LES PERFORMANCES D'ADULTES A DES TESTS EN LECTURE : COMMENT SÉPARER MOTIVATION ET COMPÉTENCES ?

Fabrice MURAT [(*)], *Philippe ZAMORA*[(**)]

[(*)] *Insee, Département de l'emploi et des revenus d'activité*
[(**)] *Dares*

Introduction

L'enquête Information et Vie Quotidienne (IVQ) s'attache à évaluer les compétences d'adultes dans le cadre d'une enquête-ménage. Les évaluations de compétences scolaires sont devenues un des outils les plus utiles de la recherche en sociologie de l'éducation mais l'expérience est bien moins développée en ce qui concerne les adultes. La psychologie cognitive a le plus souvent recours à des tests en laboratoire, dans un cadre bien particulier. Est-il possible de transposer sans risque les épreuves utilisées dans ce cadre à celui d'une enquête-ménage classique, disposant d'un échantillon important, afin de donner lieu à des analyses sociologiques ? La question se pose moins pour les évaluations en milieu scolaire car elles s'inscrivent facilement dans le contexte de la classe et recueillent sans peine l'adhésion des élèves, habitués à faire des exercices. En revanche, il peut sembler beaucoup plus délicat de proposer des exercices d'évaluation de la lecture ou du calcul à des personnes chez elles, alors qu'elles ont quitté le système éducatif depuis parfois très longtemps.

Cette préoccupation est renforcée par le bilan de la seule opération d'envergure ayant eu les mêmes objectifs, l'enquête *International Adult Literacy Survey* (IALS) menée par *Statistic Canada* et *Educational Testing Services* (ETS) dont les résultats ont été diffusés par l'Organisation de Coopération pour le Développement Economique (OCDE) ([11]). Les résultats étaient pour la France très surprenants : 40 % des français entraient dans la catégorie des plus mauvais lecteurs (vite assimilée à celle des illettrés) et la France se trouvait bien loin derrière la plupart des pays participants (en autres, l'Allemagne, la Suède, les Etats-Unis). Des investigations ont été menées (Dickes et Vrignaud [12], Kalton et alii [7]), mettant en évidence un certain nombre de problèmes méthodologiques qui ont justifié le retrait de la France de l'opération et la non-diffusion officielle des résultats.

L'enquête IVQ s'inscrit donc dans un contexte assez particulier. L'évaluation des compétences des adultes demeure un objectif important : du fait de la complexification du monde du travail et de la vie quotidienne, il est intéressant de savoir si tout le monde maîtrise les bases de la lecture et du calcul. Le précédent de IALS et les études menées sur ce cas incitent à une certaine prudence mais donnent aussi bien des pistes pour obtenir une évaluation de meilleure qualité. Le développement de l'enquête IVQ s'est fait en gardant à l'esprit les difficultés rencontrées précédemment. C'est pourquoi un effort particulier a été fait pour construire des exercices pertinents, élaborer un protocole d'enquête fiable.

Le présent article va porter sur les deux tests de l'opération qui se sont déroulés à Limoges et Nancy en décembre 2000 (test 1) et à Lyon, Nancy et Poitiers, en avril 2002 (test 2). Ces opérations diffèrent surtout entre elles et de l'enquête finale qui a eu lieu en novembre 2002, par le contenu des exercices ; l'architecture est par contre identique. Cet article comporte trois parties : dans un premier temps, le contexte et le dispositif d'enquête seront présentés rapidement ; ensuite, viendront des exemples d'analyses sur les épreuves, qui ont guidé les choix pour l'enquête finale ; enfin, nous présenterons un travail sur la prise en compte du temps passé à répondre aux questions et des observations des enquêteurs comme mesure de la motivation, ce qui constitue l'une des originalités de l'enquête.

Les objectifs et l'organisation d'IVQ : le test 2

Dans cette partie, on présentera successivement le contexte de l'enquête en particulier les travaux qui ont été menés sur l'enquête IALS et le dispositif IVQ. On en trouvera un exposé plus détaillé dans le document de travail [6], consacré à l'exploitation du test 1.

Le contexte de l'enquête

Evaluer les compétences des adultes

L'intérêt d'évaluer les compétences des adultes est multiple : il vient des mutations du monde du travail et de la vie quotidienne, qui nécessitent un accès bien maîtrisé à une information de plus en plus complexe. Les évaluations des compétences peuvent être un outil d'analyse très utile pour mieux comprendre le lien entre le diplôme et l'emploi. Un des objectifs importants de l'enquête est aussi le repérage des personnes en situation d'illettrisme. Il est apparu rapidement que cette population devait faire l'objet d'un questionnement particulier non seulement parce qu'un test classique était pour eux inadapté mais aussi parce que l'on souhaitait avoir une vision fine de ce qu'ils savent faire. Il est probable aussi que cette population recouvre une grande diversité de cas, en terme de compétences (de l'analphabète à l'immigrant de fraîche date ne maîtrisant pas encore le français) et de parcours scolaire, familial et professionnel.

L'enquête IALS

En 1994, Statistique Canada et ETS (Educational Testing Service) ont organisé l'enquête internationale IALS sur les compétences en littératie des adultes, dont les résultats ont ensuite été largement diffusés par l'OCDE. Comme on l'a dit, les résultats français ont suscité de fortes interrogations : le classement international de la France était très mauvais et les journaux ont repris des fuites alléguant que 40 % des français seraient illettrés[1]. Un certain nombre de problèmes méthodologiques sont apparus lors des expertises, justifiant le retrait de la France de cette enquête, problèmes que l'on va présenter rapidement, avec les enseignements qu'ils ont pu apporter.

[1] Il importe de signaler que la définition de l'OCDE ne correspond que partiellement à la notion d'illettrisme. Les personnes se trouvant au niveau 1 de l'échelle de littératie sont celles qui sont uniquement capables de répondre à des questions simples (avec 80 % de chances pour être précis). Il existe donc une grande hétérogénéité dans cette population entre les personnes qui réussissent les questions de niveau 1 les plus difficiles (mais pas celles qui ont été classées en niveau 2) et les personnes qui parviennent à peine à réussir les questions les plus simples. Cependant, les formulations de l'OCDE insistant sur les grandes difficultés de ces personnes, leur très faible niveau de compétence ont pu justifier l'usage du terme « illettrés » dans ce qu'il a de plus stigmatisant.

L'échantillonnage

L'échantillonnage était fondé sur la liste des numéros téléphoniques et la méthode des itinéraires pour régler le problème des numéros en liste rouge. L'enquêteur pouvait remplacer le logement en cas d'impossibilité à joindre le ménage de façon persistante. Il n'est pas certain que la procédure ait été toujours parfaitement appliquée. Le taux de réponse de l'enquête est d'ailleurs problématique : 45 % des ménages ont refusé de répondre, introduisant des biais importants dans l'échantillon en terme de diplôme notamment, que le calage corrige assez mal. Le traitement des personnes ne parlant pas bien français est aussi assez flou et c'est encore plus le cas dans certains autres pays comme l'Allemagne, où toutes les personnes interrogées étaient germanophones. Dans le cas d'IVQ, l'usage de l'échantillon maître de l'Insee garantit une meilleure représentativité de l'échantillon obtenu.

La passation de l'enquête

L'enquête IALS était une enquête assez longue : le livret proposé à l'enquêté comportait une quinzaine d'exercices. La passation de l'épreuve pouvait prendre jusqu'à deux heures. Il était difficile dans ces conditions de maintenir l'attention et la motivation de l'enquêté tout au long du questionnaire. De plus, la situation de l'enquêteur était assez inconfortable car il se trouvait complètement désœuvré, ce qui accroissait chez la personne interrogée le stress et l'impression qu'elle disposait d'un temps limité pour faire les épreuves, même si les consignes précisaient le contraire. L'usage de CAPI dans IVQ permet des interactions plus naturelles entre enquêteurs et enquêtés. De plus, les exercices étant posés un à un, il est possible d'arrêter l'enquête avant que la fatigue ne rende les réponses de la personne peu significatives. On reviendra plus loin sur l'importance de prendre en compte la motivation pour analyser les performances des individus.

La présentation de l'enquête est apparue aussi très importante. La référence au ministère de l'éducation nationale sur les livrets semble avoir réactivé chez de nombreuses personnes des souvenirs peu agréables et a contribué à donner à l'enquête un tour scolaire peu propre à maintenir la motivation des personnes interrogées. Cela nous a conduits pour notre enquête à définir un protocole d'approche le plus neutre possible, évitant autant que faire se peut, de décrire le contenu exact de l'enquête. Des questions spécifiques de mise en situation, on le verra, ont aussi été introduites dans notre module d'orientation.

La codification des réponses

Une grille de correction avait été élaborée pour les épreuves mais de nombreuses critiques lui ont été adressées. En effet, elle s'avère très grossière : dans le fichier, on ne dispose plus généralement que de trois codes possibles : bonne réponse, mauvaise réponse, non-réponse. Or l'analyse des données et le retour aux questionnaires ont montré qu'il existait un flou assez important dans la codification.

En effet, d'une part, les exercices étaient souvent très ambigus et il est apparu dans de nombreux cas possible de donner une bonne réponse autre que celles qui avaient été prévues dans les consignes. L'alternance de questions simples et de questions difficiles a pu aussi provoquer la sensation de questions-pièges sur les exercices plus évidents, provoquant des réponses plus subtiles que ce que l'on attendait. C'est pourquoi il est apparu important dans le cas d'IVQ de recueillir les réponses de la façon la plus détaillée possible, le recours à CAPI permettant de les retranscrire aussitôt. Une opération de codage fin sera ensuite entreprise, le retour aux réponses originelles étant toujours possible.

On a aussi constaté une certaine confusion entre mauvaise réponse et non-réponse. Il semble en effet que très souvent les personnes ont sauté des exercices quand il portait sur des thèmes qui ne les intéressaient pas. De façon encore plus fréquente, les personnes interrogées interrompaient l'enquête

avant la fin, du fait de sa durée et de l'investissement demandé. Le codage et le traitement effectués sont alors assez flous : dans certains cas, la non-réponse sera considérée comme un échec à la question, dans d'autres comme une réelle absence d'information, sans que la distinction s'avère toujours pertinente et constante d'un enquêté à l'autre.

La comparabilité internationale

Alain Blum et France Guérin-Pace [1 ; 2], ainsi que Pierre Vrignaud [5 ; 12][2] ont beaucoup travaillé sur la comparabilité internationale de l'enquête. Ils ont mis en évidence de nombreux problèmes de traduction, qui souvent ont pu rendre plus difficiles les questions de la version française. Mais au-delà de la qualité de la traduction, ils mettent en doute la possibilité de construire une mesure unidimensionnelle permettant des comparaisons entre pays. De nombreux facteurs peuvent provoquer ce que l'on qualifiera de biais culturels et rendre telle question plus difficile dans tel pays que dans tel autre, alors que l'inverse s'observera sur un autre exercice. Des techniques statistiques existent pour repérer ces décalages mais le traitement à apporter n'est pas évident : suppression des items problématiques (mais quel seuil de divergence accepte-t-on) ou reconnaissance du caractère multi-dimensionnel du domaine que l'on veut étudier. Ces questions, très importantes dans le cadre d'une enquête internationale, ne seront pas développées ici car IVQ n'a pas pour l'instant vocation à être passée hors de France. On donnera cependant un exemple d'analyses similaires qui peuvent être faites en comparant des groupes sociaux.

Le dispositif IVQ

Elaboration de l'enquête

L'OCDE assistée de Statistique Canada et de ETS a décidé de lancer une autre enquête sur les compétences des adultes (ALLS qui a été menée en 2002) étendant son champ d'investigation à d'autres compétences comme la « résolution de problèmes ». Comme les principes de conception de IALS pour la partie littératie de ALLS seront reconduits à l'identique malgré les nombreuses critiques qui leur ont été adressées, la France a décidé de ne pas participer. Il est en effet apparu aux institutions françaises concernées[3] que les recherches sur l'évaluation des compétences des adultes devaient en priorité être menées dans une optique nationale.

Un comité de pilotage a donc été institué, où ces différentes institutions sont représentées, pour mettre au point un protocole d'enquête et des épreuves adaptées à la population française. Un premier test de l'opération a été mené en décembre 2000 sur quelques centaines d'individus pour s'assurer que le principe même d'une évaluation à domicile était possible. Le bilan de ce test était qu'il n'y avait pas de rejet trop marqué de l'enquête de la part des personnes interrogées mais que les problèmes de motivation (attention accordée au questionnaire ou lassitude en fin d'épreuve) se posaient effectivement de façon cruciale. C'est pourquoi un effort particulier a été fait pour capter de l'information sur le degré de motivation de chaque enquêté. Le deuxième test, sur un échantillon du même ordre, en avril 2002, a permis le choix et l'amélioration des épreuves pour l'enquête finale. Celle-ci s'est déroulée du 4 novembre au 14 décembre 2002 sur un échantillon initial de 4000 logements. Il est par ailleurs prévu de proposer une version du questionnaire à un échantillon de 7000 personnes ayant répondu à l'enquête FQP, l'échantillon étant ciblé sur les moins qualifiés.

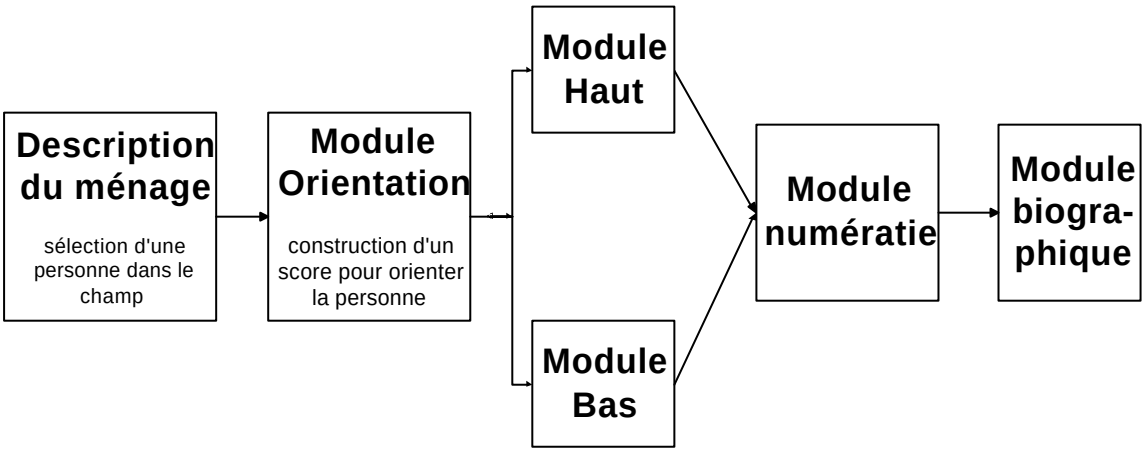
[2] Qui participent tous trois au comité de pilotage de l'enquête

[3] Dans le groupe de pilotage, se trouvent représentés l'ANLCI (Agence de lutte contre l'illettrisme), le CGP (Commissariat général au plan), le CREST (Centre de recherche en économie et en statistique), la DARES (Direction de l'Animation de la recherche et des études sociales du Ministère du travail), la DPD (Direction de la programmation et du développement du Ministère de l'éducation nationale), l'INED (Institut national des études démographiques), l'INETOP (Institut national d'étude du travail et de l'orientation professionnelle), l'INSEE (Institut national de la statistique et des études économiques).

Architecture de l'enquête

Les deux tests et l'enquête finale ont à peu près la même architecture (graphique 1). On sélectionne une personne dans un champ défini (pour l'enquête finale, il s'agit des individus de 18 à 65 ans). On lui propose un module d'orientation (dans le cas où la personne se déclare immédiatement illettrée, l'enquêteur insiste et évoque l'exercice de compréhension orale ; si la personne ne parle pas français, on se contentera de lui poser les questions biographiques, en tenant compte de cette caractéristique). Si la personne n'a pas eu de résultat suffisant, elle passera les exercices du module Bas (rebaptisé module ANLCI pour l'enquête finale) ; sinon, on lui propose les exercices du module Haut. Dans le test 2 et l'enquête finale, on propose ensuite des questions de « numératie », visant la maîtrise des compétences de base en calcul et en raisonnement logique (là encore, le niveau de difficulté des exercices est adapté en fonction des réponses à quelques questions simples insérées dans le module d'orientation). Ensuite, on recueille un certain nombre d'informations sur le parcours familial, scolaire et professionnel de la personne interrogée, sur ses pratiques de lecture et (pour le test 2 et l'enquête finale) sur d'éventuelles difficultés à accomplir certains actes de la vie quotidienne, pour les personnes ayant eu des performances médiocres aux tests.

Graphique 1 : architecture de l'enquête IVQ



Contenu des épreuves

Les épreuves diffèrent sensiblement selon l'étape de l'opération. On laissera ici de côté le module numératie non exploité dans cet article (il s'agit d'une suite de petits problèmes posés oralement). Voici une vision d'ensemble des épreuves.

test 1

Le test 1, on l'a dit, avait pour objectif de vérifier la faisabilité-même de l'enquête. Les épreuves elles-mêmes avaient été soit reprises d'évaluations scolaires, soit élaborées rapidement par les membres du groupe de pilotage.

- L'épreuve d'orientation comportait cinq textes, deux sur des supports quotidiens (une page de programme TV et des horaires de bibliothèques) et trois petits textes d'une dizaine de lignes.

- Le module Haut était constitué de huit textes répartis en quatre séries de deux textes (Famille-Société, Cinéma-Loisir-Spectacle, Littérature, Sport). La personne interrogée choisissait l'un des thèmes et passait les deux textes correspondant ; un autre thème parmi les trois restants lui était proposé.
- Le module Bas se composait de 7 séries d'exercices très simples, portant sur l'identification de mot, la compréhension de phrases ou de textes très courts.

test 2

Les épreuves du test 1 avaient relativement bien fonctionné mais des améliorations semblaient possibles. Le module d'orientation semblait un peu trop long et il est apparu préférable d'orienter plus rapidement les personnes en difficulté vers un module adapté (le seuil avait d'ailleurs était placé un peu trop bas si bien que peu de personnes avaient été orientées vers ce module au test 1). Le comité de pilotage a aussi souhaité élargir le champ des compétences évaluées : le module bas a été profondément repensé ; les supports du module Haut étaient plus diversifiés ; on a introduit l'évaluation de la numératie et de la compréhension orale. Des équipes de recherche universitaires ont été sollicitées pour mettre au point les épreuves.[4]

- Le module d'orientation se réduit maintenant à un exercice sur une page de programme TV (voir dans la partie suivante une présentation rapide de cet exercice).
- Le module Haut comporte 10 exercices sur quatre types de support : un test de compréhension orale, trois textes de la vie quotidienne (recette de cuisine, mode d'emploi d'un lave-vaisselle, une page du Guide du Routard), deux exercices sur des documents statistiques (tableaux et graphiques sur les accidents de la route), quatre textes plus classiques (articles de journaux ou un extrait de Victor Hugo). La personne passait l'exercice de compréhension orale, celui sur la recette de cuisine et quatre autres textes choisis de façon aléatoire.
- Le module bas, appelé module ANLCI du fait de la collaboration avec cet organisme, a été revu : il commence par le même exercice de compréhension orale que le module Haut (avec quelques questions supplémentaires de vocabulaire) puis la personne passe une « dictée » (le terme, très connoté, n'est bien sûr pas employé ; il s'agit d'une liste de course) et des questions d'identification de mot et de compréhension de texte sur un autre support quotidien (un CD de musique). On s'intéressait aussi à l'expression orale (exercice non repris dans l'enquête finale du fait de la difficulté du codage).

l'enquête finale

Le test 2 a permis d'affiner la composition du protocole d'exercices final.

- Le module d'orientation a été un peu aménagé : il est apparu que les consignes de codification n'avaient pas été parfaitement comprises et que de ce fait de nombreuses personnes ont été orientées vers le module ANLCI alors qu'elles auraient pu passer le module Haut (les enquêteurs ont été trop sévères dans leur correction). Des consignes plus claires ont été élaborées. L'exercice de compréhension de texte a aussi été complété et porte maintenant sur un film fictif.
- Le module ANLCI n'a pas connu de grand changement, sinon la suppression du questionnement d'expression orale, cette compétence étant trop difficile à évaluer.
- Pour le module Haut, 5 textes ont été finalement été retenus en plus de l'exercice de compréhension orale (la page du Guide du Routard, les tableaux et graphiques sur les accidents de la route, le texte de Victor Hugo, celui sur les familles recomposées et un récit de match de foot). En outre, un tiers de l'échantillon se voient proposée une série de six exercices

[4] L'équipe PsyEf de J.M. Besse (Université de Lyon 2) a ainsi mis au point le module Orientation, le module ANLCI et l'épreuve de compréhension orale, l'équipe EVA (Université de Rennes et Hôpital Ste Anne) de C. Charon et C. Meljac le module numératie et enfin, C. Chabrol (Université Paris 3) et P. Vrignaud (INETOP) ont élaboré quelques-unes des épreuves du module Haut. Tous les membres du comité de pilotage ont participé également à cette élaboration (voir notes 2 et 3).

issus de l'enquête IALS de 1994, ce qui permettra une comparaison des résultats avec des protocoles de collecte différents.

Consignes de passation

En effet, il importe de rappeler que l'effort a autant porté sur la constitution d'un protocole de collecte fiable que sur l'élaboration d'épreuves adaptées au public. Dans la présentation des critiques faites à IALS, on a évoqué les aménagements apportés au protocole d'IVQ afin d'y répondre. La capisation du questionnaire a facilité la prise d'informations. Les réponses seront connues précisément et surtout il est possible de recueillir deux types d'informations permettant de se faire une idée du degré de motivation de la personne interrogée : le temps qu'elle a consacré à répondre aux questions (pour le module Haut) et une grille d'observation de son comportement remplie par l'enquêteur.

Validité des épreuves

Cette partie est consacrée à l'analyse des épreuves du test 2. Pour avoir des éléments sur celles du test 1, on peut se reporter au document de travail consacré à cette étape de l'enquête. On ne trouvera pas ici l'intégralité des traitements effectués mais seulement trois exemples d'analyses portant sur le module d'orientation, l'exercice de compréhension orale et les textes du module Haut. Les épreuves de production écrite et de lecture du module ANLCI et la numératie ont aussi fait l'objet de travaux, non incluses ici. Toutes les études menées sur les trois domaines cités ne sont d'ailleurs pas présentées. Il s'agit plutôt de donner des exemples d'analyses :

- sur le module d'orientation, les techniques classiques de contrôle psychométrique seront appliquées aux données pour juger la qualité d'une épreuve ;
- pour l'exercice de compréhension orale, on montrera la spécificité de cette évaluation et la nécessité de procéder à un questionnement en deux temps ;
- les items du module Haut seront étudiés dans la perspective des biais culturels : y en a-t-il qui « favorisent » les hommes ou les plus diplômés ?

Le module d'orientation

Le but de l'épreuve d'orientation est de repérer très vite, même de façon un peu grossière, le niveau de la personne interrogée pour lui proposer des exercices adaptés. De plus, ce module a été conçu afin de permettre une bonne prise de contact : relevant clairement de la vie quotidienne, connu du tous (même ceux qui n'ont pas de télévision savent ce qu'est un programme TV) cet exercice permet de faire accepter le principe d'évaluation. L'épreuve comportait 13 questions que l'on peut regrouper en trois parties :

- bloc 1 : mise en situation (questions 1, 2 et 3)
  Les premières questions visent à permettre l'entrée de l'enquêté dans le principe d'évaluation. Elles sont volontairement un peu floues (difficiles à coder) et non scolaires. En montrant le programme, l'enquêteur demande : qu'est-ce que c'est ? A quoi ça sert ? Comment l'avez-vous reconnu ?
- bloc 2 : identification de mots (questions 4, 5, 6, 7 et 8)
  Dans ce bloc, il s'agit de voir si la personne parvient à lire des mots : le titre d'une émission, son sujet, une date, etc.
- bloc 3 : compréhension (questions 9, 10, 11, 12 et 13)
  Quelques questions portent enfin sur le film de 20 h 45 (le Grand Pardon), questions ponctuelles ou demandant de résumer le film.

Ce découpage est évidemment théorique : il est très difficile d'isoler les compétences mises en jeu dans la lecture : le repérage et la lecture de mot d'une part et la compréhension d'autre part. De plus, le

support est aussi un élément à prendre en considération : deux questions portant sur un support identique risquent d'être liées même si elles sollicitent des compétences différentes.

Le tableau 1 donne les principales caractéristiques de l'épreuve pour les 445 personnes ayant répondu aux questions. L'épreuve est plutôt facile : l'item le plus difficile est quand même réussi par 60 % de la population et ce taux est supérieur à 90 % pour 8 items sur 13. L'alpha de Cronbach, dépendant de l'ensemble des corrélations entre items, mesure la cohérence interne de l'épreuve. Avec une valeur de 0,71, on peut considérer que l'épreuve est d'assez bonne qualité, que les items mesurent tous à peu près la même chose. La valeur est cependant loin de son maximum (1) ce qui ne surprend pas étant donné le découpage a priori en 3 blocs. On peut aussi mener ce raisonnement sur la fidélité de l'épreuve au niveau de chaque item en étudiant la corrélation entre la réussite à tel item et le score obtenu sur l'ensemble des autres items (nombre total de bonnes réponses). On se contentera ici du coefficient de Pearson même si la nature dichotomique de la réussite le rend un peu inadapté. Les corrélations sont plutôt bonnes (supérieures à 0,3) mais certains items posent un peu problème (les items 9, 5, 12 et 13).

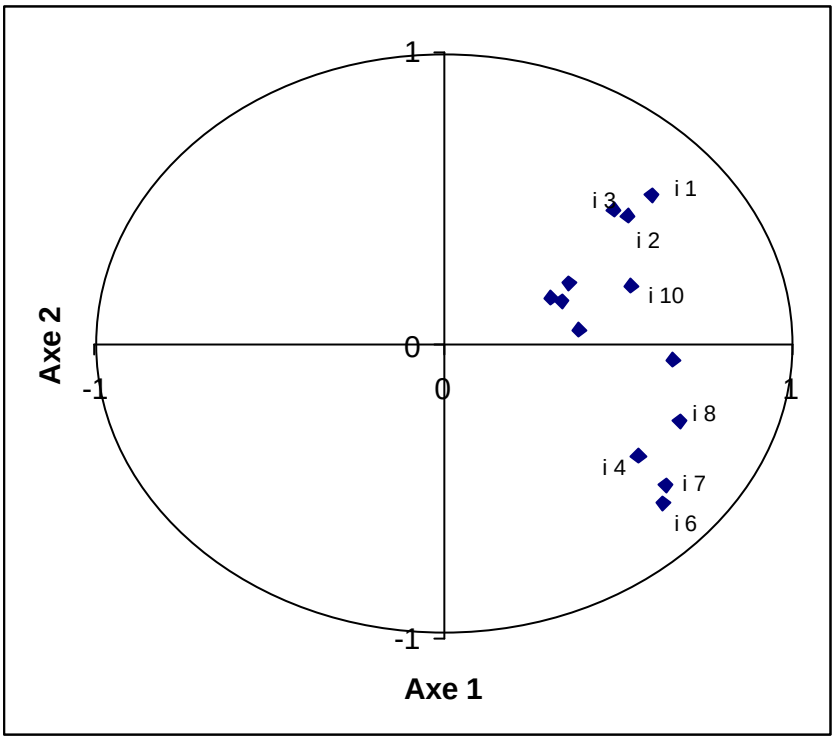
Tableau 1 : caractéristiques de l'épreuve d'orientation

N° d'item	% de réussite	corrélation item-test
1	93%	0,50
2	95%	0,43
3	87%	0,36
4	96%	0,36
5	96%	0,27
6	98%	0,41
7	95%	0,40
8	96%	0,49
9	67%	0,24
10	86%	0,43
11	97%	0,51
12	88%	0,27
13	60%	0,27

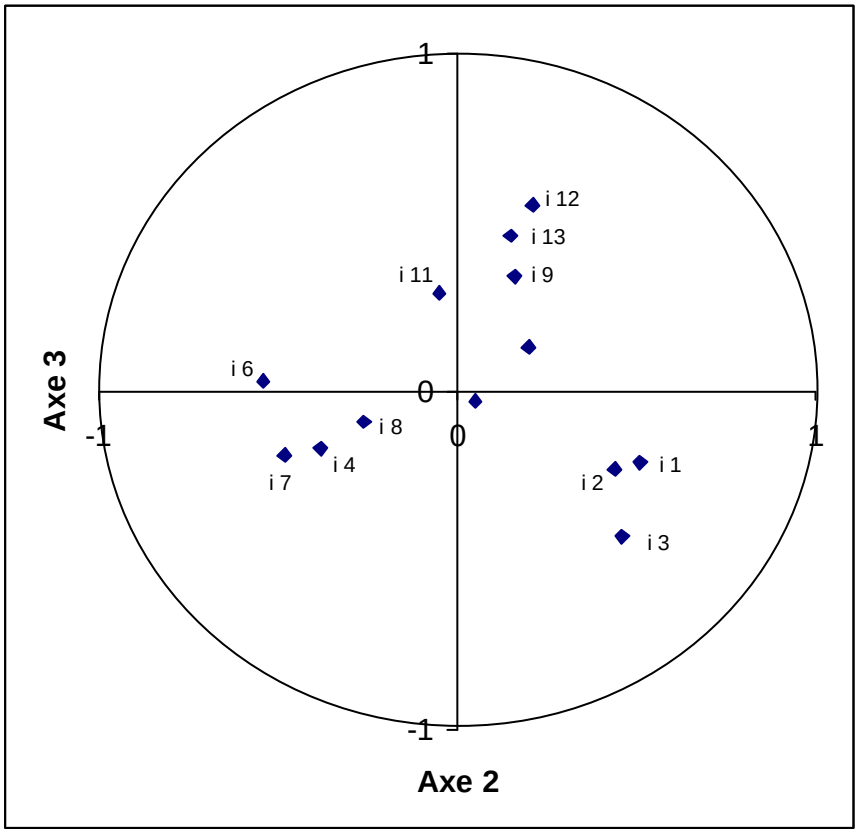
Lecture : l'item 1 du module d'orientation est réussi par 93 % de la population ; la corrélation (r de Pearson) entre la réussite à cet item et le score construit sur les autres items est de 0,5.

L'examen d'une épreuve cognitive peut utiliser d'autres outils. Par exemple, les techniques d'analyses de données ont été développées en grande partie par les psychomètres cherchant à étudier statistiquement l'« intelligence ». Appliquées aux données, ces techniques vont permettre de vérifier la qualité de l'épreuve, en tenant compte de son caractère composite. Une analyse en composante principale a donc été effectuée sur les 13 indicatrices de réussite. Les résultats confirment ce qui vient d'être dit : le premier axe explique 28 % de l'information à lui seul contre 11 % pour le deuxième et 8 % pour le troisième. Ce premier axe, sur lequel toutes les indicatrices ont une projection positive (graphique 2) est assez nettement prépondérant et correspond à un facteur « général » dans l'épreuve. Cependant, tout en étant de moindre importance, d'autres facteurs semblent à prendre en compte que l'on peut repérer sur les axes 2 et 3 (graphique 3).

Graphique 2 : projection des indicatrices de réussite sur les deux premiers axes



Graphique 3 : projection des indicatrices de réussite sur les axes 2 et 3



Le deuxième axe oppose de façon nette les items du bloc 1 de mise en situation avec ceux du bloc 2 portant sur l'identification de mot. Quant au troisième axe, il met en évidence les corrélations entre items du bloc 3 sur la compréhension qui s'opposent à tous les autres (exception de l'item 5 assez atypique sur les trois axes : il fallait donner la date du programme) et surtout à ceux du bloc 1.

Le constat de cette étude est nuancé : la forte corrélation entre l'ensemble des items peut justifier la construction d'un score global, par exemple le nombre de bonnes réponses (dans le test, les questions étaient pondérées suivant leur difficulté a priori). Il est bien licite d'utiliser ce score pour orienter les personnes vers le module ANLCI ou le module Haut (reste à déterminer un seuil). Ce score peut aussi être utilisé dans des analyses sociologiques, en le confrontant au diplôme, au sexe, à l'âge. Cependant, il est apparu clairement que ce score ne donnait qu'une vision partielle de l'information contenue dans l'épreuve. Pour avoir une vision plus complète, il sera utile de tenir compte de facteurs de deuxième ordre sur la nature des compétences mises en jeu dans l'acte de lire.

La compréhension orale

L'analyse ne portera ici que sur les questions qui étaient posées à la fois aux personnes ayant passé le module ANLCI et à celles ayant passé le module Haut. On dira un mot de la qualité psychométrique de l'épreuve après avoir tenu compte de la spécificité de l'évaluation de la compréhension orale. En effet, on devrait ici parler autant de mémoire que de compréhension, les deux compétences étant liées (on mémorise vite ce que l'on comprend bien). Par rapport aux questions écrites, l'épreuve avait donc la particularité de se dérouler en deux temps. Après l'écoute du message sonore, l'enquêteur posait une première fois les questions, puis le message était rediffusé et les questions à nouveau posées. L'enquêté savait que les mêmes questions seraient posées. Seules trois questions, communes aux deux populations, n'étaient posées qu'une fois : la première n' apparaît qu'à la première écoute (sur ce que l'on vient d'entendre : il faut la corriger immédiatement dans les rares cas d'erreurs pour pouvoir poursuivre l'interrogation) ; deux questions sont posées en plus dans la deuxième après la série de 6 questions répétées.

Tableau 2 : réponses aux questions de l'exercice de compréhension orale

n°	Toujours faux	Juste puis faux	Faux puis juste	Toujours juste
1	2,5			97,5
2-8	4,7	1,1	14,8	79,3
3-9	4,5	3,8	13,0	78,7
4-10	27,6	2,5	39,6	30,3
5-11	20,5	4,0	40,2	35,3
6-12	9,7	2,0	27,2	61,1
7-13	8,5	2,3	28,5	60,7
14	7,9			92,1
15	11,2			88,8

Lecture : l'item 1 (passé seulement en première écoute) est réussi par 97,5 % de la population ; 4,7 % de la population échouent à l'item 2 lors des deux écoutes, 1,1 % le réussissent à la première et échouent ensuite, 14,8 % se trompent puis réussissent, 79,3 % donnent une réponse juste les deux fois.

On va s'intéresser plus particulièrement aux questions qui ont été posées deux fois. A la première écoute, les résultats ne sont pas toujours satisfaisant (tableau 2) : en particulier les questions 4 et 5 sont réussies par moins de 40 % de la population. Comme la réponse était « vrai » ou « faux », on voit que

les gens répondent moins bien que s'ils répondaient au hasard[5] ! La deuxième écoute paraît donc fondamentale : en effet, elle permet d'isoler quelques personnes qui avaient juste la première fois mais pas la deuxième (environ 10 % des personnes ayant eu juste à la première écoute changent d'avis pour la question 5) : ils avaient sans doute, pour certains, répondu au hasard la première fois et ont dû reprendre cette stratégie, sans se souvenir de leur choix. Cependant, la plupart des personnes ayant bien répondu tout d'abord confirment leur réponse[6]. Il apparaît qu'au contraire la majorité des personnes s'étant tout d'abord trompé révisent heureusement leurs réponses. Seules, entre un vingtième et un quart de la population reste dans l'erreur (ou l'ignorance car on inclut ici les réponses « Ne sait pas »). En définitive, à la deuxième écoute, pour ces deux questions, les plus difficiles, le taux de réussite approche ou dépasse 70 %.

Les indicateurs psychométriques confirment l'intérêt crucial d'une seconde interrogation : si l'on travaille sur les items en première écoute (2 à 7), l'alpha de Cronbach est de 0,56 ; il passe à 0,65 si l'on utilise les items en deuxième écoute (8 à 13), ce qui est satisfaisant compte tenu du faible nombre de questions. La cohérence de l'épreuve paraît meilleure quand on utilise la deuxième vague de questions.

Cette seconde interrogation paraît ainsi indispensable pour avoir une bonne image des compétences en compréhension orale de la population. En effet, elle se rapproche plus d'une situation naturelle, dans le sens où la personne a généralement des informations sur le contexte du texte qu'elle écoute et un objectif dans cette écoute. La première écoute illustre cependant les difficultés de compréhension que peut poser un texte auprès d'un public non concentré et non préparée.

A titre d'illustration, on a construit deux scores sur les items à deux vagues. Lors de la première écoute, 24 % des personnes ont réussi au plus 2 questions ; 49 % a réussi de 3 ou 4 questions ; 27 % a commis au plus une erreur (tableau 3). Après la deuxième écoute, à une exception près, toutes les personnes se maintiennent ou améliorent leur classement : il n'y en a plus que 13 % à réussir au mieux 2 questions ; la proportion de très bons résultats passe à 61 %.

Tableau 3 : résultat à l'exercice de compréhension orale selon le niveau en lecture

	Première écoute			Deuxième écoute		
	0-2 BR	3-4 BR	5-6 BR	0-2 BR	3-4 BR	5-6 BR
Personnes orientées vers le Module ANLCI	39	48	13	29	29	41
Personnes orientées vers le Module Haut	19	50	31	8	25	67
Ensemble	24	49	27	13	26	61

Lecture : 39 % des personnes orientées vers le module ANLCI ont eu au plus deux bonnes réponses sur les 6 questions répétées de l'exercice oral, lors de la première écoute ; ce taux passe à 29 % à la deuxième écoute.

La confrontation des résultats en compréhension orale avec ceux en lecture est assez instructive : que l'on prenne les données avant ou après réécoute, la corrélation est forte mais elle s'amplifie entre les deux temps (le coefficient de contingence sur les tableaux croisés passe de 0,22 à 0,28). C'est particulièrement net si l'on isole les personnes en difficulté face à l'oral : au départ, 39 % des personnes ayant passé le module ANLCI se trouvent dans cette population (contre 19 % du reste de la population, soit un rapport de chances relatif[7] de 2,7). Lors de la deuxième vague, il en reste encore 29 % (contre 8 % parmi les autres, soit un rapport de chance relatif de 4,7). En revanche, si l'on travaille

[5] Pour bien comprendre ce résultat, il faut entrer dans le détail des questions. La passagère d'une voiture a été blessée au visage parce que la conductrice du véhicule où elle se trouvait n'avait pas respecté les distances de sécurité. Les deux questions étaient : la conductrice de la fourgonnette a été blessée au visage (Vrai ou Faux) ; la conductrice n'était pas seule dans la fourgonette (Vrai ou Faux). Les indicatrices de réussite à ces questions sont par ailleurs très fortement corrélées.

[6] Bien sûr, certaines personnes ont pu répondre au hasard juste les deux fois ou bien avoir répondu juste au hasard la première fois et de façon plus assurée la deuxième.

[7] C'est la probabilité que pour un couple Module ANLCI-Module Haut, le premier soit en difficulté face à l'oral et pas l'autre, divisé par la probabilité de l'événement inverse.

sur les individus ayant de bons résultats, on n'observe pas de changement : l'écart de chances relatif est de 0,33 au départ ; il est de 0,34 ensuite.

Ces résultats amènent plusieurs remarques : premièrement, les personnes maîtrisant la lecture ont généralement eu aussi de bons résultats à l'épreuve orale en particulier en deuxième écoute mais c'est aussi le cas d'un certain nombre de personnes ayant eu des difficultés au test d'orientation. Deuxièmement, la réécoute entraîne une nette amélioration pour les personnes sans difficulté face à l'écrit : moins de 10 % demeurent sous la barre des 3 bonnes réponses ; les deux tiers ont alors des résultats très satisfaisants. On observe aussi une forte augmentation des très bons résultats chez les personnes ayant passé le module ANLCI (de 13 à 41 %) mais la baisse de la proportion des personnes en difficulté face à l'oral est moins nette (on passe de 39 % à 29 %). On peut en conclure que la réécoute permet de mieux distinguer parmi les personnes en difficulté face à la lecture, ceux qui ont aussi des problèmes de compréhension orale et ceux qui n'en ont pas[8].

Le module Haut

Le module Haut a aussi fait l'objet d'une analyse psychométrique semblable aux précédentes mais nous allons l'utiliser ici pour traiter la question des biais culturels. Il ne s'agit pas d'étudier la comparabilité internationale des items puisque l'enquête est uniquement française mais de montrer le type de question que l'on se pose quand on veut comparer deux populations, deux pays ou, comme on va le faire, deux groupes sociaux. Le plus simple est de comparer ces deux groupes en utilisant le score obtenu sur l'ensemble de l'épreuve. Dans le cas du module Haut, on ne peut utiliser la proportion de bonnes réponses, car tout le monde ne passait pas les mêmes exercices (un choix aléatoire de quatre textes parmi huit était fait en plus du premier qui était passé par tous). Ceux qui ont passé les exercices les plus difficiles risquent d'être désavantagés. Pour tenir compte de cela, on a eu recours à un modèle de réponse à l'item, technique statistique souvent utilisée dans ce genre de cas, en particulier dans les enquêtes internationales. Il est ainsi possible de construire un score SMRI résumant au mieux l'information contenue dans les réponses d'une personne et tenant compte de la difficulté des questions qu'il a passées[9].

Ces mêmes techniques sont aussi souvent utilisées pour étudier le fonctionnement des items, les biais dont nous allons parler. On se contentera ici d'une présentation beaucoup plus simplifiée du problème. Il s'agit de voir si certaines questions « favorisent » une catégorie donnée de la population. En effet, si l'on se place sous l'hypothèse d'unidimensionnalité (hypothèse implicite dès que l'on calcule un score) il faudrait que tous les items soient cohérents quand ils mesurent l'écart entre deux groupes : si pour un item, les membres du groupe A sont meilleurs que ceux du groupe B et que c'est l'inverse pour un autre item, la conclusion est ambiguë et l'équivalence des deux groupes que l'on risque d'observer en agrégeant les deux items est aussi trompeuse. Il est donc souhaitable, si l'on veut construire un score, que l'on n'observe pas trop de ces incohérences qui remettent en cause l'hypothèse d'unidimensionnalité. Les tests portant sur le fonctionnement des items sont nombreux et la question est abondamment traitée dans la littérature. On n'utilisera pourtant ici que des techniques rudimentaires pour illustrer le problème.

Un point important est à signaler : un groupe n'est pas forcément favorisé par l'item A, si les membres de ce groupe le réussissent plus souvent. Cette réussite peut tout simplement être le signe d'une

[8] Il est aussi possible d'étudier la dispersion de ces populations (en utilisant par exemple des indicateurs dérivés de l'écart-type du nombre de bonnes réponses au test oral) : alors que cette dispersion évolue peu pour les personnes ayant passé le module Haut, elle croît sensiblement pour ceux qui ont des difficultés en lecture : la deuxième vague montre plus clairement la fote hétérogénéité de cette population en terme de compétence orale.

[9] On attribue à chaque item un paramètre de difficulté β_j et à chaque individu un paramètre de compétence θ_i, que l'on estime en postulant que la probabilité de réussite de l'individu i à l'item j est de la forme $1/(1+\exp(\theta_i-\beta_j))$. On utilise généralement une technique de maximisation de la vraisemblance pour obtenir la valeur des coefficients. Des modèles plus complexes existent, introduisant d'autres paramètres comme la discrimination de l'item ou la pseudo-chance, probabilité de réussir l'item au hasard (voir [4]).

compétence plus grande que le reste de la population. La question est de savoir si un écart d'une même ampleur s'observe sur les autres items. Dans ce cas, pour déterminer s'il y a bien un biais ou non, il faut neutraliser un éventuel écart de compétences, et comparer ce groupe avec le reste de la population, toutes choses égales par ailleurs, et voir s'il subsiste un écart significatif de performances. Pour ce faire, on peut par exemple découper l'ensemble de la population en quartiles (ou en déciles) et comparer au sein de chaque quartile la réussite à l'item des gens du groupe A avec la réussite des autres personnes. Si, la compétence mesurée par l'ensemble des items étant fixée, un écart apparaît, on peut affirmer qu'il y a biais dans la réussite, qui n'est pas justifié par un écart de compétence. La généralisation immédiate de cette méthode consiste à construire une modélisation logistique de la réussite à un item en fonction du score global et d'une indicatrice de groupe. Si des écarts significatifs apparaissent entre groupe à score donné, il y a un biais qu'il faut expliquer.

Pour commencer, on va présenter un cas simple, celui des écarts selon le sexe. En effet, si l'on utilise le score global, la différence entre les deux groupes est non significative (0,04 alors que l'écart-type du score a été fixé à 1, la moyenne se trouvant à 0). L'analyse des items est donc simple : pour aucun item, on ne devrait trouver d'écart significatif entre hommes et femmes. Il s'agit bien sûr d'une situation idéale que l'on ne rencontre jamais. Bien des facteurs peuvent expliquer des performances différentes et il est rarement possible de supprimer les questions susceptibles d'être biaisées car souvent on supprime aussi une composante fondamentale de la compétence que l'on veut mesurer.

Dans le cas de notre enquête, sur les 32 items du module Haut, 6 items présentent un écart significatif (au seuil de 10 %). Cela n'est pas beaucoup mais chaque question n'était passé que par 150 personnes environ ce qui rend les tests de significativité assez exigeants. Le tableau 4 montre d'ailleurs que les écarts mis en évidence sont en effet tout à fait sensibles. On trouve autant d'écart en faveur des femmes qu'en faveur des hommes : par construction, les biais se compensent pour retrouver un écart nul sur l'ensemble des épreuves.

Tableau 4 : les items biaisés selon le sexe

Numéro d'item	1	13	16	21	27	30
Femmes	**73**	4	**70**	65	41	**90**
Hommes	61	**11**	56	**79**	**59**	80

Lecture : 73 % des femmes réussissent l'item 1 contre 61 % des hommes

Il n'est parfois pas trop difficile de trouver une explication. Ce peut par exemple être un effet du support : la question 1 demande le nombre d'ingrédients utilisés dans une recette de cuisine et elle favorise les femmes ; la question 21 porte sur un élément de compréhension ponctuelle dans le récit d'un match de football et les hommes le réussissent mieux. Cependant, ces interprétations simples ont leur limite : sur les deux autres items utilisant la recette, il n'y a pas d'écart de plus de 6 points entre hommes et femmes ; pour les trois autres questions sur le match de foot, on observe une égalité parfaite de réussite, un écart de 6 points en faveur des hommes et un autre du même ordre en faveur des femmes (donner un titre au récit) ! De même, l'écart à la question 30 pourrait trouver son origine dans le plus grand intérêt que les femmes ont pris à un texte sur les familles recomposées mais l'une des questions du même exercice est un peu mieux réussie par les hommes ce qui devra être expliquée (9 points : on est juste au-dessus du seuil des 10 % de significativité). La question 27 demandait un calcul et l'analyse des réponses à la numératie confirme que les hommes sont plus à l'aise dans ce domaine. Quant aux questions 13 et 16, il faudrait savoir si les hommes connaissent mieux Vauban (repérer dans un texte du Guide du Routard les monuments qu'il a construits) et les femmes plus proches de Victor Hugo (contexte d'un texte sur la peine de mort).

Les mêmes difficultés d'interprétation vont apparaître quand on va comparer les diplômés de l'enseignement supérieur et les autres. Dans ce cas, il n'est plus possible de comparer brutalement les taux de réussite des deux catégories : la différence de scores globaux représente 84 % d'écart-type. En effet, dans ce cas sur la plupart des items, les diplômés auront un taux de réussite supérieur : les deux premières lignes de % du tableau 5 présentent ces taux pour les six items qui vont apparaître biaisés.

Même pour les trois items où le biais est en faveur des non diplômés du supérieur, ils ne creusent pas l'écart : au mieux (item 9), ils passent 1 point au-dessus des diplômés. On a donc recours à la modélisation psychométrique pour neutraliser cet écart de compétence. Dans ce cas, l'indicatrice de diplôme apparaît 6 fois signativement liée à la réussite à l'item. Là encore, par construction, les biais apparaissent trois fois en faveur des diplômés du supérieur, trois fois en faveur du reste de la population. Dans le tableau X, les deux dernières lignes donnent des résultats issus de la modélisation : la probabilité pour les personnes de compétence moyenne (SMRI=0) de réussir l'item si elles sont diplômées de l'enseignement supérieur ou non.

Tableau 5 : les items biaisés selon le niveau de diplôme

	Numéro d'item	2	9	11	17	18	20
Taux de réussite bruts	Non diplômés du supérieur	76%	22%	49%	70%	19%	66%
	Diplômés du supérieur	78%	21%	83%	66%	58%	98%
Taux pour les personnes avec SMRI=0	Non diplômés du supérieur	**82%**	**24%**	54%	**74%**	20%	78%
	Diplômés du supérieur	70%	7%	**73%**	57%	**46%**	**97%**

Lecture : les diplômés de l'enseignement supérieur sont 78 % à réussir l'item 2 contre 76 % de ceux qui ne sont pas aussi diplômés. Si l'on se restreint aux personnes dont les résultats aux tests sont dans la moyenne (SMRI=0) la probabilité de réussir l'item est 70 % pour les diplômés du supérieur et 82 % pour les autres (probabilités issues d'une modélisation logistique).

Il faut mettre à part la question 9 qui se trouvait dans un exercice utilisant des statistiques d'accidents de la route. La formulation de la question était très ambiguë et les personnes diplômés de l'enseignement supérieur ont souvent donné une réponse que l'on pourrait considérer comme juste (reste à savoir pourquoi le même mouvement ne s'est pas observé chez les autres). En revanche, ils ont fait preuve d'une bien meilleure aisance dans la lecture de graphiques que comportait aussi cet exercice (question 11). Ils ont par contre été gênés par un certain flou, qui était là voulu, de l'une des questions sur la recette de cuisine (question 2). Ils ont fait preuve de plus de finesse pour l'item 20 demandant le score d'une partie de football : ils ont donné le score final (2-2) alors que les moins diplômés préfèraient donner le score à la mi-temps, qui était en effet indiqué en chiffre dans le texte.
Le plus intéressant concerne les items 17 et 18 : les diplômés de l'enseignement supérieur ont très bien répondu à la question 18 qui visait la compréhension d'un des arguments que Victor Hugo exposait contre la peine de mort. On serait tenté de mettre en avant le caractère très littéraire du texte, au contraire du reste de l'épreuve, pour expliquer ce biais. Cependant, les résultats à l'item 17 indique le contraire : sur le même texte, les diplômés sont alors moins à l'aise à repérer un autre argument. Beaucoup se laissent tenter par une modalité (il s'agit d'un QCM) qui, sans être fausse, ne donne en fait qu'une vision partielle de la pensée de Victor Hugo, une autre modalité étant plus complète mais plus longue.

Des indicateurs de motivation liées aux performances

L'une des critiques les plus sérieuses adressées aux enquêtes sur les compétences telles que IALS remet en cause le manque de réalisme des conditions de passation. Ainsi que le rapportent l'ensemble des enquêteurs ayant réalisé les deux tests de l'enquête IVQ en décembre 2000 et avril 2002, il est certain que lorsque l'enquêté est placé devant un exercice artificiel, à son domicile et devant un enquêteur qu'il ne connaît pas, il montrera sans doute moins d'entrain[10] à s'employer à le résoudre que si une situation réelle - et qui comporte un réel intérêt pour lui - l'y oblige. Les performances enregistrées seront donc nécessairement plus basses que ce que ses capacités réelles lui permettraient d'atteindre. Par ailleurs, cet écart entre performances enregistrées et capacités réelles n'est pas uniforme : il variera selon le degré d'implication de l'enquêté. De multiples facteurs peuvent influer sur cette implication, globaux (le goût pour les jeux de réflexion - jeux de société, mot-croisés -,

[10] Ou alors, il ressentira davantage de stress que s'il était seul.

l'habitude des situations scolaires[11], l'adhésion aux objectifs de l'enquête), comme locaux (mauvaise humeur passagère, fatigue, environnement perturbateur, etc.).

Afin de résoudre au moins partiellement ce problème, deux stratégies peuvent être poursuivies :
- essayer d'accroître, autant que faire se peut, l'engagement des enquêtés.
- collecter des indicateurs témoins de cet engagement afin de corriger les résultats observés lors de l'exploitation ultérieure.

Deux types d'indicateurs susceptibles de rendre compte de la motivation.

On se propose dans la suite de développer quelques aspects de la deuxième stratégie. En effet, l'emploi de la technique CAPI de l'enquête permet de collecter facilement deux types d'indicateurs susceptibles de rendre compte de la motivation :

• *le temps de lecture et de réflexion pour chaque texte proposé*. Ainsi, lorsque l'enquêteur remet un texte à l'enquêté, il lui demande de lire le texte et de réfléchir aux questions associées. Lorsque l'enquêté estime être en mesure de répondre aux questions, il l'indique à l'enquêteur qui interrompt alors le décompte du temps. Précisons que la mesure du temps se déroule à l'insu de l'enquêté. Parmi les indicateurs de motivation, le temps de lecture est susceptible de constituer un bon candidat. On peut penser en effet que moins l'enquêté est intéressé par un texte donné, moins il y consacrera de temps. Le test 1 a mis en œuvre cette technique[12].

• *des indicateurs de comportement :* Dans le test 2, l'enquête intègre pour chacun des textes passés une grille de comportement. L'enquêteur essaie de qualifier le comportement de l'enquêté face au texte proposé suivant quatre dimensions : le stress, la colère, la lassitude et les signes de difficulté. la question générique (par exemple sur la dimension lassitude) se présente de la façon suivante :

L'enquêté a-t-il montré des signes de lassitude ?
1. *Oui, importants*
2. *Oui, légers*
3. *Non, pas de signes de lassitude*

Dans la suite, nous présentons quelques données de description de ces deux types d'indicateurs, puis dans la dernière partie, nous exploitons les données des deux tests afin de mettre en évidence d'éventuels liens entre ces indicateurs et les performances observées. Précisons enfin qu'un autre travail (Lollivier et Verger 2002 [9]) tente également de mettre en évidence une influence de la motivation sur les performances mais à partir d'autres indices indirects de démotivation déjà disponibles dans IALS.

Quelques statistiques descriptives sur les temps de réflexion.

Le temps de réflexion dépend de la longueur du texte et de la difficulté des questions

Le temps de lecture des textes et de résolution des questions varie manifestement selon la longueur des textes (tableau 6). Les temps médians du module Orientation - qui comprenaient des textes ou documents assez courts - s'échelonnent ainsi entre 1,5 minutes et un peu plus de 2 minutes, alors que les temps du module Haut - composé d'articles de journaux ou textes littéraires d'une demi-page

[11] Par exemple, entre autres facteurs explicatifs, l'un des arguments invoqués pour rendre compte d'écarts entre performances françaises et performances américaines reposait sur la plus grande familiarité des américains pour les exercices de type questions-réponses.
[12] Pour des raisons techniques, le test 2 n'a pas intégré cette innovation.

environ - sont compris entre 3 et plus de 5 minutes. Mais la difficulté des questions est également un facteur qui influence les temps. Ainsi, les textes Or3, Or4, Or5 sont à peu près aussi longs et contiennent le même nombre de questions mais ils sont associés à des questions de difficulté croissante, ce qui explique probablement que le temps médian passe de 70 secondes à 100 secondes. De même, c'est le texte L2 - texte de Victor Hugo sur la peine de mort -, qui a certainement posé le plus de problèmes aux enquêtés et donné lieu à un temps de réflexion parmi les plus longs.

Tableau 6 : statistiques descriptives sur les temps de réflexion par exercice

texte	Temps de réflexion médian	Ecart-type	Premier Quartile (25%)	Troisième Quartile (75%)	Nbre de questions par texte	Taux de réussite
MODULE ORIENTATION						
OR1	144	95	98	205	4	68,0%
OR2	119	72	93	168	4	69,8%
OR3	71	63	52	95	4	77,0%
OR4	92	62	68	125	4	71,9%
OR5	103	69	71	143	4	62,6%
MODULE HAUT						
Cinéma-Spectacles-Loisirs						
CSL1	326	168	250	432	6	73,0%
CSL2	192	113	147	251	5	62,0%
Famille-Société						
FS1	192	101	154	250	5	70,9%
FS2	231	142	167	292	4	77,6%
Littérature						
L1	204	84	154	250	5	74,7%
L2	300	138	233	403	4	61,3%
Sport						
S1	191	113	150	276	5	72,7%
S2	264	109	216	331	5	61,4%

Temps (en secondes)
Lecture :50% des enquêtés qui ont passé le texte CSL1 y ont consacré un temps de réflexion inférieur à 326 secondes, 25% y ont consacré moins de 250 secondes et 75% moins de 432 secondes. La moyenne de réussite aux six questions associées à ce texte est de 73%.

On peut constater également que la variance des temps de réflexion est assez importante, ce qui peut témoigner de la disparité des efforts fournis ou des attitudes face à un texte donné.

Moins de temps faibles lorsque le texte est choisi.

Nous nous sommes attachés à décrire les conditions d'apparition de temps de réflexion courts. Rappelons en effet l'hypothèse que nous voulons vérifier dans cette brève étude : les temps courts sont-ils le signe d'une moindre motivation ? Mais il faut avant toute chose définir ce qu'on appelle temps court. Cette définition dépend naturellement du texte choisi. Pour un texte donné, on appelle temps court au seuil de 5% (par exemple), les temps qui figurent parmi les 5% les plus faibles, ce qui a été choisi dans la suite.
Rappelons pour continuer que le protocole choisi lors du test 1 permettait aux enquêtés, à l'entrée du module Haut, de choisir un thème parmi les quatre suivants : Cinéma-Spectacles-Loisirs, Famille-Société, Littérature, Sport et de passer les deux textes associés. Ensuite, un thème parmi les trois restants était tiré au sort et l'enquêté passait les deux textes associés.
Une procédure logistique (Tableau 7) a donc été effectuée afin de décrire les conditions d'apparition de temps de réflexion faibles. Les variables introduites dans la régression sont les variables socio-démographiques classiques. On a également rajouté une variable qui indique si le texte a été passé en mode choisi ou aléatoire.

Tableau 7 : modélisation du fait d'accorder peu de temps à un exercice

		Coeff	Ecart-type	Stat du Chi²	Prob.
Constante	1	-2.6144	0.5173	25.5446	<.0001
Txt choisi	1	-0.6667	0.3434	3.7706	0.0522
(à opposer à texte soumis aléatoirement)					
Plus haut diplôme possédé					
BEP/CAP	1	0.0195	0.4988	0.0015	0.9688
BAC	1	0.0573	0.5598	0.0105	0.9184
DEUG et +	1	0.3813	0.5307	0.5162	0.4725
Ref : Sans diplôme/CEP/Brevet					
Position prof. Actuelle					
Prof int.	1	-1.5149	1.0816	1.9616	0.1613
Cadres	1	0.7360	0.5790	1.6157	0.2037
Employés	1	0.4596	0.5105	0.8108	0.3679
Ref : Ouvriers					
Statut					
Chômeur	1	0.2755	0.5766	0.2282	0.6328
Inactif	1	0.2497	0.4667	0.2862	0.5927
Ref : actifs occupés					

Lecture : La variable expliquée est ici « avoir consacré un temps de réflexion faible ». Le fait de passer sur un texte choisi joue positivement sur cette variable et significativement au seuil de 10%.

Les résultats (tableau 7) ne montrent aucun résultat significatif concernant les variables socio-démographiques. Il conviendra donc d'attendre une enquête sur davantage d'individus pour analyser plus avant les corrélations éventuelles. En revanche, la corrélation relative à la variable du mode de passage (choisi ou aléatoire) est presque significative au seuil de 5%. En mode choisi, la probabilité est plus faible d'observer un temps de réflexion court. Toutefois, l'interprétation de cette constatation n'est pas pour autant très claire. Deux raisons peuvent se conjuguer en effet pour expliquer ce résultat.

i. Un texte choisi a un thème qui a priori correspond davantage aux goûts de l'enquêté ou tout au moins qui lui est plus familier. Dans cette mesure, on peut penser qu'un texte choisi suscite une motivation supérieure, ce qui expliquerait que les enquêtés prennent suffisamment de temps pour les lire et les résoudre.

ii. Dans le dispositif choisi, les textes choisis se trouvent toujours avant les textes aléatoires. On peut penser également que le résultat observé s'explique par une lassitude croissante des enquêtés, qui attribuent de moins en moins de temps à la lecture des textes et à leur résolution.

En l'état actuel du protocole, rien ne permet de distinguer entre ces deux causes possibles. Toutefois, les deux causes renvoient à des phénomènes comparables : des temps inférieurs renvoient dans les deux cas à une moindre implication ou motivation.

Quelques statistiques descriptives pour les indicateurs de comportement.

Dans le tableau 8, les indicateurs sont déclinés selon l'ordre des textes passés au long de l'enquête. Précisons que les personnes ayant montré des signes importants et légers ont été groupées dans la même catégorie. La part des gens ayant montré des signes importants est en effet très faible et ne présente pas d'évolution particulière tout au long de l'enquête.
Dans leur bilan qualitatif, un grand nombre d'enquêteurs avaient rapporté que la fin de l'enquête leur avait été assez pénible, car celle-ci paraissait trop longue et fastidieuse aux enquêtés. On peut remarquer en effet que l'énervement croît tout au long de l'enquête. L'indicateur de lassitude suit un profil un peu surprenant. Dès le deuxième texte, la proportion d'enquêtés atteint presque 20% mais stagne à ce niveau jusqu'à la fin. On observe pour chacun de ces indicateurs une forte augmentation entre le premier texte et le deuxième texte. En effet, le module orientation, le texte oral et le premier texte ont des formes très différentes. En revanche, les formes des textes passés ensuite sont très proches. Si dans les premiers exercices, l'effet de surprise et d'insolite crée sans doute un certain intérêt ou curiosité chez la plupart des enquêtés, ensuite, l'effet de répétition dès le second texte écrit provoque lassitude et énervement chez un nombre non négligeable d'entre eux.

Tableau 8 : réaction des enquêtés selon l'ordre des exercices

	Texte Oral	Premier texte écrit	Deuxième texte écrit	Troisième texte écrit	Quatrième texte écrit
Part d'enquêtés ayant montré des signes d'énervement	3,0%	3,9%	15,6%	17,3%	18,5%
Part d'enquêtés ayant montré des signes de lassitude	3,8%	6,3%	19,5%	19,8%	19,5%
Part d'enquêtés ayant montré des signes de difficulté	13,9%	9,3%	33,4%	35,2%	33,4%
Part d'enquêtés ayant montré des signes de stress	15,6%	12,5%	20,8%	24,5%	24,5%

Lecture : 33,4% des enquêtés en train de passer leur quatrième texte ont montré des signes importants ou légers de difficulté.

Comment varient ces indicateurs ? Il faut garder en tête que l'introduction de ces indicateurs a pour objectif de détecter les variations de motivation susceptibles d'expliquer des défauts de performance. Ces indicateurs pourraient se révéler d'autant plus utiles qu'ils sont dotés d'une variabilité suffisante. Comme on le verra dans les paragraphes suivants, plus précisément c'est la variabilité intra-individuelle, c'est-à-dire le changement de comportement pour un individu donné, c'est-à-dire à compétence donnée, qui pourrait permettre de faire la part entre motivation et compétence dans l'analyse des performances. La formule suivante rappelle la décomposition de la variance totale en variance inter-individuelle et variance intra-individuelle.

$$\underbrace{\frac{1}{NP}\sum_{i=1}^{N}\sum_{p=1}^{P}(R_{it}-\overline{R})}_{V_{tot}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\overline{R}_i-\overline{R})}_{V_{inter}} + \underbrace{\frac{1}{NP}\sum_{i=1}^{N}\sum_{p=1}^{P}(R_{it}-\overline{R}_i)}_{V_{intra}}$$

Sur ce point, le tableau 9 nous informe que généralement, la variance inter-individuelle explique la variance totale davantage que la variance intra-individuelle, même s'il demeure une part non négligeable expliquée par la variance intra-individuelle. Par ailleurs, les indicateurs a priori les plus liés au comportement de motivation ne sont cependant pas les plus variables : en particulier, 60% des effets de lassitude sont expliqués par des différences inter-individuelles et l'indicateur d'énervement est celui qui possède la moins grande variance totale.

Tableau 9 : décomposition de la variance des indicateurs de comportement

	Indicateur « énervement »	Indicateur « lassitude »	Indicateur « stress »	Indicateur « difficultés »
Variance totale	**0,11**	**0,13**	**0,16**	**0,19**
Part de la variance inter-individuelle sur la variance totale	**54%**	**61,3%**	**63,5%**	**48,9%**

Lecture : 54% de la variance totale de l'indicateur « énervement » est expliqué par la variance inter-individuelle et 46% par la variance propre pour les individus tout au long de l'enquête.

Temps et performances

Dans la suite de cette section, nous tentons d'estimer l'impact du temps de réflexion sur les performances réalisées. Si nous supposons en effet que le temps de réflexion est un indicateur de motivation, il serait logique d'observer un impact négatif des temps courts sur le nombre de bonnes réponses. Le premier paragraphe est consacré à la mesure directe de cet impact par une équation de régression logistique sans introduire d'effets fixes « individus ».

Estimation sans effets fixes individuels.

La variable temps de lecture porte sur chaque texte et sur les 4 à 6 questions qui lui sont associées. Afin de mettre en évidence l'impact du temps de lecture et de réflexion sur la performance, on estime la régression logistique suivante.

$$\text{(eq. 1)}\quad X_{ip}^{*} = a_p 1_p + bT_{ip} + e_{ip} \text{ s.c. que } e_{ip} \text{ suit une loi logistique}$$

Dans cette équation, deux dimensions sont présentes : *i* et *p*. *p* est l'indice relatif au numéro du texte et *i* désigne l'individu.

X_{ip} =0 si l'individu *i* réussit à moins de 4 questions sur les 4, 5 ou 6 questions du texte.

X_{ip} =1 si l'individu *i* réussit à 4 questions ou plus sur les 4, 5 ou 6 questions du texte.

X_{ip}^{*} est une variable latente associée à X_{ip}. Ainsi, $X_{ip}^{*} \le 0 \Rightarrow X_{ip} = 0$; $X_{ip}^{*} > 0 \Rightarrow X_{ip} = 1$

1_p désigne une variable indicatrice associée au texte *p*. Pour chaque *p*, le coefficient a_p représente la difficulté[13] de l'ensemble constitué du texte *p* et des questions associées.

Enfin, sous quelle forme faire figurer le temps dans l'équation ?
On peut en premier lieu le prendre en compte en niveau (équation (1)). Les résultats obtenus pour b sont peu significatifs (résultats non détaillés).

Une autre tentative consiste à intégrer le temps sous forme d'une variable qualitative. En l'occurrence, pour chaque texte, cinq indicatrices $q_{ip}^{1},..q_{ip}^{5}$, ont été introduites de la façon suivante. Pour tout texte *p*, on note Q_p^1, …, Q_p^5 , les fractiles de la distribution des temps correspondant respectivement aux 5%, 25%, 50%, 75 % et 100% les plus courts.

[13] Les spécialistes de la littératie (voir par exemple [8]) ont ainsi décrit la plupart des caractéristiques de textes et de questions associées qui expliquent la difficulté de résolution ; ainsi font partie de ces facteurs le caractère narratif ou argumentatif, la longueur d'un texte, sa densité, mais également le type de recherche d'informations que la question requiert.

Si T_{ip} désigne le temps de réflexion de l'individu i pour le texte p,

$$q^{1}_{ip} = 1 \text{ si } T_{ip} \leq Q^{1}_{p} \text{ et 0 sinon.}$$

$$q^{n}_{ip} = 1 \text{ si } Q^{n-1}_{p} < T_{ip} \leq Q^{n}_{p} \text{ et 0 sinon.}$$

On estime alors l'équation suivante.

$$\text{(eq. 1)} \quad X^{*}_{ip} = a_p 1_p + \sum_{n=1}^{5} b_n q^{n}_{ip} + e_{ip}$$

Les coefficients issus de l'estimation logistique de l'équation (1) sont exposés dans le tableau 10. Le résultat le plus intéressant est relatif aux coefficients $(b_n)_{n=1..5}$. La courbe de ces coefficients est descendante : les temps de lecture-réflexion les plus longs s'accompagnent de moins bons résultats (le troisième quintile est le quintile de référence). Ainsi, les personnes figurant parmi les 25% les plus lents sur une question donnée ont 60% de chances en moins (par rapport aux temps médians) de réussir plus de 4 questions sur le texte. En revanche, les temps de réflexion courts s'accompagnent de performances significativement meilleures.

Estimation avec effets fixes individuels

Les corrélations mises en évidence de la sorte ne correspondent certainement pas à des phénomènes causaux. Il est probable en effet qu'une variable non prise en compte intervienne pour expliquer simultanément la réussite (X^{*}_{ip}) et le temps de lecture q^{n}_{ip} : la compétence en littératie[14]. Ce n'est qu'à compétence égale que l'on pourrait supposer que le temps court est un indice de démotivation.
Isoler une causalité du temps sur les résultats demande donc à prendre en compte la compétence dans l'estimation. Dans le cadre de cette hypothèse, tout se passe comme si pour une compétence donnée q, il y avait un temps de réflexion optimal $T_m(q)$ qu'il faut consacrer à un texte pour bien en comprendre le sens et pour en réussir les questions posées.

Une méthode pour introduire la compétence dans l'estimation est de rajouter des «effets individuels fixes» à l'équation précédente C'est l'estimation qui est maintenant effectuée dans la suite de ce paragraphe.

$$\text{(eq. 2)} \quad X^{*}_{ip} = a_p 1_{ip} + \sum_{n=1}^{5} b_n q^{n}_{ip} + m_i + e_{ip}$$

Les m_i mesurent donc la compétence en littératie de chacun des individus. La méthode d'estimation utilise la maximisation de la vraisemblance conditionnelle avec résidu logistique[15] (Chamberlain (1984) [3] pour une présentation récente, voir Magnac (2000) [10]).
Le résultat de l'estimation est décrit dans le tableau 10 et mis en regard avec l'estimation sans effets fixes.
Les résultats confirment les effets attendus. Les coefficients $(b_n)_{n=1..5}$ sont maintenant distribués suivant une courbe en U renversé. Les temps les plus courts sont maintenant associés à de moins bonnes performances. Si un individu consacre un temps de réflexion à compétence donnée parmi les 5% les plus faibles, il aura moins de chances de réussir plus de 4 questions qu'un individu de compétence identique qui y consacre un temps médian. Ce phénomène pourrait confirmer un impact de la motivation sur les performances. On peut considérer en effet que consacrer moins de temps à un texte qu'il n'en mérite est un signe de moindre motivation ou de lassitude manifestée à l'encontre du

[14] Cette compétence dépasse la simple procédure de déchiffrage : elle sous-entend également compréhension de l'explicite et pour les plus hauts niveaux de l'implicite, etc…
[15] En l'occurrence, c'est la procédure PHREG de SAS qui a été utilisée dans cette étude.

thème d'un texte particulier ou en raison de divers phénomènes contingents (un délai horaire à respecter par exemple). Mais cela pourrait être également attribué à « une confiance en soi » excessive. Un individu qui surestime ses compétences pourrait être amené à ne pas détecter les réelles difficultés du texte et donc y consacrer moins de temps que nécessaire[16].

Tableau 10 : modélisation de la performance en fonction du temps de réflexion

	sans effets fixes	avec effets fixes
Temps de réflexion		
Tps5	0,24	-0,99
	(4,67)	(4,24)
Tps25	-0,13	-0,42
	(1,54)	(3,90)
Tps75	-0,48	-0,38
	(20,52)	(3,95)
Tps100	-0,48	-0,33
	(20,73)	(2,13)
Référence : Tps50		
Constantes Textes		
OR1	1,68	2,53
	(3855,36)	(38,95)
OR2	1,75	2,62
	(39,48)	(41,81)
OR3	3,14	4,60
	(126,82)	(99,87)
OR4	2,22	3,29
	(63,67)	(61,89)
OR5	0,82	1,39
	(8,62)	(12,38)
FS1	1,50	2,33
	(29,20)	(31,30)
FS2	3,43	5,08
	(151,51)	(95,22)
CSL1	2,59	3,80
	(86,32)	(56,22)
CSL2	1,24	1,85
	(19,91)	(16,52)
S1	1,41	2,40
	(25,59)	(26,60)
S2	0,61	1,38
	(4,86)	(8,72)
L1	2,00	2,54
	(51,42)	(31,57)
Référence : Texte L2		
Nombre d'individus : 191		

Lecture : Les chiffres entre parenthèses représentent les statistiques du Chi2 associées à chaque coefficient de la régression logistique. Rappelons que si ces statistiques sont supérieures à 4 environ, alors le coefficient est significativement non nul au seuil de 5% et si elles sont supérieures à 3,5 environ, alors le coefficient est significativement non nul au seuil de 10%.

[16] Dans ce type d'enquêtes, on observe en particulier une baisse des performances pour les hauts de niveau de formation, qui pourrait peut-être attribuée à ce type de phénomènes.

On observe également l'effet symétrique : à compétence donnée, les individus qui consacrent un temps de réflexion compris entre la médiane et le troisième quartile pour un texte donné ont significativement moins de chances de réussir plus de 4 questions que les individus de même compétence et qui y consacrent un temps compris entre le premier quartile et la médiane. Il s'agit là vraisemblablement d'un phénomène de difficultés locales. Un individu peut buter sur un exercice ou un thème particuliers. Malgré le temps plus long qu'il lui consacre, il y réalise des performances plus mauvaises. On peut observer que l'estimation avec effets fixes de l'impact des temps longs est plus faible que celle sans effets fixes. Cela se justifie par le fait que temps de réflexion et compétence sont corrélés. A texte et à questions donnés, les moins compétents consacrent plus de temps que les plus compétents. Une partie de l'effet temps longs a donc été réintégrée dans le «m_i» dans la seconde estimation.

Les indicateurs de comportement sont-ils également utilisables pour corriger de la motivation ?

Le même exercice a été réalisé en intégrant des indicateurs de comportement à la place du temps de réflexion. Dans les équations (3) et (4), C_{ip} désigne respectivement un indicateur d'énervement, de lassitude, de stress (ou nervosité au sens large) et enfin de difficultés visibles. $C_{ip}=1$ signifie que l'enquêteur a jugé que l'enquêté manifestait des signes légers ou importants liés à chacune de ces réactions émotionnelles, 0 sinon.

$$\text{(eq.3)} \qquad X_{ip}^{*} = g_{p} 1_{ip} + C_{ip} + e_{ip}$$

$$\text{(eq.4)} \qquad X_{ip}^{*} = g_{p} 1_{ip} + C_{ip} + m_{i} + e_{ip}$$

Le tableau 11 consigne les résultats obtenus. Les méthodes d'estimation sont identiques à celles retenues dans la partie précédente.

Deux résultats sont notables. La première estimation (sans effets fixes) met en évidence une forte corrélation entre les indicateurs de comportement et les performances mesurées. Précisons que l'indicateur lié aux signes de difficultés observés ne peut en toute rigueur être considéré comme indicateur de motivation : il est dans sa définition même synonyme de la variable expliquée. Il a donc été intégré seulement dans cette première estimation sans effets fixes, simplement pour confronter indicateurs subjectifs et résultats objectifs.. Pour l'indicateur de difficultés, la régression avec effets fixes n'aurait pas vraiment de sens : en effet, cette variable de comportement doit être considérée comme complètement endogène En tout état de causes, la forte corrélation constatée dans les régressions sans effets fixes montre que les performances sont bien liées avec les manifestations émotionnelles et que celles-ci sont relativement bien décrites par les enquêteurs..

La seconde estimation à l'inverse de celle sur les temps de réflexion ne met pas en évidence d'impact significatif de la colère ou de stress sur les performances. En revanche, la lassitude a un impact non négligeable sur les performances (significatif au seuil de 10%). Ce résultat appelle plusieurs commentaires.

- d'abord, il s'agit pour l'instant de données de test, qui incluent un faible nombre d'observations. L'indicateur de lassitude pourrait se révéler significatif au seuil de 5% sur un échantillon de plus grande ampleur.

• la lassitude est le comportement le plus lié avec ce que l'on cherche à mesurer, à savoir le manque de motivation. Il semble bien que le manque de motivation (qui survient notamment en fin d'enquête) ait un impact sur les performances observées. Ce résultat pourrait confirmer l'interprétation de l'impact des temps de réflexion constatés dans les paragraphes précédents.

Tableau 11 : modélisation de la performance en fonction des indicateurs de comportement

	énervement (sans effets fixes)	énervement (avec effets fixes)	stress (sans effets fixes)	stress (avec effets fixes)	lassitude (sans effets fixes)	lassitude (avec effets fixes)	difficultés (sans effets fixes)
Constante	1,31	-	1,25	-	1,29	-	1,27
	(78,48)	-	(70,12)	-	(76,30)	-	(73,40)
Constantes Textes.							
L12	-1,86	-3,53	-1,82	-3,53	-1,85	-3,54	-2,01
	(26,97)	(11,36)	(25,84)	(11,41)	(26,56)	(11,46)	(30,25)
L13	-1,05	-1,52	-1,02	-1,48	-1,07	-1,55	-1,21
	(11,01)	(10,07)	(10,33)	(9,65)	(11,31)	(10,35)	(14,04)
L14	-2,62	-3,57	-2,62	-3,52	-2,61	-3,60	-2,77
	(50,92)	(19,85)	(50,82)	(20,10)	(50,60)	(20,71)	(55,02)
L23	-1,89	-3,04	-1,93	-3,00	-1,89	-2,99	-1,96
	(32,51)	(14,68)	(33,25)	(14,66)	(32,57)	(14,94)	(34,29)
L24	-2,32	-3,11	-2,42	-3,22	-2,37	-3,27	-2,45
	(41,24)	(16,59)	(43,45)	(17,31)	(42,55)	(17,32)	(44,67)
L34	-1,93	-2,53	-1,95	-2,50	-1,98	-2,68	-2,24
	(33,57)	(18,72)	(33,79)	(18,35)	(34,69)	(19,33)	(41,28)
Q2	-17,84	-19,52	-17,93	-19,53	-17,93	-19,94	-18,06
	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)	(0,00)
Q3	1,80	2,41	1,83	2,42	1,82	2,44	1,72
	(16,61)	(14,04)	(17,11)	(14,18)	(16,93)	(14,37)	(15,11)
G1	1,07	1,71	1,10	1,76	1,06	1,68	0,96
	(9,55)	(9,95)	(9,95)	(10,60)	(9,27)	(9,52)	(7,60)
G2	-0,87	-1,07	-0,85	-1,04	-0,88	-1,10	-0,99
	(14,92)	(11,19)	(14,03)	(10,98)	(15,26)	(11,76)	(18,57)
indicateur de comportement	1,04	0,78	1,00	0,70	1,03	1,14	1,03
	(16,53)	(1,42)	(21,31)	(1,64)	(18,94)	(3,47)	(26,88)
Nombre d'individus : **326**							

Lecture : Les chiffres entre parenthèses représentent les statistiques du Chi2 associées à chaque coefficient de la régression logistique. Rappelons que si ces statistiques sont supérieures à 4 environ, alors le coefficient est significativement non nul au seuil de 5% et si elles sont supérieures à 3,5 environ, alors le coefficient est significativement non nul au seuil de 10%.

Conclusion

Deux aspects des évaluations de compétences ont été abordés ici : les techniques spécifiques d'analyses des épreuves et la prise en compte de la motivation dans la performance. A travers les exemples d'analyses donnés dans cet article, le lecteur aura pu se faire une idée de la diversité du protocole IVQ et de la bonne qualité des exercices. Cependant, nous n'assimilons pas qualité d'une épreuve et unidimensionnalité. Il importe de tenir compte de la variété des compétences mise en œuvre dans une activité comme la lecture ou le calcul.

De même, la question des biais d'items (le terme de biais est peut-être lui-même trop négatif) ne doit pas conduire à rechercher une épreuve parfaitement unidimensionnelle en supprimant tous les items « favorisant » un groupe donné, parce que sinon, l'épreuve risquerait d'être injuste. Ces biais sont donnent souvent lieu à des analyses aussi intéressante que celle du score global. Certes les tentatives d'explication ici exposées, peut-être difficiles à suivre sans les documents, montrent à quel point

l'analyse des biais d'items est délicate. Pourtant, l'apparition d'écarts significatifs, alors même que l'on travaille sur de très petits effectifs, montre qu'il y a sans doute derrière ces biais, des phénomènes intéressants à étudier, qui permettraient de dépasser la comparaison brutale des niveaux de compétences entre groupes. C'est sans doute l'enseignement à tirer de cette analyse : les biais d'items sont à peu près inévitables et en fait souvent porteurs d'une information qu'il convient de dégager pour avoir une vision plus claire des différences de compétences entre groupes sociaux.

Dans cette étude, nous avons aussi essayé de mettre en évidence l'impact de la motivation sur les performances dans les enquêtes sur les compétences. Deux types d'indicateurs ont été introduits : le premier indirect, le temps de réflexion et les seconds plus directs des indicateurs de comportement par rapport à chaque texte passé. Paradoxalement, il apparaît que c'est l'indicateur indirect qui semble manifester l'impact le plus significatif sur les performances. Parmi les indicateurs de comportement, c'est le comportement de lassitude qui possède le plus significativement un impact sur les performances, ce qui semble bien confirmer que la motivation est une composante essentielle à prendre en compte dans l'analyse de ces résultats. Si l'indicateur du temps de réflexion ressort plus significativement de cette analyse, c'est sans doute qu'il possède une plus grande variabilité que l'indicateur de comportement. Les deux indicateurs sont donc complémentaires l'un de l'autre et leur analyse conjointe sur des échantillons plus grands sera probablement féconde.

Si cette étude met en évidence un impact significatif de la motivation sur les performances, il est un point sur lequel nous n'avons pour l'instant pas avancé : comment effectivement corriger les performances de la motivation ? Est-ce que la mise en place de cette correction modifie substantiellement les profils de littératie d'une population ? Ces questions s'inscrivent dans le débat autour de la controverse suscitée par les résultats de l'enquête IALS : une des pistes d'explication des résultats français reposait sur l'hypothèse d'une moindre motivation des enquêtés français.

Bibliographie

[1] Blum A., Guérin-Pace F. « L'illusion comparative : les logiques d'élaboration et d'utilsation d'une enquête internationale sur l'illettrisme », *Population, 54 (2), pp. 271-302*

[2] Blum A., Guérin-Pace F. *Des Lettres et des Chiffres,* Fayard 2000.

[3] Chamberlain G. (1984) : « Panel Data », in *Handbook of Econometrics*, ed. by Z. Griliches and M.D. Intriligator. Amsterdam : North-Holland Publishing Co.

[4] D'Hautfoeuille X., Murat F., Rocher T., « La mesure des compétences : la logique contradictoire des évaluations internationale », Actes des journées de méthodologie statistiques des 4 et 5 décembre 2000, novembre 2002, Insee.

[5] Dickes P., & Vrignaud P. « Rapport sur les traitements des données françaises de l'enquête internationale sur la littératie » *Rapport pour le Ministère de l'Education Nationale. Direction de l'Evaluation et de la Prospective, 1995.*

[6] Insee, «Enquête méthodologique Information et Vie Quotidienne. Tome 1: bilan du test 1», Document de travail de l'Insee, série Méthodologie d'Enquête, n° 2, décembre 2002, Insee.

[7] Kalton G., Lyberg L., Rempp J.- M. « The international Adult Literacy Survey : a review of methodology » *Rapport d'expertise,* Décembre 1995

[8] Kirsch I. S., Jungeblut A., Mosenthal P. B, *"The Measurement of Adult Literacy"* in *Adult Literacy in OECD countries : Technical Report on the first International Adult Lirteracy Survey.*1998

[9] Lollivier S., . Verger D. : «L'influence de la motivation sur les performances aux tests de littéracie » *Document de travail 2002 (à paraître).*

[10] Magnac T. : « L'apport de la micro-économétrie à l'évaluation des politiques publiques » *Cahiers d'économie et de sociologie rurales* n°54,2000.

[11] OCDE – Statistique Canada : « La littératie à l'ère de l'information : rapport final de l'enquête internationale sur la littératie des adultes » *2000.*

[12] Vrignaud P. « Evaluations sans frontières : comparaisons interculturelles dans le domaine de la cognition » in M. Huteau & J. Lautrey (Eds). *Les figures de l'intelligence*. Paris : Editions et Applications Psychologiques.