi le même type d'ajustement sur les paramètres de discrimination, est alors :

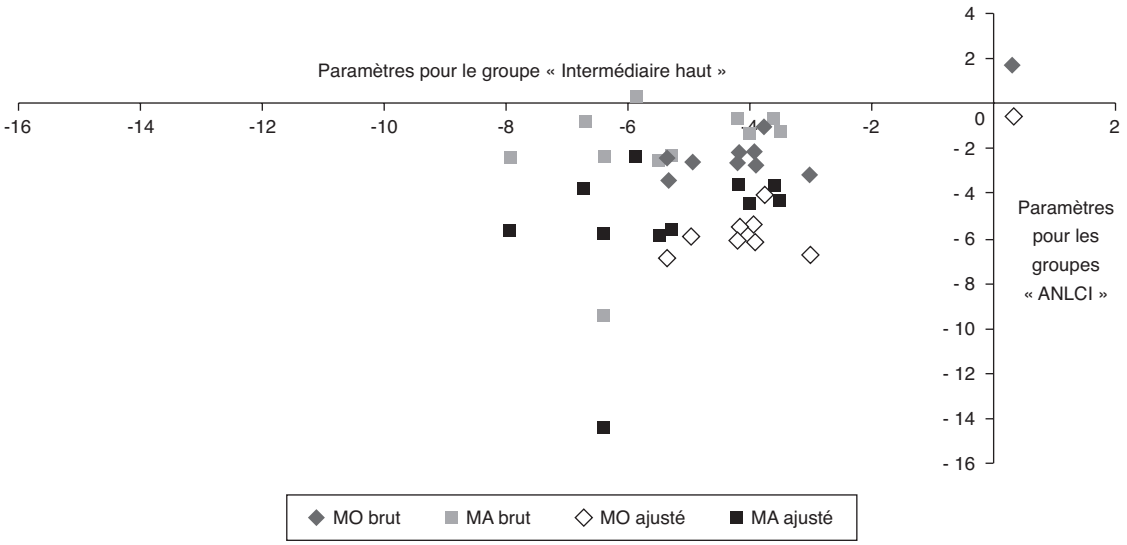
$$DIFFIC_{\text{pour les cas 1\&2 sur l'échelle cas 3}} = -3{,}07 + 1{,}24 \times DIFFIC_{\text{pour les cas 1\&2 sur l'échelle cas 1\&2}}$$

Les paramètres pour les groupes « ANLCI » sont recalés pour avoir une moyenne à peu près identique à celle obtenue sur le groupe « Intermédiaire Haut » (cf. tableau 5) : en effet, elle est alors de - 4,49 pour le groupe « ANLCI » contre - 4,62 pour le groupe « Intermédiaire Haut ». Il existe cependant un décalage dans la hiérarchisation des modules d'orientation et ANLCI. Pour le groupe « Intermédiaire haut », le module ANLCI paraît nettement plus facile, ce qui est normal puisque par construction les individus de ce groupe réussissent ces items. Pour les groupes « ANLCI », même après recalage, les deux modules paraissent de difficulté comparable.

On applique alors, pour les individus relevant des groupes « ANLCI », la même transformation au score obtenu avec le MRI initial, afin de les mettre sur la même échelle que les individus du groupe « Intermédiaire Haut ».

La même procédure est appliquée au groupe « Haut direct » pour les recaler sur l'échelle du cas « Intermédiaire Haut » (cf. graphique IV). La hiérarchie des items du module d'orientation n'apparaît pas vraiment proche sur les deux groupes, alors que la cohérence est très nette sur les items du module Haut. C'est pourquoi on n'a calculé les coefficients de Stocking et Lord que sur cette série d'items, du fait qu'ils étaient en nombre suffisant.

Graphique III
Paramètres des modules d'orientation (MO) et ANLCI (MA) estimés avant et après ajustement, sur les groupes « ANLCI », d'une part et le groupe « Intermédiaire Haut », d'autre part



Lecture : chaque point représente un item avant et après ajustement. En abscisse, on lit sa difficulté dans l'estimation sur le groupe « Intermédiaire haut » ; en ordonnée, la difficulté dans l'estimation sur les groupes « ANLCI ».
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Tableau 5
Difficulté moyenne des modules d'orientation et ANLCI avant et après ajustement

	MRI initial sur le groupe « Intermédiaire haut »	MRI initial sur les groupes « ANLCI »	MRI sur les groupes « ANLCI » recalé sur l'échelle du groupe « Intermédiaire haut »
Module d'orientation	- 3,79	- 1,10	- 4,43
Module ANLCI	- 5,38	- 1,18	- 4,54
Ensemble	- 4,62	- 1,14	- 4,49

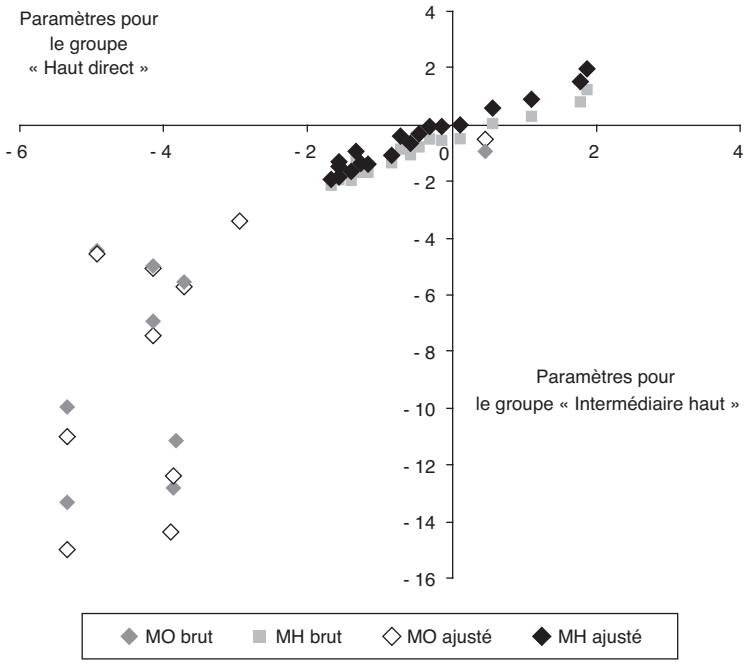
Lecture : la difficulté moyenne de chaque épreuve est obtenue en faisant la moyenne des difficultés des items qui la compose.
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Confrontation des scores

La confrontation entre les différents scores a été simplifiée en fixant leur moyenne à 0 et leur écart-type à 1. Les corrélations sont fortes, sans être toujours très proches de 1, ce qui montre que les méthodes ne sont pas équivalentes (cf. tableau 6).

Une certaine proximité apparaît entre le score issu de l'analyse factorielle PLS (*SM3*) et le score économétrique (*SM2*), alors que le score PLS est nettement moins bien corrélé avec les autres scores (par exemple, la corrélation n'est que de 0,80 avec le « score simple brut »). Les deux scores issus des modèles de réponse à l'item sont assez fortement liés entre eux, mais

Graphique IV
Paramètres des modules d'orientation (MO) et Haut (MH) estimés avant et après ajustement, sur le groupe « Intermédiaire Haut » d'une part et le groupe « Haut direct » d'autre part



Lecture : chaque point représente un item avant et après ajustement. En abscisse, on donne sa difficulté dans l'estimation sur le groupe « Haut direct » ; en ordonnée, la difficulté dans l'estimation sur le groupe « Intermédiaire Haut ».
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Tableau 6
Corrélations entre les différents scores calculés

Type de score		SMO	SMA	SMH	SM1V1	SM1V2	SM2	SM3	SM4V1	SM4V2
Scores bruts	SMO	1								
	SMA	0,48*	1							
	SMH	0,35**	0,25***	1						
Imputation simple	Brut (SM1V1)	0,78	0,81*	0,97**	1					
	Normalisé (SM1V2)	0,74	0,84*	0,91**	0,93	1				
Imputation économétrique (SM2)		0,83	0,74*	0,96**	0,88	0,93	1			
Score PLS (SM3)		0,80	0,64*	0,98**	0,80	0,87	0,97	1		
Scores MRI	Estimation séparée (SM4V1)	0,78	0,86*	0,97**	0,98	0,96	0,93	0,88	1	
	Estimation concourante (SM4V2)	0,76	0,85*	0,95**	0,93	0,98	0,95	0,91	0,98	1

** : Seulement pour les groupes 1, 2 et 3.*
*** : Seulement pour les groupes 3 et 4.*
**** : Seulement pour le groupe 3.*

Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

aussi avec les scores les plus frustres, et cela plus nettement pour l'estimation séparée (*SM4V1*).

L'écart entre les quatre groupes diffère selon les méthodes (cf. tableau 7) : ainsi pour le score « simple brut », le groupe « ANLCI direct » se trouve à - 1,78 écart-type de la moyenne, alors que les personnes relevant du groupe « Haut direct » sont à 0,55 au-dessus, soit un écart de 2,32 écarts-types. Pour la plupart des autres scores, l'écart est inférieur à 2 écarts-types. Ce sont surtout les personnes qui appartiennent au groupe « Intermédiaire ANLCI » qui apparaissent les plus sensibles à la méthodologie utilisée : pour le premier score, elles se trouvent presque aussi éloignées de la moyenne que les personnes du groupe « ANLCI direct » (1,58 écart-type en-dessous), alors que pour le score issu de l'analyse PLS, l'écart est 3 fois moins grand (seulement 0,5 écart-type).

Les scores diffèrent aussi selon leur distribution (cf. graphique V). Comme attendu, le score « simple brut » se caractérise par une forte bimodalité : les groupes « ANLCI » se trouvent concentrées en bas de la distribution, tandis que les groupes « Haut direct » et « Intermédiaire Haut », nettement séparés des premiers, se trouvent en haut. Le score « simple normalisé » conserve cette distinction, mais rapproche les populations et donne, par construction, à la distribution une forme normale. Le score économétrique et le score PLS, dont on a vu la proximité par l'analyse des corrélations entre scores, ont une distribution asymétrique, qui tient en partie au fait, qu'au contraire des deux scores précédents, les personnes relevant des groupes « ANLCI » se trouvent mieux réparties dans la distribution (un nombre non négligeable de ces personnes se trouvent au même niveau que des personnes appartenant au groupe « Haut direct »). Les scores des modèles de réponse à l'item maintiennent eux une distinction assez nette selon le résultat de la procédure d'orientation. Ils se distinguent assez nettement par la forme de la distribution obtenue. L'estimation séparée a un caractère légèrement bimodal, qui rappelle celui du score « simple brut », avec lequel elle est effectivement assez corrélée. Il est probable que ce phénomène soit dû à la difficulté que nous avons eu à relier les groupes ANLCI et le groupe « Intermédiaire Haut ». La hiérarchisation des items communs selon leur difficulté était en effet assez différente entre les deux groupes.

Corrélation des différents scores avec le diplôme, l'âge et le sexe

La confrontation avec les variables sociodémographiques disponibles dans l'enquête va apporter un éclairage intéressant à la fois pour juger de la qualité des scores et pour avoir une idée de la sensibilité des résultats à la méthodologie retenue. La corrélation avec le niveau scolaire est élevée (cf. tableau 8) : entre un tiers et deux cinquièmes de la variance des scores peut être expliqué par le niveau scolaire (17). Cette forte corrélation était attendue, du fait que la compétence mesurée (la compréhension de l'écrit) se développe principalement à l'école. Cette corrélation varie cependant assez nettement d'un score à l'autre : le coefficient de détermination est proche du tiers pour le score « simple brut »

17. Ce niveau comporte 8 modalités : jamais scolarisé ou sans diplôme et n'a pas dépassé le primaire, pas de diplôme et a dépassé le primaire, CEP, BEPC-CAP-BEP, Bac sans études supérieures, Bac avec études supérieures mais pas de diplôme à leur issue, Bac+2, Bac+3 ou plus. Les personnes en cours d'études ont été exclues. Les scores moyens par niveau n'ont pas été présentés, mais pour tous les scores, on trouve une hiérarchisation très nette des moyennes.

Tableau 7
Scores moyens des quatre groupes selon les différents scores calculés

		Groupe			
		ANLCI direct	Interm. ANLCI	Interm. Haut	Haut direct
Imputation simple	Brut (SM1V1)	- 1,78	- 1,58	0,15	0,55
	Normalisé (SM1V2)	- 1,50	- 1,18	- 0,12	0,47
Imputation économétrique (SM2)		- 1,56	- 0,81	0,05	0,39
Score PLS (SM3)		- 1,55	- 0,50	- 0,03	0,34
Scores MRI	Estimation séparée (SM4V1)	- 1,70	- 1,38	0,16	0,50
	Estimation concourante (SM4V2)	- 1,49	- 1,18	- 0,05	0,46

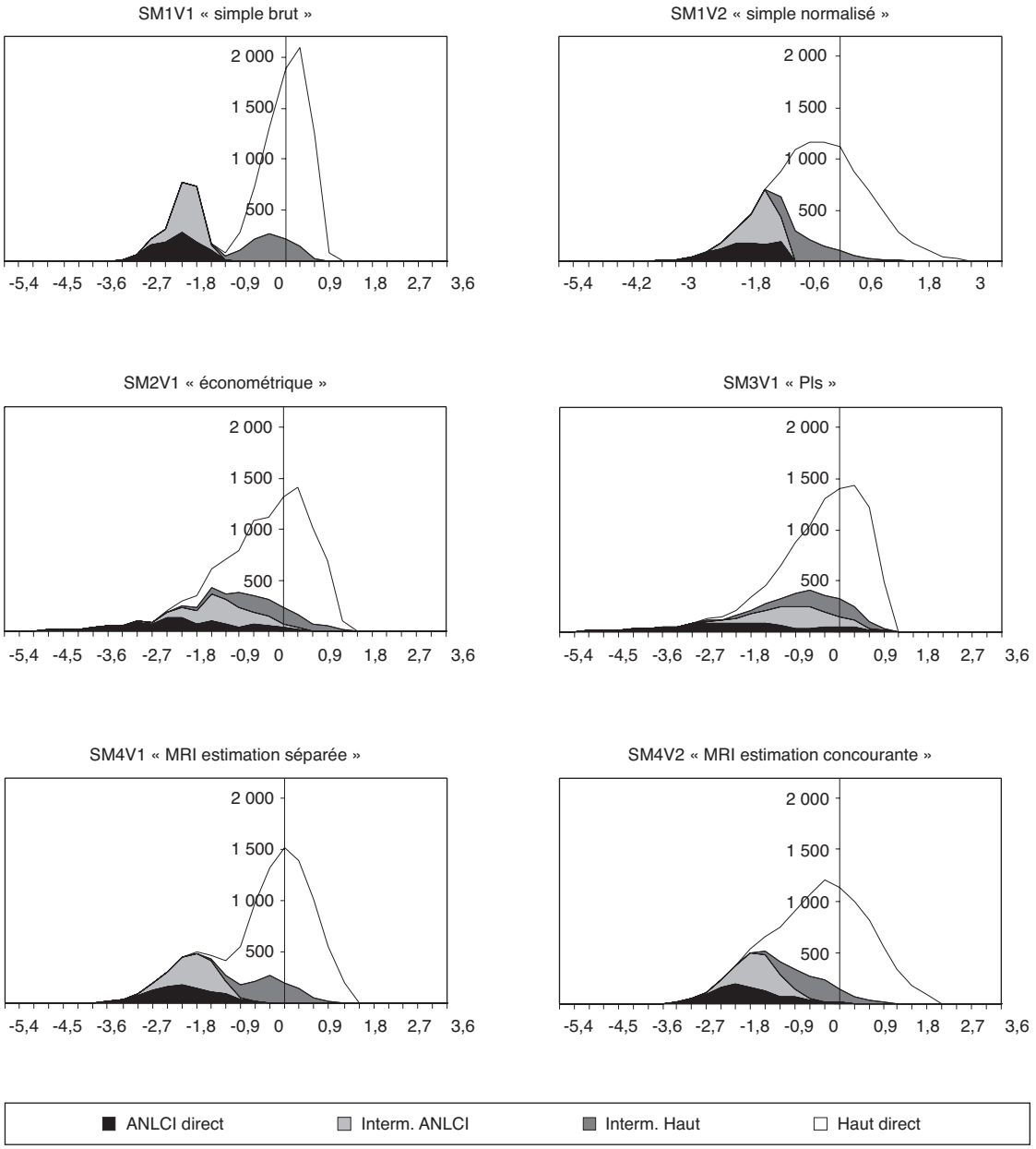
Lecture : comme la moyenne de chaque score a été fixée à 0 et l'écart-type à 1, le groupe « ANCLI direct » se trouve à 1,78 écart-type en dessous de la moyenne pour le score SM1V1.
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

et le score PLS, des deux cinquièmes pour le score MRI utilisant l'estimation concourante.

Les premières analyses sur *IVQ* ont montré que, quel que soit le domaine, les résultats décroissent avec l'âge (Micheaux et Murat, 2006). Les scores donnent des résultats concordants sur ce point, mais là encore avec une ampleur variable pour caractériser ce phénomène : le score « simple brut » et le score PLS donnent une nouvelle fois les corrélations les plus faibles (autour de 6,5 % pour le coefficient de détermination), tandis que l'on dépasse 8 % pour l'estimation MRI concourante.

Les résultats à l'écrit sont différents selon le sexe des personnes évaluées (Djider et Murat, 2006). Les femmes ont mieux réussi le module d'orientation que les hommes : elles sont donc moins souvent orientées vers le module ANLCI et considérées comme en difficulté à l'écrit ; en revanche, plus nombreuses à passer le module

Graphique V
Distribution des différents scores calculés



Lecture : pour chaque score, on a calculé le nombre d'individus par tranche de 0,3 écart-type, en distinguant les 4 groupes. Les graphiques présentent la distribution de ces effectifs.
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

Haut, elles y obtiennent de moins bons résultats. Cette inversion peut trouver son origine dans le type d'épreuve proposée : en effet, le rapport s'inverse selon les supports et les procédures évaluées (les femmes sont distancées sur les cartes et les graphiques statistiques ; elles font jeu égal avec les hommes quand les questions portent sur un texte d'une page). Il se peut aussi que la population masculine soit plus dispersée, à la fois plus souvent en difficulté et plus performante sur les exercices complexes. Les scores vont être sensibles à ce phénomène. Ainsi, le score « simple brut », qui suit de façon scrupuleuse le processus d'orientation en distinguant nettement ceux orientés vers le module ANLCI et ceux orientés vers le module Haut, accorde un poids fort aux résultats dans le module d'orientation et va donc « favoriser » les femmes, qui y obtiennent de bons résultats : l'écart qui les sépare des hommes est significatif au seuil de 1 %. En revanche, pour les autres scores (à l'exception de l'estimation MRI séparée, d'ailleurs proche du score « simple brut »), l'écart est bien moins net et non significatif, même pour le score « simple normalisé ».

À l'issue de ces analyses nos préférences se portent sur deux scores. Le score issu du MRI avec l'estimation concourante présente une distribution régulière et est le plus corrélé avec le diplôme. La variante normalisée de la première méthode est aussi assez bien corrélée avec le diplôme et présente l'avantage d'être d'une conception et d'un calcul plus aisés.

Comparaison des méthodes sur données fictives

Cependant, les écarts entre les scores ne sont pas suffisamment grands pour pouvoir trancher facilement. Des simulations ont donc été réalisées pour tester la robustesse des méthodes. Elles se fondent sur une situation où l'ensemble de l'information est disponible, sans aucun processus de sélection. L'analyse classique et les modèles de réponse à l'item sur ces données complètes vont donner deux synthèses, assez proches, que l'on considèrera comme des références. Ensuite, on introduira des « trous » dans les données, de façon similaire à *IVQ*, pour mettre en œuvre les différentes méthodes qui viennent d'être présentées. La confrontation entre les scores ainsi obtenus et les scores de référence permettra de mieux comprendre les limites de chaque méthode.

On peut effectuer deux types de simulations :

- Des simulations sur données fictives : on construit complètement la matrice des réponses à partir de quelques hypothèses. L'intérêt de cette méthode est de permettre des variantes pour les différents paramètres : taille et difficulté des différentes épreuves, hypothèses sur la distribution des compétences, modification du processus d'orientation, etc. L'inconvénient est le caractère artificiel des données et le risque que l'on valide une méthode proche du mode de construction de la matrice.

- Des simulations sur données réelles : on part d'un jeu de données existant, de préférence assez proches de la structure des réponses à l'enquête *IVQ*. Il est bien sûr plus satisfaisant de partir de données réelles, mais les marges de manœuvre sont aussi moins grandes. Les résultats peuvent être sensibles à la longueur et la difficulté des épreuves, au type de population, au domaine étudié, etc. Il faut de plus trouver une base de données contenant un nombre important d'items

Tableau 8
Corrélation des différents scores avec le diplôme, l'âge et le sexe

		Lien avec le niveau d'études (R^2) en %	Lien avec l'âge (R^2) en %	Lien avec le sexe		
				Score des hommes	Scores des femmes	Test
Imputation simple	Brut (SM1V1)	33,1	6,3	- 0,06	0,04	***
	Normalisé (SM1V2)	38,1	7,7	- 0,01	0,01	n.s.
Imputation économétrique (SM2)		36,0	8,1	- 0,01	0,01	n.s.
Score PLS (SM3)		32,0	6,5	- 0,01	0,00	n.s.
Scores MRI	Estimation séparée (SM4V1)	36,7	7,3	- 0,03	0,02	***
	Estimation concourante (SM4V2)	39,2	8,1	- 0,01	0,01	n.s.

Lecture : l'analyse de variance du score SM1V1 par le niveau d'études en 8 postes permet de rendre compte de 33,1 % de la variance de ce score. La part de variance expliquée est 6,3 % quand on ne prend en compte que l'âge. Toujours pour ce score, les hommes ont un score inférieur de 0,06 écart-type à la moyenne (et se distingue significativement des femmes).
Champ : personnes âgées de 18 à 65 ans vivant en ménage ordinaire, en France métropolitaine.
Source : enquête Information et Vie Quotidienne *2004, Insee.*

pour simuler des données relativement proches de celles d'*IVQ*. Compte tenu de ces difficultés, ce type de simulation sera développé dans un travail ultérieur.

Dans un premier temps, les données fictives vont être construites de façon à être le plus proche possible de celles d'*IVQ*. On part des résultats du MRI avec estimation concourante, qui semble le score le plus pertinent, si on en juge par la corrélation avec le niveau d'études et qui permet facilement la construction de données fictives. Pour chaque individu, on garde le score qui a été estimé (*SBASE* : « score de base ») ; pour chaque item, on reprend les paramètres de difficulté et de discrimination (c'est-à-dire la sensibilité de l'item à une variation du niveau de compétence, cf. encadré 4). Avec ces valeurs, il est possible de calculer la probabilité p_{ij} d'un individu *i* de réussir l'item *j*, en utilisant la formule définissant le modèle. À l'aide de cette probabilité, on va construire une matrice complète de réponses : pour chaque couple (*i*,*j*), on tire au sort de façon uniforme un nombre entre 0 et 1. Si ce nombre est inférieur à p_{ij}, on donne la valeur 1, sinon, on donne la valeur 0.

Une fois, les réponses simulées, on a calculé sur l'ensemble de la matrice :

- *STOT* ou « score classique de référence » : nombre de bonnes réponses ;

- *SMRITOT* ou « score MRI de référence » : score issu d'un modèle de réponse à l'item à deux paramètres.

Les deux scores sont très fortement liés entre eux (cf. tableau 9). La corrélation est moins nette avec le score de base qui a servi à construire la matrice, du fait des facteurs aléatoires introduits lors de la construction des indicatrices de réussite par item. Les données ont ensuite été tronquées selon un processus proche d'*IVQ*. Un score a été calculé sur les dix premiers items, qui a distingué trois groupes de tailles approximativement équivalentes à celles observées sur *IVQ* pour les personnes orientées vers le module ANLCI, le module Intermédiaire et le module Haut. Ensuite, le score sur les 11 items suivant permet de distinguer parmi les personnes du groupe médian celles qui doivent « passer » le Haut et celles qui doivent « passer » le module ANLCI. Les quatre cas ainsi définis vont servir à « creuser » la matrice : par exemple, pour le groupe 1, « ANLCI direct », on met à blanc les 21 derniers items ; de même pour le groupe 2, « Intermédiaire ANLCI » ; pour le groupe 4, « Haut direct », ce sont les items de 11 à 21 qui sont mis à blanc.

Tableau 9
Corrélation entre les scores de référence

	SBASE	STOT	SMRITOT
Score de base	1		
Score classique de référence	0,92	1	
Score MRI de référence	0,94	0,98	1

Source : données simulées, Insee-DEPP.

Une cohérence d'ensemble assez bonne

Les différentes méthodes présentées *supra* sont alors appliquées aux données pour obtenir les scores *SM1V1*, *SM1V2*, etc. Les différents scores sont assez bien corrélés avec ceux calculés sur l'ensemble des données (cf. tableau 10). Ce

Tableau 10
Corrélation entre les scores calculés avec les « trous » et les scores de référence

		Score de base	Score classique de référence	Score MRI de référence
Scores bruts	SMO	0,71	0,80	0,76
	SMA (1)	0,77	0,85	0,83
	SMH (2)	0,84	0,96	0,94
Imputation simple	Brut (SM1V1)	0,86	0,93	0,91
	Normalisé (SM1V2)	0,90	0,94	0,96
Imputation économétrique (SM2)		0,90	0,96	0,96
Score PLS (SM3)		0,87	0,92	0,92
Scores MRI	Estimation séparée (SM4V1)	0,91	0,97	0,97
	Estimation concourante (SM4V2)	0,92	0,96	0,98

1. *sur les individus ayant passé le module MA.*
2. *sur les individus ayant passé le module MH.*

Source : données simulées, Insee-DEPP.

sont les scores issus des MRI qui sont les plus proches, mais cela tient peut-être à l'utilisation d'un tel type de modèle pour la construction des données.

Malgré l'ampleur des corrélations, les différentes méthodes ne sont pas exemptes de biais (cf. tableau 11). Les scores moyens en fonction du déroulement de l'épreuve diffèrent selon les modèles. Les scores de référence présentent des écarts plus grands que le score de base ayant servi à la construction. En effet, la division en sous-populations est fondée sur des sous-scores de *STOT*, score classique de référence. Il est donc normal que les différences soient très tranchées selon ce score. Entre le score classique de référence et le score de base, il y a l'introduction de facteurs aléatoires, pour simuler l'erreur de mesure propre à toute évaluation, qui rendent donc la liaison moins nette pour le score de base. Ce phénomène est amplifié par le caractère facile de l'épreuve : de forme dissymétrique, le score classique de référence tend à accroître les écarts au bas de la distribution. Le score issu des modèles de réponse à l'item donne des écarts plus faibles entre les groupes, car, de forme logistique, il aboutit à une distribution symétrique plus proche de celle du score originel.

Les méthodes appliquées sur les données incomplètes donnent généralement des écarts plus grands que sur données complètes : les compétences des individus du groupe « ANLCI direct » sont sous-estimées et celles des individus du groupe « Haut direct » surestimées. La situation des individus des groupes intermédiaires est variable selon les scores : par exemple, les individus du groupe « Intermédiaire ANLCI » semblent un peu sous-estimés dans la méthode « simple normalisée » et surestimés dans les méthodes économétrique et PLS.

Sur les données complètes, il est possible de se faire une idée de la pertinence des hypothèses, assez fortes, qui sous-tendent les imputations simples. Ainsi, dans les données simulées complètes, les individus des groupes ANLCI réussissent presque 7 items sur 21 dans le module Haut. Or l'imputation simple se fonde sur le principe de leur attribuer un score de 0 sur 21. Inversement, d'après cette simulation, les individus passant directement le module Haut, réussissent « seulement » 9,7 items sur 11 du module ANLCI.

Les simulations permettent aussi de mieux comprendre les biais qui affectent la méthode économétrique. La reconstruction du score *SMA* pour les personnes orientées vers le module Haut est plutôt bonne (cf. graphique VI). En effet, le score *SMA* prédit par la régression pour les valeurs 9 et 10, pour lesquelles il n'est pas calculable sur les données incomplètes, est très pro-

Tableau 11
Scores moyens des 4 groupes selon les différents scores calculés

A- Scores de référence

Groupe	Score de base	Score classique de référence	Score MRI2 de référence
ANLCI direct	- 1,63	- 1,91	- 1,71
Intermédiaire ANLCI	- 1,00	- 1,07	- 1,06
Intermédiaire Haut	- 0,12	- 0,05	- 0,14
Haut direct	0,46	0,50	0,48
R^2 (en %)	51,2	64,7	56,9

B- Scores calculé avec les « trous »

Groupe	Scores simples		Score économétrique	PLS	MRI	
	Brut	Normal.			2 par séparée	2 par concourante
ANLCI direct	- 1,92	- 1,68	- 2,02	- 2,05	- 1,79	- 1,73
Intermédiaire ANLCI	- 1,58	- 1,10	- 0,83	- 0,43	- 1,23	- 1,20
Intermédiaire Haut	0,04	-0,29	- 0,20	- 0,15	- 0,19	- 0,20
Haut direct	0,58	0,52	0,49	0,41	0,54	0,53
R^2 (en %)	88,1	60,3	62,8	52,5	68,1	64,8

Lecture : comme la moyenne de chaque score a été fixée à 0 et l'écart-type à 1, la sous-population 1 se trouve à - 1,63 écart-type en dessous de la moyenne pour le score de base. La variable distinguant les 4 groupes explique 51,2 % de la variance du score de base.
Source : données simulées, Insee-DEPP.

che du score moyen observé sur données complètes. En revanche, la régression entre le score *SMH* et *SMO*, obtenue sur données incomplètes uniquement avec les groupes « Intermédiaire haut » et « Haut direct », est assez mal estimée. En effet, comme il a été expliqué plus haut, le score moyen *SMH* associé aux valeurs intermédiaires du score d'orientation (7 et 8 sur 10) est surestimé sur données incomplètes, par la mise à l'écart des individus du groupe « Intermédiaire ANLCI ». Dans cette méthode sur données incomplètes, cette population va se voir attribuer un score moyen de 11,1 (pour ceux qui ont eu 7 sur 10 au module d'orientation) ou 12,6 (pour ceux qui ont eu 8 sur 10), alors que les « vraies » valeurs, sur données complètes, sont respectivement 7,4 et 8,7. Cette sous-population va donc être surévaluée par cette méthode.

Des simulations pour tester des améliorations possibles de la structure de l'épreuve

Cette première simulation a permis de comparer, sur des données proches de celles d'*IVQ*, les différentes méthodes et de montrer que l'estimation « concourante » du MRI à 2 paramètres semblait la plus efficace. Un exercice plus ouvert de simulation va chercher à étudier dans quelle mesure on pourrait modifier la structure du protocole pour améliorer la mesure finale. Pour cela, nous allons, sur une population fictive de 10 000 individus dont la compétence suit une loi normale, construire une série d'épreuves fictives qui vont se différencier selon :

- la taille du module d'orientation (le nombre d'items),

- la difficulté de ce module,

- la proportion d'individus orientés vers les différents modules à l'issue de ce module d'orientation,

- la difficulté du module Bas,

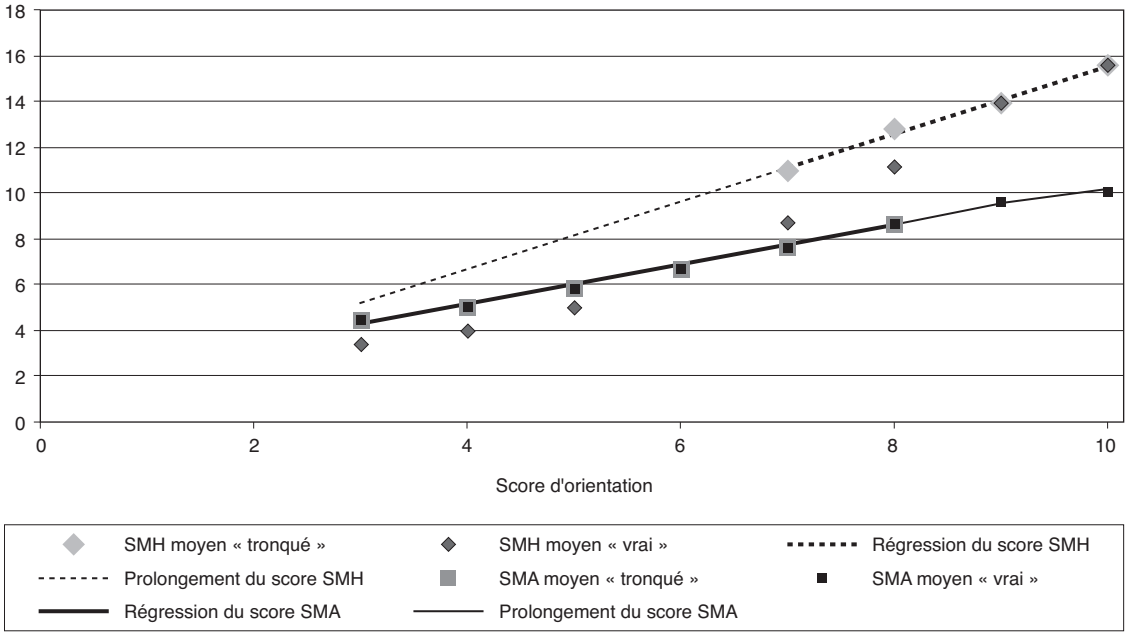
- la difficulté du module Intermédiaire,

- la difficulté du module Haut.

La matrice des réponses est construite de la façon suivante :

- Pour l'exercice d'orientation, le nombre d'items (*Nori*) et la difficulté moyenne (*Bom*) sont les paramètres de la simulation. La disper-

Graphique VI
Régressions des scores aux modules ANLCI (SMA) et Haut (SMH) sur le score au module d'Orientation (SMO)



Lecture : les carrés et losanges indiquent les scores moyens observés pour chaque valeur de SMO respectivement pour SMA et SMH. Les gros carrés et losanges indiquent les scores moyens sur données incomplètes, tandis que les petits indiquent les scores sur données complètes. Les droites de régression ont aussi été indiquées sur les graphiques (le trait est plus fin quand la droite est prolongée sur un champ où elle n'a pas été calculée). Par commodité, on a présenté les deux régressions sur le même graphique, bien que SMA et SMH soient sur des échelles distinctes.
Source : données simulées, Insee-DEPP.

sion des difficultés des items (*Bos*) a été fixée au même niveau pour l'ensemble des simulations. De même, les paramètres de discrimination ont une moyenne (*Aom*) et une dispersion (*Aos*) identiques dans toutes les simulations. Une fois ces valeurs fixées (18), on va tirer autant de paramètres de difficultés et de discriminations qu'il y a d'items, dans une loi normale de caractéristiques (*Bom*, *Bos*) pour les difficultés et une loi log-normale de caractéristiques (*Aom, Aos*) pour les discriminations. Pour chaque individu, on simule, comme précédemment, la réussite ou l'échec à l'item en fonction de sa compétence et des caractéristiques de l'item.

- Il est alors possible de construire un score d'orientation et une procédure respectant la contrainte concernant les proportions d'individus passant les différents modules.

- Pour les individus passant le module Bas, on va procéder comme pour l'exercice d'orientation. La taille de l'épreuve est fixée en supposant que ces individus, fragiles dans ce domaine, ne peuvent répondre à une épreuve trop longue. En ajoutant les items de l'orientation, le nombre total d'items proposés est fixé à 35.

- Il en va de même pour les modules Intermédiaire et Haut : le nombre total d'items est fixé à 40 pour les individus passant le module Intermédiaire et 50 pour les individus passant le module Haut. Notons que le module Intermédiaire est ici un module à part entière et non un repêchage de l'orientation comme dans le cas d'*IVQ*.

Pour chacun de ces facteurs, nous avons fixé cinq valeurs, sauf pour les seuils d'orientation, pour lesquels il n'y a que trois scénarios distincts :

- la difficulté moyenne de l'épreuve d'orientation prend les valeurs (- 2 ; - 1,5 ; - 1 ; - 0,5 ; 0) ;

- la difficulté moyenne du module Bas prend les valeurs (- 3 ; - 2,5 ; - 2 ; - 1,5 ; - 1) ;

- la difficulté moyenne du module Intermédiaire prend les valeurs (-2 ; 1,5 ; - 1 ; - 0,5 ; 0) ;

- la difficulté moyenne du module Haut prend les valeurs (- 0,5 ; 0 ; 0,5 ; 1 ; 1,5) ;

- le nombre d'items du module d'orientation (ou taille du module) prend les valeurs (5, 10, 15, 20, 25) ;

- les proportions d'individus orientés vers les modules Bas et Intermédiaire prennent les valeurs (10 %, 20 %), (15 %, 30 %) et (20 %, 40 %).

Des épreuves pour toutes les combinaisons de ces différentes valeurs ont été construites. Il est alors possible pour chacune d'entre elles, de mettre en œuvre les techniques présentées précédemment sur ces différentes simulations. Pour simplifier, nous n'avons conservé que l'estimation séparée du MRI à 2 paramètres. La qualité du résultat (qui juge maintenant celle de l'épreuve et non plus de la méthode) est mesurée par 2 critères :

- l'erreur d'adéquation globale, qui se rapproche de la distance entre le vecteur des compétences et celui des scores estimés :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\theta_i - \hat{\theta}_i)^2} \quad (19) ;$$

- la proportion d'individus mal classés par le score estimé par rapport à ce que donne la compétence, en fixant un critère de repérage des individus en difficulté (avoir une compétence ou un score estimé inférieur à un certain seuil).

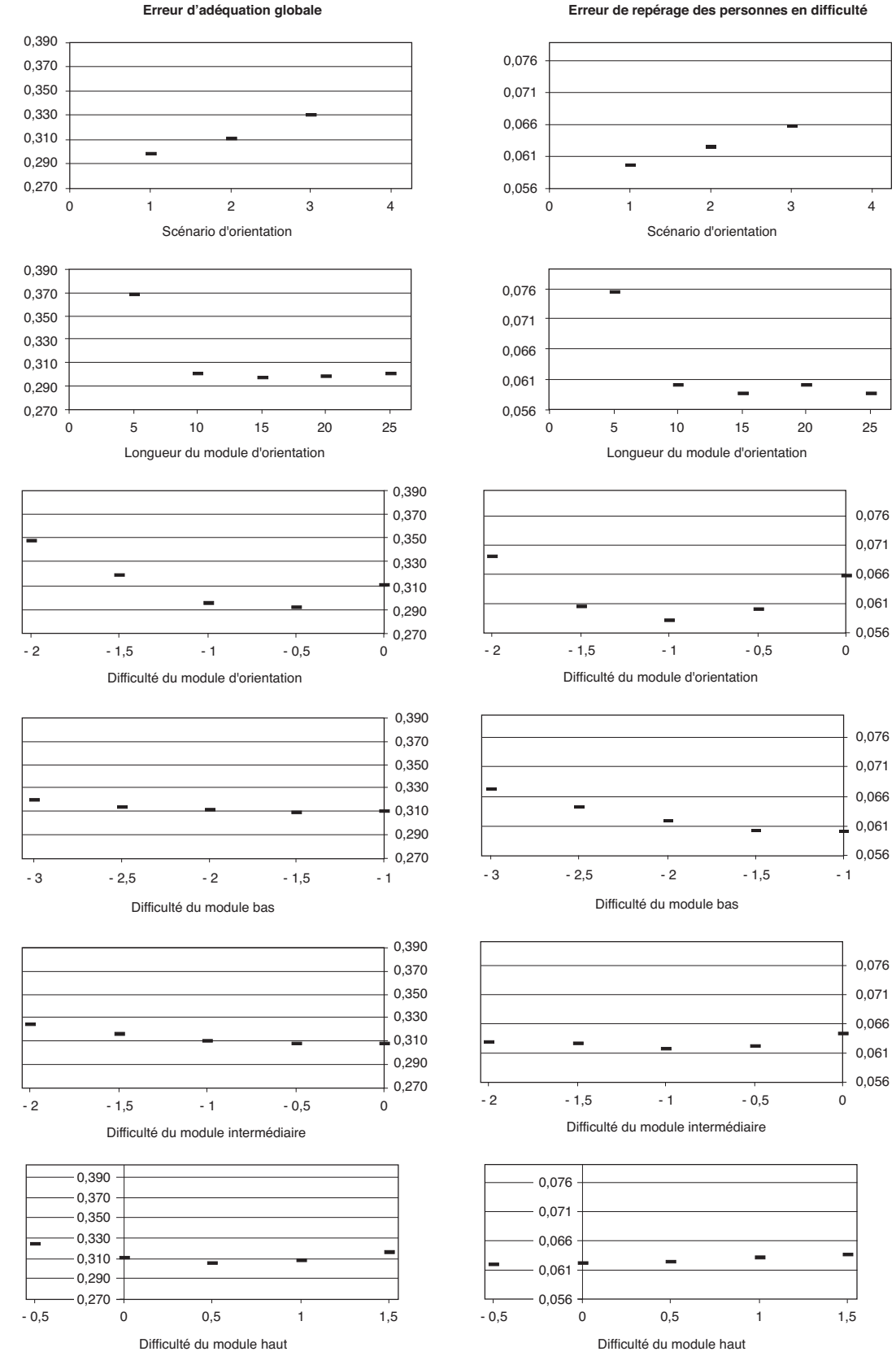
Pour les deux critères, le premier scénario d'orientation, celui qui cible davantage, donne des résultats sensiblement meilleurs (cf. graphique VII). Concernant la longueur de l'épreuve d'orientation, si des épreuves très courtes (5 items) sont à proscrire, il n'apparaît pas de différences très nettes entre celles qui ont 10 items et celles qui en ont 25. Le gain de précision pour les épreuves d'orientation longues est compensé par la diminution de la taille des épreuves suivantes. Une épreuve d'orientation plutôt facile paraît souhaitable : l'erreur d'adéquation globale est minimale, autour de - 0,5. En effet, dans le cas présent, la procédure d'orientation cible les personnes en difficulté ; une orientation ciblant l'élite inverserait le signe. Compte tenu du fait que la population a une compétence moyenne de 0, cela signifie qu'elle doit réussir environ 66 % des items de l'orientation. Pour le repérage des personnes en difficulté, le minimum se situe plutôt autour de - 1, qui est d'ailleurs le seuil utilisé pour classer les personnes comme en difficulté, ce qui correspond à un taux de réussite de près de 80 %.

---

18. Il est bien sûr envisageable de faire varier aussi Bos, Aom et Aos, mais cela multiplie les simulations et les essais que nous avons faits à ce sujet n'ont pas montré une influence très forte de ces paramètres.

19. La corrélation entre le score estimé et la compétence du sujet a aussi été utilisée et donne des résultats très proches.

Graphique VII
Impact des caractéristiques des épreuves sur la qualité de la mesure



Lecture : pour les simulations relevant du scénario 1 d'orientation (10 % vers le module bas et 20 % vers le module intermédiaire), l'erreur d'adéquation globale (distance entre le score estimé et la compétence) prend la valeur 0,298. Dans ces simulations, lors du repérage des individus en difficulté (compétence inférieure à - 1), 6 % des individus sont mal classés par le score estimé par rapport au classement obtenu avec la compétence.
Source : données simulées, Insee-DEPP.

La difficulté du module Bas a relativement peu d'impact sur l'adéquation globale. Il semble cependant préférable que cette épreuve ne soit pas trop facile. En revanche, pour le repérage des personnes en difficulté, un niveau de difficulté entre - 1,5 et - 1, donc seulement légèrement inférieur à l'épreuve d'orientation, semble souhaitable. Cependant, l'utilisation d'un critère cherchant à mesurer si à l'intérieur de cette population le niveau est bien estimé, pour bien distinguer les personnes très en difficulté de celles aux difficultés légères, nuancerait ce résultat et favoriserait des épreuves plus faciles.

Pour le module Intermédiaire, un niveau de difficulté légèrement supérieur à celui de l'orientation (autour de 0) semble préférable pour assurer l'adéquation globale, mais, pour améliorer le repérage des personnes en difficulté, l'optimum se situe plutôt autour de - 1. Il en va de même pour le module Haut. Privilégier le repérage des personnes en difficulté pousserait à se fixer une épreuve plutôt facile, alors que l'analyse de l'adéquation globale donne plutôt une valeur de 0,5 (correspondant à un taux de réussite de 33 % environ sur l'ensemble de la population).

Une analyse plus fine, pour des combinaisons particulières de ces différents paramètres, permet de préciser certains résultats. Ainsi, l'écart entre les trois scénarios est plus fort pour les épreuves d'orientation très faciles, en faveur de celui ciblant les personnes en difficulté. En revanche, le rapport s'inverse pour les épreuves d'orientation assez difficile : pour une difficulté de 0 (soit un taux de réussite de 50 % pour l'ensemble de la population), le troisième scénario paraît alors préférable aux deux autres.

* *
*

Ces simulations donnent une première image de la fiabilité et des limites de la procédure d'orientation utilisée dans *IVQ* et des traitements psychométriques qu'elle implique. Dans l'ensemble, la perte d'information ne semble pas trop importante : les scores sur les données incomplètes sont assez bien corrélés avec ceux sur données complètes. Cependant, la corrélation n'est pas parfaite et la procédure d'orientation provoque des biais pas tout à fait négligeables. Ces biais sont de sens et d'ampleur variables selon la méthode utilisée, mais ils affectent le plus souvent le groupe « Intermédiaire ANLCI ». Dans cette population, on trouve des individus « tangents » : leurs résultats aux modules d'orientation et intermédiaire, insuffisants sans être très mauvais, ont justifié le passage du module ANLCI, mais la passation complémentaire du module Haut aurait sans doute permis de mieux estimer leurs compétences. Par contraste, les sous-populations « ANLCI direct » et « Haut direct » semblent peu affectées par la procédure d'orientation, même si le niveau de compétences de la sous-population « Haut direct » paraît légèrement surestimé sur données incomplètes par la plupart des méthodes.

D'autre part, les simulations faisant varier les paramètres d'items donnent aussi des pistes d'amélioration. Elles confortent les choix faits pour la taille du module d'orientation (environ 10 items) et la procédure d'orientation (10 % vers le module Bas et 20 % vers le module Intermédiaire). En revanche, elles suggèrent que des épreuves plus difficiles seraient sans doute souhaitables, en particulier pour le module d'orientation. En effet, la valeur optimale de la difficulté pour ce module se situe autour de - 0,5 (soit un taux de réussite de 66 % pour l'ensemble de la population), alors que pour les répondants à l'enquête *IVQ*, la difficulté des items du module d'orientation est inférieure à - 2. □

BIBLIOGRAPHIE

Accardo J. et de Saint Pol T. (2009), « Qu'est-ce qu'être pauvre aujourd'hui en Europe ? L'analyse du consensus sur les privations », *Économie et Statistique*, n° 421, pp. 3-27.

Bernier J.-J. et Pietrulewicz B. (1997), *La psychométrie. Traité de mesure appliqué*, Gaëtan Morin éditeur, Montréal.

Blais J.-G. et Laurier M. (1997), « La détermination de l'unidimensionnalité de l'ensemble des scores à un test », *Mesure et évaluation en éducation*, vol. 20, n° 1, pp. 65-90.

Bonnet G., Braxmeyer N., Horner S., Lappalainen H.P., Levasseur J., Nardi E., Rémond M., Vrignaud P. et White J. (2001), *The Use of National Reading Tests for International*

Comparisons: Ways of Overcoming Cultural Bias, MEN-DPD.

Bock R.D. et Zimowski M.F. (1998), *Feasibility Studies of Two-Stage Testing in Large-Scale Educational Assessment: Implications for NAEP*, American Institute of Research, NAEP Validity Studies.

du Toit M. (éd.) (2003), *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*, SSI, Lincolnwood.

Dickes P., Tournois J., Flieller A. et Kop J.-L. (1994), *La psychométrie. Théorie et pratique de la mesure en psychologie*, PUF, Paris.

Djider Z. et Murat F. (2006), « Des chiffres pour les hommes… … des lettres pour les femmes », *Insee Première*, n° 1071.

d'Haultefœuille X., Murat F. et Rocher T. (2002), « La mesure des compétences : les logiques contradictoires des évaluations internationales », *Actes des Journées de méthodologie statistique 2000*, Insee.

Gould S.J. (1987), *La mal-mesure de l'homme*, Éditions Odile Jacob, Paris.

Guillevic C. et Vautier S. (1998), *Diagnostic et tests psychologiques*, Nathan Université, Paris.

Hanson B.A. et Béguin A.A. (2002), « Obtaining a Common Scale for Item Response Theory Parameters Using Separate Versus Concurrent Estimation in the Common-Item Equating Design », *Applied Psychological Measurement*, vol. 26, n° 1, pp. 3-24.

Hanson B., Zeng L. et Cui Z. (2004), *ST: A Computer Program for IRT Scale Transformation*.

Huteau M. et Lautrey J. (1999), *Évaluer l'intelligence, Pyschométrie cognitive*, PUF, Paris.

Juhel J. (éd.) (1999), *Le modèle de la réponse à l'item*, numéro spécial de Psychologie et Psychométrie, vol 20, n °2-3.

Lord F.M. (1980), *Application of Item Response Theory to Practical Testing Problems*, Lawrence Erlbaum Associates, Hillsdale, NJ.

Merle P. (1996), *L'évaluation des élèves. Enquête sur le jugement professoral*, PUF, Paris.

Micheaux S. et Murat F. (2006), « Les compétences à l'écrit, en calcul et en compréhension orale selon l'âge », *Données Sociales 2006*, pp. 195-202, Insee.

Mislevy R. J. (1987), « Exploiting Auxiliary Information about Examinees in the Estimation of Item Parameters », *Applied Psychological Measurement*, vol. 11, n° 1, pp. 81-91.

OCDE (2005), *PISA 2003* Technical Report.

Rasch G. (1960), *Probabilistic models for some intelligence and attainment test*, Nielsen & Lydiche, Copenhagen.

Rocher T. (2003), « La méthodologie des évaluations internationales de compétences », *Psychologie et Psychométrie*, vol. 24, n° 2/3, Éditions EAP, pp. 117-146.

Rocher T. (2004), « Les évaluations en lecture dans le cadre de la journée d'appel de préparation à la défense. Année 2003 », *Note d'évaluation*, n° 04.07, ministère de l'Éducation nationale, de l'Enseignement supérieur et de la Recherche, direction de l'Évaluation et de la Prospective.

Stocking M.L. et Lord F.M. (1983), « Developing a Common Metric in Item Rsponse Theory », *Applied Psychological Measurement*, vol. 7, n° 2, pp. 201-210.

Stout W. (1990), « A New Item Response Theory Modeling Approach with Applications to Unidimensionality Assessment and Ability Estimation », *Psychometrika*, vol. 55, n° 2, pp. 293-325.

Tenenhaus M. (1998), *La régression PLS - Théorie et pratique*, Éditions Technip, Paris.

Umetri AB (1998), *SIMCA 7.0, Graphical Software for Multivariate Modeling*, Umetri AB, Suède.

Vallet L.-A., Bonnet G., Emin J.-C., Levasseur J., Rocher T., Blum A., Guérin-Pace F., Vrignaud P., d'Haultfœuille X., Murat F., Verger D. et Zamora P. (2002), « Enquête Méthodologique Information et Vie Quotidienne - Tome 1 : bilan du test 1, novembre 2002 », *Document de travail, série Méthodologie statistique*, n° C0202, Insee.

Wainer H. (éd.) (2000), *Computerized Adaptive Testing: A Primer (Second Edition)*, Lawrence Erlbaum Associates, Mahwah, NJ.