# NON REPONSE PARTIELLE DANS LES ENQUETES DE MESURE DES COMPETENCES / L'EXEMPLE DE PIAAC

*Nicolas. JONAS (*), Alexandre. LEBRERE (**)*

*(*) INSEE, Division « Emploi »*
*(**) INSEE, Unité « Méthodes Statistiques »*

## Introduction

L'idée d'exprimer un niveau global de compétence d'une population donnée et de relier ce niveau à certaines variables physiologiques et sociologiques remonte au 19ème siècle. Dans le contexte eugéniste très marqué d'une Europe en plein mutation industrielle, le développement de la psychologie a fait naître le projet de tenter de caractériser les groupes sociaux « problématiques » (les ouvriers, les populations noires, les femmes, les enfants, certains groupes ethniques…) par leur niveau moyen d'intelligence. C'est dans cette optique que des travaux précurseurs comme ceux de Galton (1874) ont favorisé l'émergence d'un courant de recherche psychométrique qui voulait établir les lois héréditaires de l'intelligence. Malgré le non sens scientifique et les impasses évidentes de ces approches eugénistes, la mesure objective des niveaux macrosociologiques de l'intelligence a longtemps conservé un grand interêt dans des franges relativement larges de la population et de la communauté scientifique.

Il a fallu attendre la deuxième moitié du 20ème siècle pour que cette tadition académique soit peu à peu abandonnée grâce à l'émergence de nouvelles catgories de pensée. Sous l'influence de quelques chercheurs américains, un glissement s'est opéré vers une approche moins déterministe. La constitution progressive des sciences de l'éducation comme discipline autonome du champs universitaire américain (Herpin et Jonas, 2011), impose les concepts de compétence et d'apprentissage à la place de ceux d'intelligence et d'hérédité. En parvenant à se rendre indispensables à la définition et à l'évaluation des politiques d'éducation (Walters 2007), les spécialistes en psychométrie ont obtenu les moyens de dévellopper des outils de mesure à grande échelle et de mettre en place les premières grandes enquêtes de mesure des compétences.

Il faut dire que cette évolution a bénéficié d'un contexte international particulièrement favorable. La diffusion des approches économiques néoclassiques et la généralisation des économies de marché, ont joué un grand rôle dans la multiplication des travaux portant sur les conditions nécessaires au développement économique à long terme. Une variable (en partie exogène) joue un rôle de plus en plus important dans les projections de croissance définies par ces travaux : celle exprimant le capital humain. Dès lors, le degré de maîtrise des compétences fondamentales peut apparaître comme une approximation prometteuse de ce capital humain. C'est en tout cas dans cette optique que l'OCDE (l'organisation pour la coopération et le développement économique) a soutenue fortement l'émergence de ces enquêtes de mesure des compétences des adultes au niveau international. Le lien entre niveau de compétence et développement économique est d'ailleurs très clairement assumé dans les différentes publications de cette organisation[1].

Devant l'importance prise par ce type de problématique dans le débat public, il est donc plus que jamais nécessaire de regarder de près la méthodologie de construction de ces enquêtes. C'est l'objet de cet article qui se propose détudier le lien entre non réponse et estimation des scores de compétence. En effet, en se replaçant dans le contexte des enquêtes auprès des ménages, il ne faut pas oublier que les niveaux de compétence par pays sont établis à partir de données de nature très différentes des autres variables explicatives utilisées habituellement dans les études de projection de croissance. En particuliers, la multiplicité des comportements possibles en situation d'enquête rend discutable la définition d'un score unidimensionnel transposable et comparable à ceux obtenus dans

[1] C.f. par exemple : OCDE et Statistique Canada, *La littératie à l'ère de l'information. Rapport final de l'enquête internationale sur la littératie des adultes*, Paris, OCDE, 2000.

d'autres pays et dans d'autres contextes. La question se pose notamment pour les données manquantes en cours d'entretien, c'est-à-dire pour les non réponses partielles. C'est cette approche que nous étudierons à partir des données de l'enquête pilote PIAAC.

On revient donc ici sur l'importance de la prise en compte des non réponses partielles dans les enquêtes de mesure des compétences, surtout lorsqu'elles concernent la population adulte. Puis nous présentons plus particulièrement les logiques de l'enquête PIAAC et la nature des données collectées. En nous appuyant sur ces données, nous passons enfin en revue quelques modèles de traitement des non réponses partielles susceptibles d'être utilisés lors de l'enquête principale.

# 1. Réponses et non réponses dans les enquêtes de mesure des compétences

## 1.1. Le contexte du pilote PIAAC en France

Les enquêtes de mesure des compétences sont apparues assez tôt au niveau international, principalement sur la population des élèves, mais ne se sont vraiment institutionnalisées qu'à partir de la deuxième moitié des années 1990. Si les premières initiatives, pendant les années 50, sont à mettre au crédit de l'UNESCO, ce sont surtout deux entités qui ont piloté ces programmes de grande envergure : l'IEA (*International Association for the Evaluation of Educational Achievement*) d'abord, notamment avec les enquêtes TIMMS[2] et PIRLS[3] et l'OCDE ensuite avec l'enquête PISA[4] qui en est à sa cinquième édition et qui est devenue un outil incontournable de pilotage et/ou d'évaluation des politiques nationales d'éducation.

Toutes ces enquêtes ont démontré la faisabilité et la légitimité des mesures directes et comparatives des compétences des élèves en littératie et en numératie et ont donc encouragé la définition de projets plus ambitieux portant cette fois sur les populations adultes.

### 1.1.1. Les enquêtes internationales sur les adultes

Les premières enquêtes représentatives sur les adultes sont donc plus récentes et ont eu lieu aux Etats-Unis, d'abord sur les jeunes adultes (souvent encore en étude) avec l'enquête YALS (*Young Adult Literacy Survey* – 1986), puis sur l'ensemble des majeurs de moins de 65 ans avec l'enquête NALS (*National Adult Literacy Survey* – 1992). Ces deux premières expériences ont été déterminantes pour la suite de la période pour au moins quatre raisons :

- Ces enquêtes ne font pas référence directement aux compétences en mathématiques qui, jusque là (et pour les élèves en tout cas) constituaient le domaine d'évaluation de prédilection.
- Elles ont mis en évidence la possibilité d'effectuer des tests à grande échelle sur des adultes dans le cadre des enquêtes auprès des ménages
- Elles ont imposées une approche des compétences en terme de « littératie » (*literacy*) qui se définit comme l'ensemble des compétences nécessaires à la compréhension et à l'utilisation d'informations écrites sous des formats divers (textes, tableaux, graphiques…)
- Elles ont consacré l'institut privé américain ETS (*Educational Testing Services*) qui devient un acteur incontournable de ce type d'enquête pour la conception des épreuves psychométriques et la définition et la mise en œuvre des techniques statistiques de traitement des scores de compétences.

[2] *Third International Mathematics and Science Study* réalisée en 1995, 1999, 2003, 2007 et 2011 à la suite des enquêtes FIMS (*First International Mathematics Study* - 1964) et SIMS (*Second International Mathematics Study* - 1977)
[3] *Progress in International Reading Literacy Study*, réalisée en 2001, 2006 et 2011
[4] *Programme for International Student Assessment*, réalisée en 2003, 2003, 2006, 2009 et 2012

Fort de ces premiers succès, ETS, associé à l'Office national de statistiques canadien (*Statistics Canada*) a proposé à l'OCDE un programme d'enquête international baptisé IALS (*International Adult Literacy Survey*). Centrée sur les individus âgés de 16 à 65 ans, cette enquête s'est déroulée la première fois en 1994 dans huit pays, dont la France, puis a été menée par de nouveaux participants de sorte que, en 1997, 25 pays avaient finalement conduit cette enquête.

Les épreuves psychométriques, construites sur le même modèle que celles de l'enquête NALS, cherchaient à mesurer les capacités de lecture et de compréhension à partir de documents inspirés de la vie quotidienne, conformément aux principes de la littératie. Trois « sous-domaines » de compétence étaient évalués : la compréhension de textes quantitatifs, la compréhension de textes schématiques et la compréhension de textes suivis. Pour chacune de ces trois catégories, cinq niveaux de compétences étaient définis. Les publications de l'OCDE présentaient la répartition des personnes interrogées pour chacun des cinq niveaux de ces trois sous-domaines pour chaque pays participant. On pouvait ainsi « comparer » les performances des huit (puis des vingt-cinq) pays engagés dans le programme.

Les chiffres les plus attendus concernaient la compréhension des textes suivis dont les résultats ont été, pour le moins, surprenants. Près de 40% des français apparaissaient dans le niveau 1 de compétence, celui constitué par les personnes ne maîtrisant pas les compétences de base nécessaires à la compréhension d'un texte simple, et 75% dans les deux niveaux les plus bas. Ces données sont particulièrement alarmantes lorsqu'on pense que le niveau moyen (niveau 3) est défini comme « le niveau minimum pour répondre aux exigences du monde moderne ». La France, au même niveau que la Pologne, apparaît donc très en retrait des autres pays participants, surtout des pays anglo-saxons, en obtenant des scores sans commune mesure avec ceux attendus par les spécialistes de ces questions.

*Tableau 1 : Répartition par pays et par niveau de compétence en lecture de textes suivis*

| **Niveaux** | **1** | **2** | **3** | **4 et 5** |
|---|---|---|---|---|
| Allemagne | 14 | 34 | 38 | 13 |
| Canada anglophone | 19 | 26 | 31 | 24 |
| Canada francophone | 28 | 26 | 38 | 9 |
| Etats-Unis | 21 | 26 | 32 | 21 |
| France | 41 | 34 | 22 | 3 |
| Pays-Bas | 13 | 31 | 42 | 14 |
| Pologne | 43 | 35 | 20 | 3 |
| Suède | 13 | 23 | 36 | 27 |
| Suisse alémanique | 19 | 36 | 36 | 9 |
| Suisse romande | 18 | 34 | 39 | 10 |

Source : NCES, 1998, cité par A. Blum et F. Guérin-Pace[5]

### 1.1.2. Les conséquences de l'enquête IALS

Une fois les résultats connus, mais non encore publiés, le France, par l'intermédiaire du Ministère de l'Education Nationale, a négocié avec *Statistics Canada* un embargo sur les chiffres français. L'idée était moins d'imposer une censure à de mauvais résultats que de refuser de s'associer à des chiffres manifestement fantaisistes issus d'une enquête scientifiquement contestable, le temps, au moins, d'en comprendre les raisons.

[5] *Des lettres et des chiffres. Des tests d'intelligence à l'évaluation du « savoir lire »*, Fayard, Paris, 2000, page 79.

Très rapidement, cependant, ces chiffres ont « fuités » provoquant un scandale de grande ampleur accru par cette volonté d'embargo, interprétée d'emblée comme une volonté de dissimuler une vérité dérangeante. La presse anglo-saxonne d'abord, puis les journaux nationaux français, ont titré abondamment sur ce chiffre honteux[6], ce « chiffre noir »[7] ou cette faillite du système de formation français[8] en passant sous silence les innombrables faiblesses du protocole d'enquête et en identifiant abusivement le groupe de niveau 1 de l'enquête IALS au groupe d'illettrés.

Après avoir été interpellé à l'Assemblée Nationale sur ces mauvais résultats, le ministre de l'Education Nationale a commandé une mission d'expertise pour établir clairement les biais de IALS. Un ensemble d'éléments ont rapidement été identifiés, comme les principes d'échantillonnages, la faiblesse de la qualité de traduction des épreuves psychométriques, les pratiques de correction des exercices ou encore la mauvaise prise en compte des non-réponses. La France a fait part de ses conclusions à l'OCDE mais sans que celles-ci aient réellement été acceptées, notamment par ETS qui n'a pas souhaité remettre en cause des tests et des pratiques ayant déjà fait leurs preuves lors des expériences passées.

Malgré les réserves exprimées, l'OCDE a lancé en 2000 un nouveau programme : le programme ALLS (*Adult Literacy and Lifeskills Survey*). La France n'a pas souhaité participer à cette opération car la méthodologie adoptée était quasiment identique à celle de IALS. De nombreux autres pays se sont également abstenus. Seuls sept États (ou entités) y ont finalement prêté leur concours.

En contrepartie de cette non-participation, la France a cherché à développer ses propres outils. L'enjeu consistait toujours à comprendre ce qui n'avait pas fonctionné dans IALS, afin de réussir à se doter d'une information valide au niveau national. Le projet « Information et Vie Quotidienne (IVQ) » a vu le jour dans le cadre d'un important collectif d'acteurs dépassant la sphère technique et même la sphère publique habituelle. Y ont notamment été associés l'Insee, la Depp, la Dares, l'agence nationale de lutte contre l'illettrisme (ANLCI), le ministère de la Culture, des centres de recherche. En 2002, un premier test méthodologique a permis de valider le protocole. En 2004, une première enquête nationale IVQ a été conduite en France métropolitaine auprès de plus de 13 000 logements. Elle a totalisé 10 000 répondants. Des extensions régionales en Aquitaine, dans le Nord-Pas-de-Calais et dans le Pays de la Loire ont permis de faire des zooms sur certaines parties du territoire. L'enquête IVQ a également été réalisée en Martinique, à la Guadeloupe et à la Réunion, avec une représentativité pour chacun des Dom enquêtés. Des spécificités territoriales ont été mises en évidence. L'enquête IVQ a obtenu un assez fort retentissement dans la société civile, grâce à des articles de presse qui ont pris le temps de décortiquer les résultats. Les médias ont énormément relayé cette enquête, qu'il s'agisse du test de 2002, qui donnait des résultats sur l'ensemble de la population illettrée, ou de l'enquête de 2004, qui a permis de cibler plus précisément l'analyse sur certaines catégories de la population (comme les actifs, par exemple).

### 1.1.3. Les objectifs du pilote PIAAC

Au milieu des années 2000, un Consortium d'entreprises, toujours mené par ETS, soumet un nouveau programme à l'OCDE baptisé PIAAC (*Programme for the International Assessment of Adult Competencies)* qui sollicite donc ses adhérents pour un engagement sur cette nouvelle opération censée être reconduite tous les cinq ans. Devant les garanties apportées par l'OCDE sur un certain nombre de points délicats (standardisation de l'échantillonnage, processus de qualité des traductions, participation des pays à l'élaboration des test cognitifs…), les innovations de ce nouveau projet (utilisation de tests informatisés) et l'importance de la mobilisation (27 pays participants au premier cycle), les acteurs français, échaudés par IALS, hésitent tout de même à participer. C'est finalement une décision interministérielle qui officialise l'engagement Français en associant l'INSEE, la DEPP (Ministère de l'Education Nationale) et la DARES (ministère de l'Emploi).

Mais la participation française s'accompagne de quelques exigences de garanties. En premier lieu, l'enquête IVQ est reconduite parallèlement à PIAAC afin de disposer d'un référent stable et légitime

[6] Guillaume Malaurie, « France : l'illettrisme honteux », *L'Express*, 26 décembre 1996.
[7] « La France garde secret son illettrisme », *Libération*, 7 décembre 1995.
[8] Rémy Prud'homme, *Le Figaro*, 7 mars 1996.

concernant la mesure des compétences de la population française. En second lieu, la participation ne sera définitivement actée qu'après la réalisation et l'expertise d'un test à grande échelle.

Le pilote de PIAAC, dont certains des résultats seront discutés dans cet article, constitue ce test préliminaire. Il comporte donc en lui-même plusieurs objectifs :

- Fournir des éléments permettant d'accepter ou non une participation définitive à PIAAC
- Préparer, comme pour tous les tests des enquêtes ménages, la réalisation de l'enquête principale en mettant à l'épreuve du terrain les procédures de collecte et de traitement statistique
- Evaluer les progrès accomplis depuis IALS en matière de coordination et de standardisation internationales
- Anticiper les résultats que la France pourrait obtenir lors de l'enquête principale en appliquant les méthodes statistiques prônées par ETS

C'est dans ce dernier objectif que s'inscrit cet article en revenant sur la signification et l'importance des non réponses partielles dans les enquêtes de mesure des compétences des adultes, en général, et en analysant le poids de ces non réponses partielles dans l'établissement des scores de compétence à partir des données du pilote PIAAC, en particulier.

## 1.2. Mesurer les compétences des adultes

La prise en compte des non-réponses est un problème classique dans les enquêtes auprès des ménages mais elle prend une dimension particulière dans le cadre des mesures de compétences et plus encore lorsque la population étudiée est déjà en grande partie entrée dans la vie active.

### 1.2.1. Elèves vs adultes

Le succès d'un programme comme PISA ne suffit pas à assurer celui de PIAAC, notamment parce que le fait d'étudier des adultes plutôt que des élèves pose de redoutables questions méthodologiques et même scientifiques.

La première particularité à prendre en compte concerne la spécificité des enquêtes auprès des ménages. Lorsque l'on propose des tests aux élèves, ceux-ci sont en quelques sortes « captifs » puisque les tests sont administrés par les établissements scolaires. Il n'est donc pas nécessaire de les convaincre puisque, même s'ils refusent de jouer le jeu honnêtement, ils n'ont pas la possibilité physique de se soustraire à l'évaluation. Auprès des ménages, au contraire, la première difficulté est d'identifier le ménage échantillonné, puis de parvenir à entrer en contact avec lui, puis de le convaincre de participer. Et on ne se pose qu'en dernier ressort la question de leur réelle implication dans les tests d'évaluation. Cette distinction est d'une importance cruciale dans la mesure où elle introduit, outre les questions de définition de la base de sondage et de règles d'échantillonnage, des pratiques de collecte et des techniques de pondération et de redressement de la non-réponse. Dans le cas d'une enquête internationale, ce sont donc autant de domaines, pouvant jouer sur la signification et la significativité des données, qu'il faut pouvoir standardiser ou, du moins, contrôler pour être certain de produire des données comparables.

La seconde particularité dont il faut être conscient concerne l'acte évaluatif. Il n'est pas évident de proposer à des adultes, ayant achevé leur formation initiale depuis plus ou moins longtemps et étant à des niveaux de qualification très hétérogènes, des exercices qui, malgré tous les efforts des concepteurs, conservent une forme scolaire. Alors qu'auprès des élèves ce type d'activités détient une forme de légitimité (ou de validité « écologique ») il peut apparaître comme incongru, déplacé voir totalement artificiel dans le contexte des enquêtes ménages. Il n'est en tout cas pas « naturel ». Il peut même être vécu comme « violent » dans la mesure où il renvoie des individus, autonomes et établis, à leur expérience scolaire qui ne s'est pas toujours bien déroulée. En conséquence, la question de la « prise de distance » (*i.e.* de la motivation) des enquêtés vis-à-vis de l'enquête se pose avec beaucoup plus d'acuité.

Enfin, la troisième particularité concerne le contenu-même de l'évaluation. Il est toujours possible de s'entendre sur un socle commun ou sur un programme sur lequel évaluer des élèves de tel niveau scolaire. Bien que la question soit déjà loin d'être évidente, elle renvoie en tout cas à des apprentissages scolaires identifiables et généralisables. Pour des adultes, une telle définition est beaucoup plus problématique. Qu'est-ce qu'un adulte est censé savoir faire ? Qu'est-ce qu'un adulte est censé connaître ? Doit-on se centrer sur des compétences purement cognitives ou sur des compétences utiles dans la vie quotidienne ? Etant donnée la variété des métiers et des situations individuelles, est-ce qu'il est possible d'identifier des compétences génériques propres à la population adulte ?

L'ensemble de ces particularités a conduit à l'émergence de concepts propres aux enquêtes psychométriques destinées aux adultes. C'est ainsi que, plutôt que de parler de mathématiques, de compréhension de texte ou de logique, on fait référence à deux domaines principaux d'évaluation : la littératie et la numératie.

### 1.2.2. Les compétences en question

Anglicisme dérivé du terme *literacy*, la littératie constitue une approche « pratique » des compétences littéraires, mise en œuvre aux Etats-Unis, dans une forme plus basique, dès le XIX$^{ème}$ siècle lorsque, pendant la guerre de sécession, les candidats à l'immigration dans les états de l'Union devaient subir un test d'évaluation en langue anglaise pour s'assurer qu'ils comprenaient correctement les droits et devoirs de leur nouveau statut de citoyens américains. La littératie se définit comme l'ensemble des compétences permettant de comprendre d'extraire et de réutiliser de l'information provenant de textes écrits dans des formats utilisés communément dans la vie quotidienne (textes littéraires, brochures publicitaires, tableaux comparatifs, graphiques…). Un des sous domaines de la littératie mesuré dans IALS, la compréhension de textes quantitatifs, constitue aujourd'hui, un domaine propre d'évaluation baptisé « numératie » (*numeracy*). La numératie suppose la lecture de chiffres et la maîtrise des raisonnements mathématiques fondamentaux permettant de comprendre, d'extraire et de réutiliser des informations numériques ou chiffrées présentées sur des supports écrits usuels de la vie quotidienne (facture, étiquettes de prix…).

Ces grandes enquêtes se construisent donc désormais autour de processus cognitifs très généraux (raisonnement inférentiel, compréhension littérale…) appliqués à des supports inspirés de la vie réelle. Cette approche a un double avantage, puisqu'elle permet de s'affranchir de la définition stricte d'un programme de référence pour établir les tests des enquêtes de mesure des compétences et puisqu'elle permet surtout d'éviter le caractère trop scolaire des exercices psychométriques. On propose, en quelques sortes, des mises en situation (lire un plan de métro, repérer un numéro de téléphone…) pour emporter l'adhésion des enquêtés et pour se rapprocher, au plus près, des compétences effectivement mobilisées dans la vie de tous les jours. Mais cette façon d'aborder la question, très semblable à la technique des scénarios utilisée dans de nombreux domaines (sondages d'opinion, psychologie comportementale, micro sociologie…) présente tout de même un certain nombre d'inconvénients. Même si les exercices sont construits « comme si » il s'agissait de situations de la vie réelle, ils ne sont pas complètement semblables à ceux rencontrés dans la vie réelle. Ce léger décalage suffit souvent à révéler l'artificialité de l'interrogation. Par ailleurs, en proposant de répondre par écrit, à partir des seuls éléments fournis par l'enquêteur, le questionnaire réduit considérablement le spectre des compétences qu'un individu peut mettre en œuvre pour trouver une solution à un problème qui se présenterait dans sa vie quotidienne (le bruit, les couleurs, le comportement de l'environnement, les autres outils possibles d'aides à la décision…) et appauvrit donc la notion même de compétence.

Enfin, ces enquêtes partent tout de même d'un présupposé non négligeable. En effet, les concepteurs de ces tests se proposent de donner une image des compétences d'une population dans en littératie et en numératie. Mais on peut se demander si LA compétence dans ces deux domaines existe de façon aussi objective que des données physiologiques, ou si, au contraire, on ne mesure que des performances à un test. En effet même si on accepte, au vu de l'analyse statistique, l'existence de facteurs latents pouvant être évalués, il n'en reste pas moins que la réalisation d'un questionnaire international suppose un travail intense et détaillé de tous les pays participants pour que les exercices

retenus soient suffisamment bien construits pour que leur difficulté relative soit stable d'un pays à l'autre.

### 1.2.3. Peut-on comparer les performances ?

Mais une telle condition est-elle réalisable et, plus globalement, est-il possible de comparer des compétences entre des ères culturelles très dissemblables ? Au moins trois points délicats, voire trois limites, sont à prendre en compte.

#### *Les déterminants culturels des compétences*

Fondamentalement, les notions de littératie et de numératie considèrent qu'il existe un stock de compétences nécessaires à acquérir pour pouvoir s'intégrer et pour pouvoir répondre efficacement aux défis du monde économique moderne. Or, il est tout à fait envisageable que ce ne soient pas les mêmes compétences qui seront valorisées ou recherchées dans les différents pays. Sauf à considérer, dans une forme d'orthodoxie évolutionniste, que l'uniformisation des modes de vie ait atteint un niveau de perfection absolue, on peut penser que l'organisation économique, les rapports sociaux au sein de chaque espace culturel ou encore les processus d'intégration sociale conservent encore suffisamment de spécificités nationales pour encourager, au minimum, un travail non pas simplement de traduction mais de transposition des épreuves psychométriques pour chaque pays participant.

#### *Les biais culturels*

Sans aller aussi loin dans la remise en cause de la légitimité des comparaisons internationales, il est au moins une question à laquelle il faut être particulièrement attentif : celle des biais culturels. En effet, même si on peut s'accorder sur un certain nombre de prérequis universels nécessaires à l'intégration économique et sociale des individus, on ne peut pas nier que ces compétences de base s'expriment dans des univers très différents d'un pays à l'autre (A Blum et F. Guérin-Pace, 1999). Par exemple, dans le cas de PIAAC, pour un exercice portant sur la mesure de températures, il n'est pas neutre d'exprimer l'énoncé de cet exercice en degrés Fahrenheit. Si cette unité de mesure est usuelle dans un certain nombre de pays, ce n'est pas le cas, par exemple, en France. Or, on touche bien là un point central : la traduction pure et simple de Fahrenheit en Celsius, ou même la conversion des degrés Fahrenheit en degré Celsius ne suffirait pas à assurer l'équivalence de difficulté entre la question rédigée en français et celle, par exemple, rédigée en Américain. En effet, les ordres de grandeurs numériques utilisées dans l'énoncé ne seraient pas exactement du même ordre, ce qui pourrait jouer sur la difficulté mathématique à résoudre la question posée. Que faut-il alors penser de cet exercice portant sur la comparaison des quantités de vin consommé par habitant et par pays dont les chiffres sont énoncés en gallons (et non pas en litres) ? Autre exemple : dans un exercice nommé « Guadeloupe », le questionnaire propose un texte, dans un format de type encyclopédique, présentant quelques éléments généraux sur la Guadeloupe, en précisant, notamment, qu'il s'agit d'un territoire français et francophone. La première question est la suivante : « quelle est la langue parlée en Guadeloupe ? ». La majorité des non-français ne connaissaient pas cette île avant que cette question leur soit posée. Ces personnes vont donc rechercher la réponse correcte dans le texte et vont répondre, comme le guide de correction s'y attend : « le français ». Pour une bonne part des français, au contraire, cette question leur semblera tellement simple, que certains répondront « le créole » et leur réponse sera considérée comme fausse.

#### *Les difficultés de traduction*

Enfin, le dernier point qui rend les comparaisons internationales délicates provient de la qualité des traductions. Toutes les enquêtes internationales de mesure des compétences des adultes partent d'un questionnaire source américain rédigé en langue anglaise. Il est donc nécessaire de le traduire avant de pouvoir l'administrer. Or la traduction modifie très directement la difficulté des énoncés. Par exemple, la langue anglaise accepte volontiers les répétitions ou la voie passive. C'est beaucoup moins le cas en Français. Il faut donc trouver des synonymes dans le premier cas ou modifier la structure grammaticale dans le second cas pour traduire l'énoncé dans un français correct. Un questionnaire traduit additionne donc une multitude de petites variations qui, au final, modifient la

répartition des difficultés au sein des exercices. Ces variations sont tellement fréquentes que, même pour une même langue, les versions traduites peuvent être très différentes. Pour illustrer ce point, on peut reprendre un exemple développé par A. Blum et F. Guérin-Pace[9]. Lors de l'enquête IALS, une des questions portant sur une brochure de présentation de piscine ouverte au public était formulée en anglais de la façon suivante : « *What is the possible latest time you could enter to go swimming ?* ». En France, la traduction était : « A quelle heure a lieu la vente des derniers billets d'entrée à la piscine ? », en Suisse elle est devenue « Jusqu'à quelle heure peut-on acheter un billet pour aller se baigner ? », et au Canada, elle a été traduite par « Quelle est l'heure limite à laquelle vous pouvez entrer pour aller vous baigner ? ». Au final, 88% de suisses romans ont donné la bonne réponse, contre 84% des canadiens francophones et 80% des français.

Pour résumer, la transcription d'un questionnaire d'une langue source vers une autre langue introduit un certain nombre de variations qui interdisent de considérer les deux questionnaires comme absolument équivalents. Les difficultés de la traduction modifient obligatoirement la difficulté des questions : même à un degré faible, celles-ci deviennent ou plus faciles ou plus difficiles à résoudre et ou plus faciles ou plus difficiles à comprendre. Cette instabilité de la difficulté engendre forcément des réponses de nature très différentes d'un pays à l'autre et notamment des non-réponses (ou du moins des natures de non-réponses) différentes.

## 1.3. Le statut des non réponses

### 1.3.1. Le traitement des non réponses

Les enquêtes de mesure des compétences distinguent les différentes catégories de réponses possibles dans les exercices psychométriques. La distinction principale sépare trois catégories de réponses : les bonnes réponses, les réponses fausses et les non réponses.

Les bonnes réponses, de façon évidente, correspondent aux questions qui, conformément aux guides de correction, sont codées comme bonnes (*i.e*. comme justes). Notons tout de même que l'existence d'un guide de correction standardisé ne suffit pas toujours, loin de là, à garantir une codification uniforme des réponses. Les correcteurs possèdent tous une marge d'interprétation des règles de correction qui introduit toujours une certaine variabilité dans les appréciations des règles. Par ailleurs, ces « effets correcteurs » sont amplifiés par le nombre de nationalités et de langues de traduction du questionnaire. Par exemple, le coréen ou le japonais ont des structures grammaticales tellement différentes des langues germaniques ou latines, que les stimuli (*i.e*. les énoncés de exercices) et les réponses possibles ne peuvent pas être traduites à l'identique.

Les mauvaises réponses sont, au contraire, les réponses des enquêtés qui ne remplissent pas les conditions exigées par le guide de correction. Mais, il faut souligner que, pour qu'une réponse soit comptée comme fausse il faut que l'enquêté ait tout de même donné une réponse. Là encore, il y a une certaine marge d'appréciation. Est-ce que, par exemple, le fait d'avoir barré un exercice doit être considéré comme une réponse fausse (contraire au guide de correction) ou un refus de réponse ? Dans le cas de l'enquête IALS, on a pu noter des pratiques très différentes, même entre les régions françaises. Dans certaines régions, le fait de barrer l'ensemble des questions d'un même exercice était considéré comme un refus de répondre (donc une non réponse) et dans d'autres il était considéré comme une absence de bonne réponse (donc une réponse fausse).

La dernière catégorie comprend les non-réponses. Pour cette catégorie, les enquêtes de mesure des compétences se distinguent nettement des autres enquêtes ménages. Il convient en effet de distinguer trois sous-catégories de non réponses :

- les non réponses totales : aucun exercice n'a été rempli. Il est donc nécessaire soit de définir un programme d'imputation si on parvient à établir un modèle de non-réponse à partir des données d'échantillon ou des données biographiques collectées dans la

[9] *Op. cit*. 2000, pages 112-113.

première partie du questionnaire, soit de classer l'ensemble du questionnaire en déchet pour ne pas le prendre en compte dans l'établissement des scores de compétence.
- Les non-réponses partielles (ou finales[10]) : seul le début du questionnaire psychométrique a été rempli. Ces situations traduisent le plus souvent des cas d'abandon en cours de questionnaire. Il est en général possible d'imputer les réponses manquantes.
- Les non réponses partielles intermédiaires[11] : en cours de questionnaire, certaines questions n'ont pas été traitées.

Cette dernière catégorie est celle qui pose le plus de soucis puisque du sens qu'on lui donnera dépendra son influence sur l'établissement des scores de compétences.

### 1.3.2. L'importance de la non réponse : l'exception française ?

Pourquoi un enquêté répond-il à certaines questions et pas à d'autres ? Sous cette interrogation anodine se dissimule une difficulté irréductible de ce type d'enquête. Un des points négatifs soulignés lors des enquêtes IALS et ALLS concerne justement le traitement de cette catégorie de non réponse. Dans la logique d'ETS, une question non répondue est une question qui doit être considérée comme fausse. La méthodologie des enquêtes précédentes admettait en effet qu'à partir du moment ou un enquêté a répondu à la question *n* et à la question *n+2*, s'il n'a pas répondu à la question *n+1* alors il est normal de considérer qu'il n'est pas parvenu à trouver la bonne réponse à cette question. En conséquence la non réponse *n+1* devient une réponse fausse. Au final, les non réponses partielles intermédiaires seront toutes codées comme des réponses fausses. Un enquêté n'a donc que deux choix lorsqu'on lui propose une exercice : soit de bien répondre soit de mal répondre.

Et pourtant, une non réponse peut avoir des motivations très diverses qui, en tout cas, ne peuvent pas se réduire à de simples stratégies de contournement ou de dissimulation d'une difficulté cognitive non avouée car non avouable. Une absence de réponse peut également signifier :

- un refus de chercher une réponse. Dans ce cas l'enquêté exprime un manque de motivation, un ennui ou une lassitude vis-à-vis d'un exercice en particulier. On peut retrouver de tels exemples lorsque l'exercice est trop long (un texte littéraire par exemple), trop redondant par rapport à une question précédente, trop éloigné des préoccupations quotidiennes de l'enquêté ou pas assez légitime (trop facile, trop naïf, trop orienté…)
- une impossibilité de donner une réponse. Ce cas se rencontre lorsque l'enquêté pense avoir trouvé une autre réponse logique à la question posée que celles qui sont proposées ou alors, plus souvent, lorsque l'exercice n'est pas compris (formulation ambiguë, texte peu clair, consigne imprécise).

Dans ce deuxième cas, le phénomène peut être amplifié par les difficultés de traduction évoquées dans la partie précédente. En effet, la traduction d'un exercice peut conduire à obscurcir un énoncé, rendant sa compréhension beaucoup moins immédiate. Même si la difficulté de résolution demeure inchangée, l'implication demandée à l'enquêté n'est pas la même : il lui faut un surcroît de motivation pour chercher une réponse convenable. Mécaniquement, les non-réponses qui relèvent de cette catégorie seront donc plus fréquentes dans les pays dont la langue de questionnement sera la plus éloignée de celle de rédaction du questionnaire source. Ces pays seront alors désavantagés lors du calcul des scores de compétences puisque les mauvaises réponses seront surévaluées par rapport au niveau réel de leur population.

L'expérience de IALS a montré que la France présentait un profil tout à fait particulier vis-à-vis de ces non réponses partielles intermédiaires qui y sont plus fréquentes que dans les autres pays[12]. Deux explications peuvent être avancées. 1) d'une part, dans le cas de IALS, mais peut-être même plus généralement, la traduction de l'anglais américain au français est la traduction en langue d'alphabet latin qui produit le plus de variations par rapport aux intentions des concepteurs du questionnaire originel (soit par ce quelle a été de plus mauvaise qualité, soit parce à cause de la distance culturelle,

[10] « *Not reached* », dans la terminologie des enquêtes anglo-saxonnes.
[11] « *Omitted* », dans la terminologie des enquêtes anglo-saxonnes.
[12] Siobhán Carey, *Measuring Adult Literacy. The International Adult Literacy Survey in the European Context*, Office for National Statistics, Londres 2000.

les biais culturels sont plus fréquents, soit parce que la grammaire de texte et la grammaire de phrase sont très dissemblables de celles de l'anglais américain). 2) d'autre part, le mode d'évaluation (des cahiers d'exercices composés d'une succession de petites épreuves brèves et souvent sous forme de QCM) est très éloigné des modes d'évaluation pratiqués dans le système scolaire français mais très similaires aux pratiques d'un certain nombre d'autres pays. Il existerait une sorte « d'habitus scolaire » assez spécifique à la France qui rendrait les français moins aptes à être performants à ce type de tests (décontenancés par la présentation des tests, pas habitués à ce genre d'exercice qui suppose de devoir répondre à toutes les questions…).

Quoi qu'il en soit, cette spécificité française qui, comme on l'a vu, peut avoir une telle importance dans les résultats finaux à cause des particularités des enquêtes internationales de mesure des compétences des adultes en littératie et en numératie, justifie qu'une attention toute particulière soit accordée à l'importance, aux déterminants et au traitement des non-réponses partielles intermédiaires dans la nouvelle enquête PIAAC.

# 2. Le pilote PIAAC et l'évaluation des compétences : données et méthode

## 2.1. Présentation du pilote

### 2.1.1. Elaboration de l'enquête

L'élaboration du questionnaire PIAAC et du protocle d'enquête a suivi des règles beaucoup mieux établies que lors des enquêtes précédentes. Si la conception était prise en charge par un Consortium dirigé par ETS associant de nombreux acteurs internationaux (L'institut Tudor, l'IEA, Westat, Statistics Canada…), les orientations stratégiques du projet étaient définies à l'OCDE. Concrêtement, les représentations diplomatiques des pays participants étaient regroupées au sein d'un bureau (BPC – *Board of Participating Countries*) qui jouait le rôle d'un organe executif chargé d'arbitrer et de valider les grandes étapes de conduite du projet PIAAC et déléguant un mandat de coordinateur à l'OCDE.

Par ailleurs, chaque pays a désigné un chef de projet (NPM – *National Project Manager*) dont le rôle était de transposer les directives internationales validées par le BPC dans le contexte national de chaque pays et de rendre compte de ces adaptations au Consortium qui doit contrôler la standardisation et la comparabilité des pratiques. L'idée maîtresse de cette configuration était de garantir à la fois l'implication des pays dans le déroulement de l'enquête et à la fois, pour répondre aux critiques adressées à IALS, le respect de règles communes de bonnes pratiques statistiques.

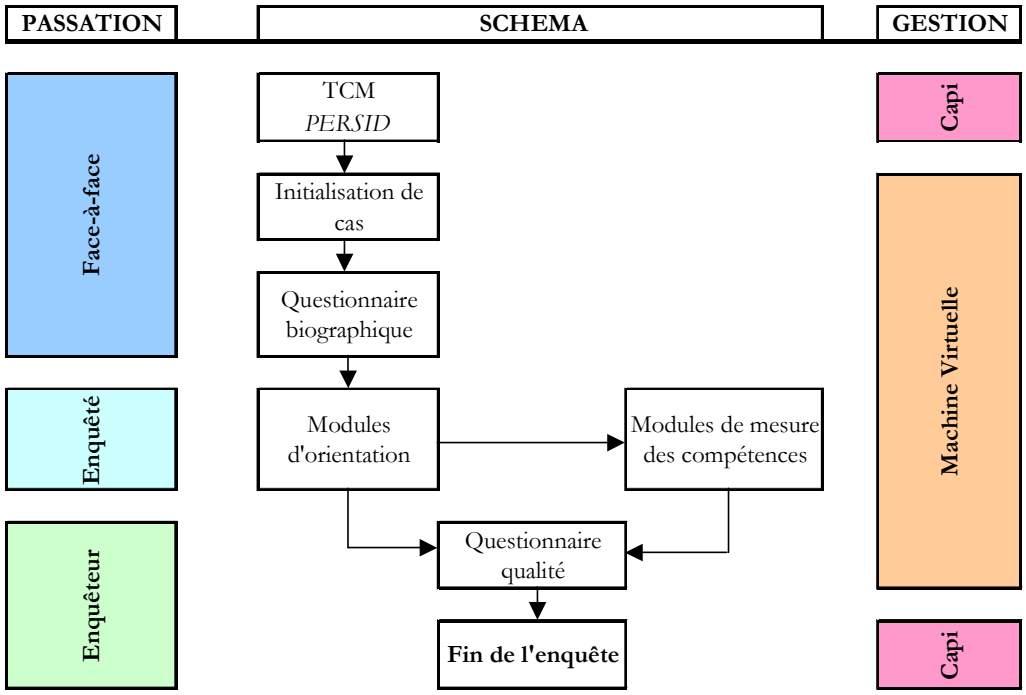
Cet effort de coordination ne s'est pas limité à la gouvernance du projet puisqu'il s'est étendu à tous les aspects techniques de conception de l'enquête. Le progrès le plus notable concerne sans doute les efforts consentis pour la qualité des traductions des épreuves psychométriques. Chaque exercice devait être traduit par deux organismes différents avant qu'une version de conciliation soit finalement adoptée. La traduction retenue était ensuite contrôlée par le Consortium qui pouvait exiger des modifications en cas d'écart significatif avec la version source.

Par contre, la conception des épreuves elles-mêmes n'a pas bénéficié de progrès aussi notables. S'il est vrai que deux groupes indépendants d'experts internationaux (un pour la littératie et un pour la numératie) ont été sollicités pour élaborer les items, le processus de conception est assez discutanle. Déjà, les pays participants n'avaient pas de droit de regard sur les items retenus ni même sur les évolutions souhaitables selon les résulatst du pilote. Ensuite, la majeur partie des experts sollicités étaient les mêmes que ceux qui avaient imaginé les exercices tant décriés de IALS et ALLS. Enfin, une partie non négligeable des exercices de PIAAC ne sont que des reprises d'exercices des enquêtes passées. L'OCDE, ETS et certains pays participants ont toujours souhaité, par ce moyen, s'assurer d'une comparabilité entre PIAAC, IALS et ALLS pour produire des analyses en évolution des compétences des adultes.

### 2.1.2. Architecture de l'enquête

Le questionnaire du pilote PIAAC se compose d'une partie biographique, administrée en face-à-face par l'intermédiaire d'un enquêteur, d'une partie psychométrique, autoadministrée par l'enquêté et une partie qualité, remplie par l'enquêteur seul (cf. schéma 1)

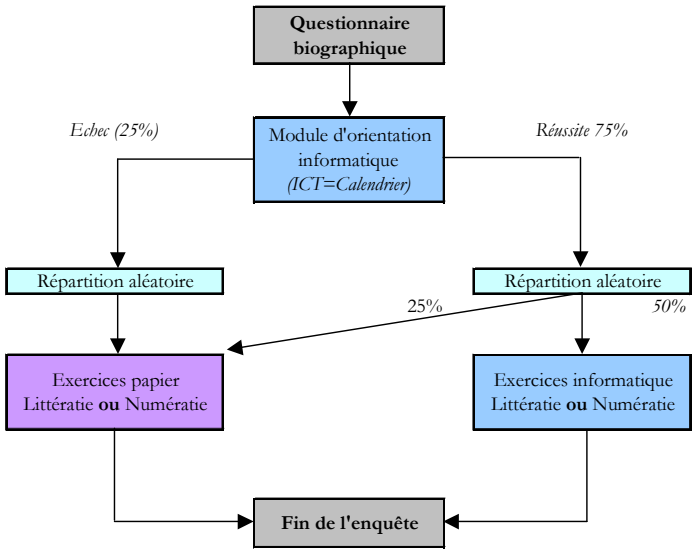
*Schéma 1 : Organisation générale du pilote PIAAC.*




La partie psychométrique propose à l'enquêté de répondre à des exercices qui serviront à estimer un niveau de compétence en littératie et en numératie. Mais l'originalité de PIAAC est de présenter de préférence, pour la première fois, ces exercices sous un format informatisé. Un logiciel dédié (une Machine Virtuelle ou VM) a été développé spécifiquement pour cette enquête. Cette VM permet de simuler un environnement virtuel semblable à celui que doivent utiliser quotidiennement la plupart des habitants des pays développés. Les possibilités ouvertes par cette VM sont très prometteuses. Les concepteurs peuvent ainsi chercher à approcher les même compétences que celles qui sont mises en œuvre lors de la manipulation de sites Internet, lors de l'utilisation de programmes bureautiques ou lors de la gestion d'une boîte mail. Un avantage supplémentaire est de pouvoir désormais s'affranchir des livrets papier pour construire des exercices beaucoup plus ludiques, d'autant que PIAAC ne se contente pas de transposer les exercices papier sur un ordinateur : des exercices nécessitant des savoir-faire spécifiques à la manipulation de l'outil informatique ont été élaborés (cliquer, surligner, déplacer des objets…). L'usage de la VM, en toute logique, devrait donc être un élément incitatif susceptible de réduire les non-réponses partielles intermédiaires liées à la lassitude ou au désintérêt.

Bien que de plus en plus généralisé, l'usage régulier de matériels microinformatiques n'est pas encore universel. Pour éviter que les résultats de l'enquête traduisent davantage la plus ou moins grande facilité de chacun à utiliser un ordinateur plutôt que les compétences en littératie ou en numératie, un questionnaire papier a été maintenu pour les personnes les moins à l'aise face à un ordinateur. L'orientation vers le questionnaire papier ou vers le questionnaire informatique se décide grâce à un module d'introduction qui permet d'estimer la familiarité des enquêtés avec les tâches les plus basiques de la manipulation informatique (cf. schéma 2). Ce module d'orientation est composé de cinq épreuves (pointer/cliquer, déplacer, surligner, utiliser une barre de défilement, ouvrir un menu déroulant). Il fallait réussir au moins 5 épreuves pour être orienté vers le questionnaire informatique. 75% des répondants français ont réussi ce module d'orientation lors du pilote.

*<u>Schéma 2</u> : L'importance du module d'orientation*



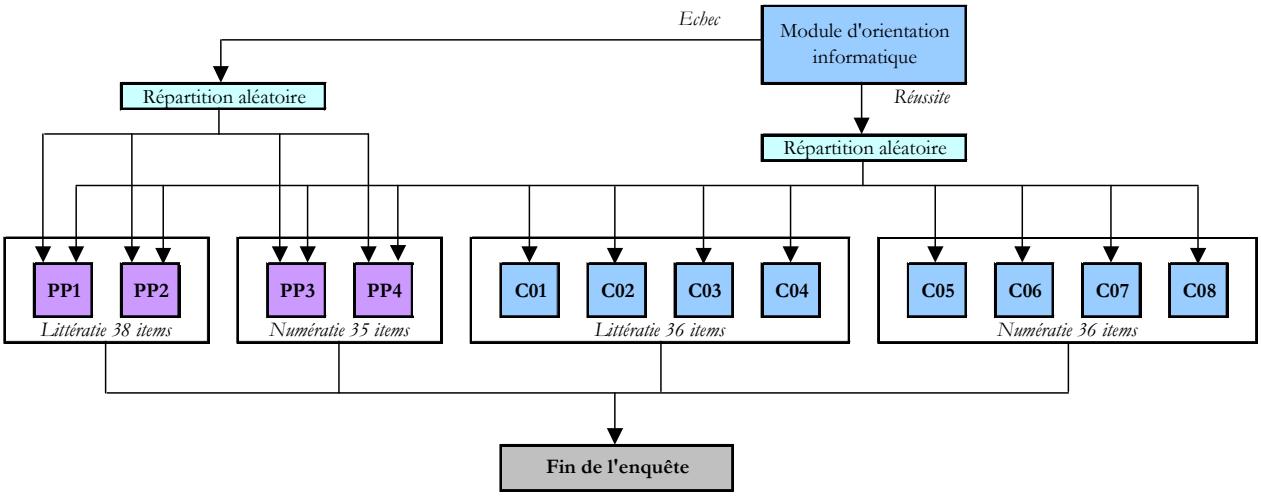

Utiliser deux supports aussi différents pour le questionnaire psychométrique pose un problème inédit : est-ce que l'usage de l'informatique « dénature » la prise d'information sur les compétences en littératie et en numératie ? Cette question est centrale dans la mesure où non seulement elle détermine la possibilité (ou non) de comparer les performances des deux sous populations séparées par le module d'orientation, mais surtout elle détermine la comparabilité de PIAAC avec les enquêtes précédentes d'ETS. De nombreux pays souhaitent en effet mesurer les évolutions du niveau général de leur population depuis IALS.

Bien que le Consortium soutient l'idée d'une neutralité du support utilisé et bien que cette neutralité soit plus ou moins démontrées dans un certain nombre d'études (Wang S. *et al.*, 2008), au moins pour les niveaux de compétences les moins problématiques (Hochlehnert A. *et al.*, 2011), l'un des objectifs du pilote PIAAC était d'évaluer les éventuels biais introduits par cette utilisation d'un double support. C'est pour cette raison qu'une partie des enquêtés (un tiers) ayant réussi le module d'orientation a tout de même été dirigée vers la version papier du questionnaire. De cette façon, il est possible de mettre en regard les résultats aux tests psychométriques de deux sous populations comparables (*i.e.* compétentes en informatique) ayant passé les épreuves sur deux supports différents.

### 2.1.3. Contenu des exercices

Comme le montre le schéma n°3, le questionnaire psychométrique se composait de 12 jeux d'exercices dont 4 sur support papier (PP1 à PP4) et 8 sur support informatique (C01 à C08). Le domaine de compétence évalué dépendait du numéro du jeu d'exercices. Les enquêtés répondaient donc soit à des questions de littératie (livrets PP1, PP2 et C01 à C04) soit à des questions de numératie (PP3, PP4 et C05 à C08). Par ailleurs, les livrets ne sont pas organisés par ordre de difficulté. Une épreuve évidente peut très bien être précédée d'une épreuve très difficile, et inversement. Cette construction est de nature à favoriser les interruptions en cas de désintérêt (quand l'exercice est trop simple) ou de découragement (lorsque l'exercice est trop compliqué). Notons enfin que tous les exercices de PIAAC ne sont pas nouveaux. De nombreux items ont été repris de IALS et de ALL pour que des travaux de comparaisons temporelles puissent être établis. Ces multiples reprises sont regrettables dans la mesure où plusieurs de ces items avaient déjà été sévèrement critiqués au moment de l'expérience malheureuse de IALS.

*<u>Schéma 3</u> : Composition des cahiers du pilote PIAAC*




Pour assurer la comparabilité entre livrets, chaque exercice était forcément présent dans au moins deux livrets. Par ailleurs, pour renforcer la comparabilité entre supports, un certain nombre d'exercices étaient communs au papier et à l'informatique. Mais ces items communs aux deux supports présentaient tout de même quelques divergences, notamment dans les modalités de réponse proposées. Par exemple, on pouvait demander sur papier, selon la préférence du répondant, d'écrire, de souligner ou d'entourer la bonne réponse, alors que, sur la VM, l'enquêté ne pouvait que surligner la bonne réponse. La pertie informatique réclame donc un surcroît d'investissement de l'enquêté qui est davantage contraint que sur papier (Noyes J*. et al.,* 2004).

La dernière remarque que l'on peut faire sur cette architecture globale du questionnaire concerne la place relative du module biographique. Pour l'enquête IVQ, le choix a été fait de le placer après les épreuves psychométriques alors que, pour PIAAC, c'est l'option inverse qui a été choisie. Les raisons de ce choix s'expliquent surtout par les propriétés méthodologiques de PIAAC. Comme le questionnaire psychométrique est organisé par cahiers tournants, c'est-à-dire que tous les enquêtés ne passent pas les mêmes épreuves ni d'ailelurs le même type d'épreuves (littératie ou numératie), il est nécessaire de recueillir le plus tôt possible des données biographiques pour imputer des réponses en cas d'abandon. Par ailleurs, dans la version finale de l'enquête, une partie du questionnaire psychométrique sera filtrée par certaines données du module biographique (le niveau d'étude, la langue parlée à l'enfance et l'usage de l'ordinateur au travail et au domicile). Dans IVQ, au contraire, on a estimé qu'il était préférable de placer le module biographique en fin de questionnaire pour que les enquêtés répondent aux exercices dans un meilleurs état de fraîcheur, ce qui permet non seulement d'éviter une dégradation des performances mais surtout de limiter le risque de pertes de données psychométriques (moins fatigués, les répondants ont moins tendance à se lasser de l'enquête et donc ont moins de chance d'interrompre les exercices en cours de route). Malgré les demandes de la France, soutenue par d'autres pays comme la Norvège, soucieuse toujours de restreindre les non réponses partielles intermédiaires, cette option n'a jamais été sérieusement envisagée par le Consortium.

Au final, l'architecture globale de l'enquête PIAAC présente quelques points susceptibles de limiter les effets de lassitude trop peu pris en compte dans les enquêtes précédentes (grâce à l'usage de l'informatique et grâce au module d'orientation), mais qui sont en partie contrebalancés par quelques défauts structurels dont il faut étudier la porté : la reprise d'exercices de IALS, l'utilisation de deux supports de nature différente, l'absence de progressivité dans la difficulté des épreuves, la place du questionnaire biographique… Autant d'éléments qui, associés à la longueur du questionnaire et aux mauvaises fonctionnalités de la VM, ne favorisent pas l'adhésion des répondants à l'enquête.

## 2.2. Réalisation du pilote

L'utilisation novatrice d'une Machine Virtuelle a obligé les acteurs nationaux à réaliser de lourds investissements sur le matériel de collecte (micro-ordinateurs compatibles) et sur les formations. De plus, pour prendre en compte certaines des critiques adressées au protocole de l'enquête IALS, un effort conséquent de standardisation des pratiques d'enquête a été entrepris. En particulier, le plan de sondage devait répondre à des normes très strictes, contrôlées par le Consortium de PIAAC.

### 2.2.1. Echantillonnage

Le champ de l'enquête est composé des personnes âgées de 16 ans à 64 révolus au moment de l'enquête (donc nées entre le 01/04/1994 et le 01/06/1945), résidant en France métropolitaine et n'habitant pas en « communauté institutionnelle » (*institutional collective dwelling units or group quarters*). Cette définition exclut donc du champ de l'enquête les prisons, les hôpitaux, les foyers de placement et les bases militaires. Elle inclut cependant les résidences universitaires, les internats, les communautés religieuses et les foyers de travailleurs. Les individus présentant des handicaps auditifs ou visuels lourds, et plus généralement toute difficulté physique incapacitante, sont exclus du champ de l'enquête.

L'échantillon, composé de 4 000 individus, a été tiré dans 7 régions métropolitaines : Languedoc-Roussillon, Centre, Île-de-France, Pays de la Loire, Lorraine, Nord Pas de Calais et Rhône-Alpes. Ces régions ont été choisies de telle sorte que la proportion régionale de ménages équipés d'ordinateurs (estimée à partir de l'enquête TIC) soit suffisamment variée (régions très équipées, régions peu équipées), à partir du Fichier de l'Impôt sur le Revenu (FIP). Par ailleurs, elles cumulent plus de 50 % de la population métropolitaine et offre des profils de population diversifiés. Le tirage s'est effectué au niveau individuel (et non pas au niveau du ménage).

Certains éléments de ce plan de sondage n'ont pas été repris à l'identique par tous les pays participants. Les normes du Consortium portaient surtout une obligation de résultats en terme de qualité de l'échantillonnage (taux de couverture de la population, nombre de degré de tirage, échantillon équipondéré…), et non pas sur une obligation de moyens. Cette contrainte indirecte s'explique par la diversité des pratiques et des traditions nationales : recourt à un registre, aux données du recensement, aux données fiscales…

### 2.2.2. Bilan de collecte

La collecte s'est finalement déroulée en France du 19 avril au 24 juin 2010, mobilisant 182 enquêteurs formés pendant 3,5 jours en groupes restreints. Sur les 3 998 personnes échantillonnées, 2 331 ont au moins commencé les exercices psychométriques, soit un taux de réussite relativement faible à 58 % en raison du nombre important de déchets. En effet, la base de sondage, expérimentale, a produit un nombre important de logements vacants. Par ailleurs, de nombreuses personnes ont refusé l'enquête en raison de sa lourdeur. Enfin, il était difficile de suivre les déménagements, le temps de collecte étant limité à deux mois.

En revanche, le taux d'abandon est relativement faible à 10 % malgré un temps de passation de l'enquête compris entre 1 et 3 heures pour 91 % des cas, avec un temps moyen de 1 heure 50 pour le questionnaire informatisé. Le temps de passation est généralement plus faible pour les personnes orientées vers le support papier.

*Tableau 2 : temps de passation des questionnaires*

| | **Médiane** | **Moyenne** |
|---|---|---|
| Enquêtes avec exercices informatiques | 1h35 | 1h50 |
| Enquêtes avec exercices papier | 1h22 | 1h38 |

Source : Pilote PIAAC, 2010

L'enquête a donc été bien acceptée par les enquêtés malgré la durée du questionnaire. Elle l'a été également par les enquêteurs malgré l'utilisation d'un outil informatique peu optimisé. En revanche, nous avons rencontré des difficultés informatiques récurrentes pendant la collecte. Nous avons notamment perdu 224 des 1 033 exercices informatiques, soit un taux de « bug » de 22 %, les dysfonctionnements de la machine ayant rendu ces fichiers illisibles, vides ou corrompus. Le taux de bug est moins important dans les pays nordiques qui bénéficiaient de registres et n'avaient donc pas à importer de données d'échantillonnage, ce qui réduisait les interactions entre systèmes. Au Canada, il s'élève à 40 %.

En revanche, la France obtient les meilleurs résultats en termes d'entretiens complets et de taux de réponse après l'Irlande dont le test portait sur un échantillon très faible. Seuls quatre pays ont obtenu un taux de réponse supérieur à 50 % alors que l'OCDE avait fixé un objectif de 70 % au début du pilote, taux en deçà duquel les résultats risquaient de ne pas être publiés sauf à proposer une analyse solide du biais d'échantillonnage.

Ces bons résultats français, au moins en ce qui concerne les taux de réponse peuvent surprendre si on considère la diversité des pratiques de collecte. En effet, suivant les recommandations de l'OCDE, de nombreux pays ont fourni des gratifications aux enquêtés pour limiter les refus et pour améliorer l'investissement personnel des répondants (sous forme d'argent, le plus souvent). En France, de telles pratiques ne sont pas autorisées et ne sont sans doute pas souhaitables étant donné le biais non contrôlé qu'induisent les gratifications pécuniaires. Elles n'ont de toute façon pas été nécessaires. En terme de collecte, c'est sans doute le professionnalisme du réseau d'enquêteurs qui a permis d'aussi bons résultats.

## 2.3. Réponses et non réponses : l'exemple de la littératie

### 2.3.1. Résultats généraux

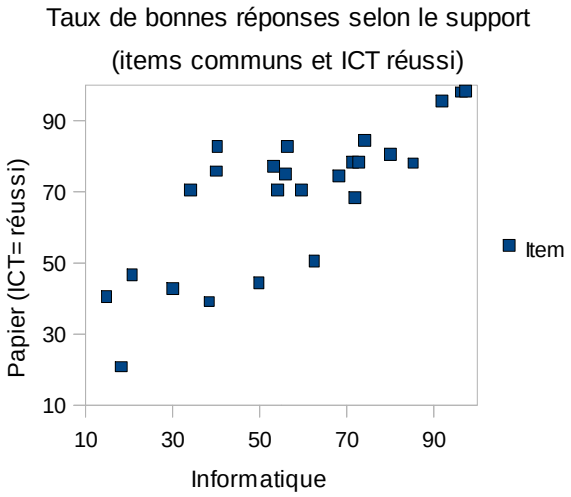
Le taux de bonnes réponses est très variable d'un item à l'autre sur support informatique. La question la plus difficile n'obtient par exemple que 3,12% de bonnes réponses. L'item le plus simple obtient quant à lui 97,61% de bonnes réponses. On peut d'ores et déjà noter que les items les moins bien notés ont tous un point commun : la réponse à ces items faisait intervenir un surlignage. Au moins trois difficultés relatives au surlignage ont largement pesé sur ces mauvais scores :

- Le principe du codage des bonnes réponses n'était pas précis (et parfois faux). Pour chaque item, le concepteur doit définir une zone de bonne réponse minimale, une zone de bonne réponse maximale et une zone de mauvaise réponse. Tout empiètement sur cette zone de mauvaise réponse conduit à une mauvaise réponse. Par exemple, si un répondant sélectionne un mot correct pour la bonne réponse mais qu'il induit un espace (avant ou après ce mot) dans son surlignage qui appartient à la zone de mauvaise réponse, sa réponse sera codée fausse.
- La technique du surlignage n'est pas intuitive. Elle diffère nettement de celle utilisée sur les logiciels de bureautique usuels (Word, Openoffice, Notepad…)
- Même sur Internet, le surlignage n'est pas un mode de réponse utilisé dans la vie courante. Il peut nous être demandé de cliquer, de sélectionner, de cocher ou de remplir une zone texte ou une zone numérique, mais à aucun moment il ne peut être demandé (dans un formulaire ou une interrogation scolaire) de surligner une bonne réponse sur support informatique. Il ne s'agit donc pas d'une pratique courante et commune de validation d'une réponse dans la vie quotidienne.

Le taux de bonnes réponses est plus homogène sur support papier, avec une borne inférieure à 18,28% et une borne supérieure à 98,14%. L'item qui ne compte que 18% de bonnes réponses se détache largement des autres items.

Comme le montre le graphique suivant, à niveau de compétence égal, les bonnes réponses sont presque systématiquement plus fréquentes sur papier que sur informatique.

*Graphique 1 :*




Outre la question de mécanismes cognitifs différents qui sont mis en œuvre lors d'une lecture sur papier ou sur un écran, il est probable que les difficultés engendrées par la Machine Virtuelle posent un réel problème de comparaison entre les deux supports. Il faut noter que le Consortium est conscient de ce problème et que la plupart des bugs (au moins ceux concernant le mauvais codage des réponses, les bugs informatiques et le temps de réaction de la VM) a déjà été en partie corrigée en vue de l'enquête finale. Quoi qu'il en soit, on peut au moins regarder le comportement des exercices repris des enquêtes précédents dans le protocole de PIAAC.

Les items communs entre IALS et PIAAC sont bien moins nombreux que les items communs ALLS/PIAAC, puisqu'on n'en compte que 12. Par ailleurs, seuls 6 d'entre eux servaient à établir l'échelle de littératie (*compréhension de textes suivis* ou échelle *Prose*) et les 6 autres ont alimenté l'échelle de compréhension des textes schématiques. Aucun item de textes « au contenu quantitatif » n'a été repris pour PIAAC. Par ailleurs, si ces 12 items sont bien présents sur support informatique, on n'en retrouve que 5 sur support papier, dont 3 relevant de la « compréhension de textes suivis » (*cf.* tableau 3).

*Tableau 3: Caractéristiques dans IALS des items de IALS repris dans PIAAC*

| EXERCICE | Item IALS | ECHELLE | | HIERARCHIE | | |
|---|---|---|---|---|---|---|
| | | Suivis | Schématiques | Suivis | Schématiques | Total |
| **Elections** | *COREQ2S1* | 89 | | 5 | | 10 |
| **Dépannage** | *B1Q1S1* | 60 | | 2 | | 6 |
| | *B1Q2S1* | 74 | | 3 | | 8 |
| **Canco/CIEM** | *B1Q10S1* | | 68 | | 5 | 7 |
| | *B1Q11S1* | | 23 | | 2 | 2 |
| **Distance Mexique** | *B5Q12S1* | 53 | | 1 | | 4 |
| **Femmes enseignantes** | *B7Q1S1* | 93 | | 6 | | 11 |
| | *B7Q3S1* | 80 | | 4 | | 9 |
| **Contact employeur** | *B7Q10S1* | | 56 | | 4 | 5 |
| | *B7Q11S1* | | 21 | | 1 | 1 |
| **Aspirine** | *B4Q1S1* | | 94 | | 6 | 12 |
| | *B4Q2S1* | | 49 | | 3 | 3 |

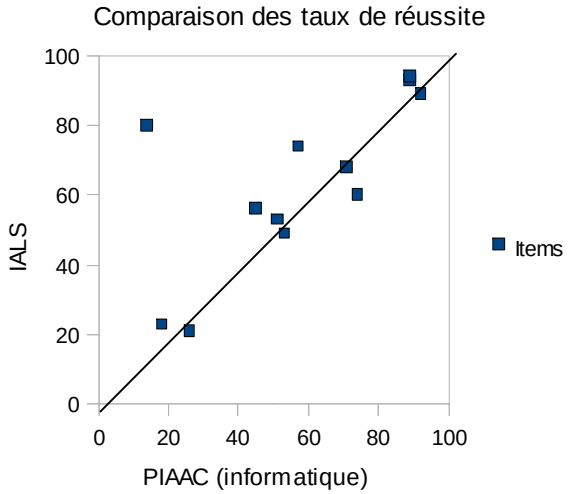
Les mêmes éléments sont calculés pour PIAAC en distinguant le support sur lequel les items ont été posés (tableau 4).

*Tableau 4: Caractéristiques dans PIAAC des items de IALS repris dans PIAAC*

| | Item PIAAC | | SUPPORT | | HIERARCHIE | |
|---|---|---|---|---|---|---|
| | PAPIER | INFORMATIQUE | P | I | P | I |
| **Elections** | *N302C02* | *D302C02S* | 92 | 92 | 5 | 12 |
| **Dépannage** | *N314101* | *D314101S* | 75 | 74 | 4 | 9 |
| | *N314102* | *D314102S* | 72 | 57 | 3 | 7 |
| **Canco/CIEM** | *N306110* | *D306110S* | 65 | 71 | 2 | 8 |
| | *N306111* | *D306111S* | 18 | 18 | 1 | 2 |
| **Distance Mexique** | | *D315512S* | | 51 | | 5 |
| **Femmes enseignantes** | | *D311701S* | | 89 | | 10 |
| | | *D311703S* | | 14 | | 1 |
| **Contact employeur** | | *D304710S* | | 45 | | 4 |
| | | *D304711S* | | 26 | | 3 |
| **Aspirine** | | *D307401S* | | 89 | | 11 |
| | | *D307402S* | | 53 | | 6 |

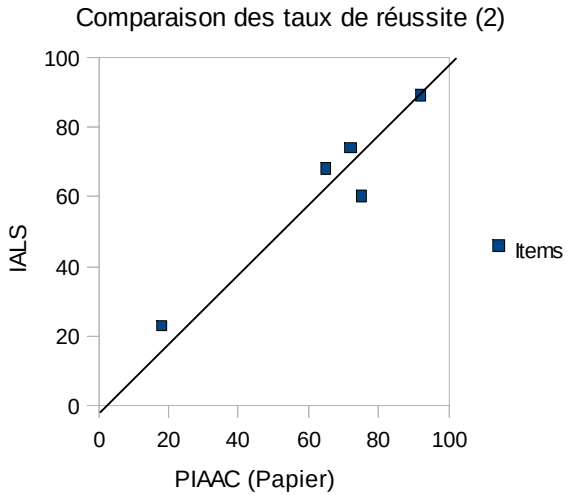
On peut maintenant comparer les résultats des deux enquêtes en ne conservant que le support informatique. On remarque que les résultats sont, sauf pour un item, très semblables entre les deux opérations. Mais comme les items sont posés sur support informatique dans PIAAC, les résultats sont sans doute biaisés par les difficultés de manipulation de la machine virtuelle. Mais, en sens inverse, ces items ont été traités par des individus ayant réussi le module d'orientation, donc possédant un niveau de compétence moyen supérieur à celui de l'ensemble de la population étudiée, biaisant vers le haut les résultats.

*Graphique 2:*




On reproduit la même comparaison en se limitant cette fois-ci aux 5 items du support papier commun avec IALS. Cette fois encore les résultats sont très comparables entre les deux enquêtes, bien que les personnes les moins compétentes soient sur représentées parmi les répondants au support papier.

*Graphique 3:*




Au final, les items issus de IALS produisent des résultats assez comparables dans PIAAC à ceux de l'enquête précédente. Mais on ne peut pas en déduire pour autant les performances françaises seront aussi mauvaises en 2013 qu'elles l'étaient en 1996. La construction des scores de compétences dépendront surtout : des caractéristiques des nouveaux items, du traitement des non réponses partielles et, plus que tout, du comportement de l'ensemble des items dans les autres pays participants.

### 2.3.2. Comparaisons internationales

Malheureusement, les « bons » résultats français en terme de collecte, n'ont pas pu être confirmés sur le sujet des performances aux tests psychométriques. Le pilote PIAAC n'était en effet pas destiné à produire des données par pays, mais seulement des données très agrégées par item. A cause de ce manque d'information il n'est pas possible de mettre en regard le fonctionnement des items en France (notamment ceux repris de IALS) avec leur fonctionnement dans d'autres pays et donc d'anticiper correctement les résultats de l'enquête principale.

L'OCDE a cependant accepté de fournir à la France quelques informations générales sur l'importance des non-réponses partielles intermédiaires, reprises ans le tableau suivant.

*Tableau 5: Proportion de non-réponses partielles intermédiaires*

| | | Informatique | Papier |
|---|---|---|---|
| **Littératie** | France | 9,31 | 9,22 |
| | *Moyenne* | *7,14* | *3,57* |
| **Numératie** | France | 10,40 | 10,45 |
| | *Moyenne* | *7,13* | *5,99* |

Ces résultats partiels confirment la spécificité française : le taux de non réponses intermédiaires y est bien plus élevé (surtout sur support papier) que dans les autres pays participants. Finalement, l'introduction de l'informatique ne semble pas avoir changé fondamentalement le comportement des répondants dans ce domaine. Face à l'ampleur du phénomène, la question de l'influence du traitement statistique des non réponses partielles intermédiaires sur l'établissement des scores de compétence se pose sérieusement dans l'objectif de parvenir, lors de l'enquête principale, à une estimation raisonnable de la distribution des compétences au sein de la population française.

# 3. Traiter les non réponses partielles : résultats et discussion

## 3.1. L'établissement des scores de compétence

### 3.1.1. Présentation du modèle MRI à deux paramètres

L'analyse des données issues des enquêtes de mesure de compétences est souvent faite en utilisant des « modèles de réponse à l'item (MRI) ». Ces modèles sont des modélisations économétriques de la réponse aux exercices proposés prenant en compte des caractéristiques de la personne répondante et des caractéristiques des exercices (les *items*). La modélisation économétrique permet de dépasser la simple description des taux de réussite (qui était entreprise dans la « théorie classique des items ») et d'analyser la covariation des caractéristiques du questionnaire des questions et des compétences individuelles.

Le principe d'un modèle de réponse à l'item est qu'une personne donnée a une probabilité bien définie de répondre correctement à un item de caractéristiques données. On donne ainsi une mesure de la probabilité que la réponse à l'item *i* par la personne *j* conditionnellement à l'ensemble des caractéristiques de l'individu et de l'item. A partir de ce principe, de nombreuses variantes sont possibles. La plus simple est souvent appelée modèle de Rasch, et prend en compte une variable de compétence individuelle et une variable de difficulté de l'item.

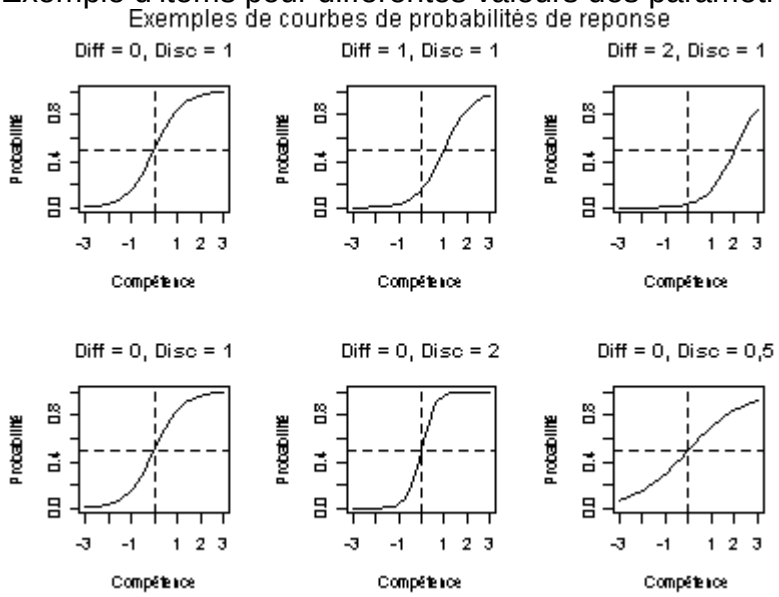
Le modèle que l'on utilise ici peut être considéré comme un développement du modèle de Rasch. Il prend en compte une caractéristique individuelle de compétence, mais deux paramètres d'item. L'un de ces paramètres sera un indicateur de difficulté de l'exercice. L'autre est appelé discrimination de l'item, et représente la capacité de l'item à séparer finement les populations respectivement des personnes de capacité inférieure et de capacité supérieure à la difficulté de l'item. Plus concrètement, un item est fortement discriminant si une personne ayant une compétence faiblement inférieure à la difficulté de l'item n'a qu'une probabilité faible de résoudre correctement l'exercice, alors qu'une personne de compétence légèrement supérieure à la difficulté donnée aura une forte probabilité de répondre correctement.

La formulation précise du modèle s'énonce comme suit. Notons *j* l'indice des personnes et *i* l'indice des items du questionnaire. Notons d'autre part $x_{ij}$ la variable égale à 1 si la réponse de la personne *j* à l'item *i* est correcte et 0 sinon. Notons enfin $\theta_j$ le paramètre de compétence individuelle de l'individu *j*, $b_i$ et $a_i$ respectivement les paramètres de difficulté et de discrimination de l'item *i*. Alors, la probabilité que la personne *j* réponde correctement à l'item *i* conditionnellement aux paramètres individuels et d'item est donnée par :

$$P(x_{ij} = 1/\theta_j, a_i, b_i) = \frac{1}{1+\exp(-Da_i(\theta_j - b_i))}$$

Ce modèle est appelé modèle logistique à deux paramètres. D'après l'expression précédente, on voit que les paramètres de difficulté et les paramètres de compétences individuelles sont mesurés sur la même échelle. Ces paramètres doivent être estimés simultanément. On remarque d'autre part que ce type de modèle suppose que le trait modélisé est unidimensionnel. L'unidimensionnalité peut être vérifiée par des mesures de cohérence interne du questionnaire. Le graphique 4 montre quelques exemples de courbes présentant la probabilité de réponse suivi le paramètre de compétence individuelle sachant quelques valeurs des paramètres de discrimination et de difficulté des items.

Graphique 4 : Exemple d'items pour différentes valeurs des paramètres




L'estimation est faite par maximum de vraisemblance. On note que sans hypothèses supplémentaires le modèle précédent n'est pas identifiable. L'hypothèse utilisée est alors une hypothèse sur la répartition de la compétence dans l'ensemble de la population sous-jacente à l'enquête. On suppose que le trait $\theta$ est réparti dans la population comme dans un échantillon de loi normale centrée réduite. L'estimation donne alors une liste des paramètres d'item d'une part, et une liste des paramètres individuels de compétence d'autre part. Les deux échelles de difficultés sont reliées par le fait qu'un individu de niveau de compétence donné répond correctement à un item de même difficulté avec une probabilité 0,5.

Après cette étape d'estimation, il reste l'étape délicate de l'interprétation des résultats. En effet, la méthode présentée ci-dessus permet d'assurer la cohérence des échelles des difficultés et des compétences, mais cette échelle elle-même est arbitraire et dépend de l'hypothèse faite sur la population. Afin de pouvoir donner un sens à l'estimation, il est nécessaire d'avoir une information extérieure permettant de fixer le sens de l'échelle de difficulté.

La possibilité de changer l'échelle des scores de compétence provient de propriétés de linéarité du modèle. Plus précisément, considérons une transformation linéaire du paramètre de compétence $\theta$ :

$$\theta' = M\theta + X$$

si l'on transforme simultanément les paramètres d'item comme :

$$a_i' = a_i / M$$

$$b_i' = Mb_i + X$$

on obtient alors :

$$P(x_{ij} = 1/\theta_j', a_i', b_i') = P(x_{ij} = 1/\theta_j, a_i, b_i)$$

Cette propriété permet d'étalonner le résultat de l'estimation sur une échelle de compétence obtenue par des informations externes.

Une première méthode d'étalonnage est celle qui a été suivie dans l'enquête IALS : il s'agissait pour un groupe d'experts de définir des items et de leurs associer des difficultés à partir d'estimations et d'études antérieures. On a ensuite décidé d'une échelle dite « de présentation », qui donnait les résultats finaux de compétence sous la forme d'un nombre entre 0 et 500. D'autre part, en même temps que les difficultés des items ont été définies des classes de scores permettant de repérer des groupes ayant des compétences homogènes et dont les compétences atteintes étaient interprétables socialement.

Une seconde possibilité se présente lorsqu'une première estimation de certains items a été obtenue précédemment. Cette solution sera retenue pour l'enquête PIAAC. En effet, un certain nombre des

items retenus ont été utilisés (une ou plusieurs fois) dans des enquêtes précédentes et comparables de mesure des compétences. La comparabilité entre enquêtes, et en particulier avec IALS, est un des objectifs principaux de PIAAC. Les items pour lesquels des valeurs sont disponibles peuvent être utilisé dans une estimation contrainte, ou encore servir *a posteriori* pour « ancrer » la nouvelle estimation sur l'item moyen de l'estimation précédente.

### 3.1.2. Quelques conditions spécifiques à PIAAC

L'enquête PIAAC se distingue par un certains nombre de traits présentant des difficultés particulières. Ces difficultés reposent sur les hypothèses de comparabilité qui ont été faites. En premier lieu, la comparabilité dans le temps est un point de principe de l'enquête. En particulier, comme vu *supra*, on suppose que les valeurs des items retenus dans l'estimation finale et communs avec les enquêtes précédentes pourront être repris sans modification. D'autre part, la comparabilité entre pays est à la base même de cette enquête. Lors de l'enquête IALS, les résultats Français avaient été considérés avec une certaine circonspection. D'autre part, toujours dans le cas Français, on note que des valeurs d'items vont être reprises bien que la France ne fît pas partie de l'échantillon de l'estimation précédente.

Enfin, l'hypothèse de comparabilité entre supports va être retenue pour l'estimation finale. Cette contrainte est très spécifique à PIAAC et est l'une des avancées principales de l'enquête. On suppose ainsi que les items sont identiques entre une passation avec carnet papier et une passation informatique. C'est une hypothèse forte, qui peut n'être pas vérifiée. En particulier, dans le cadre de l'enquête pilote, de simples raisons pratiques peuvent influencer les résultats : lire de longs texte sur l'écran d'ordinateur peut être fastidieux, voire difficile. D'autre part, l'action de répondre à la question par exemple en entourant une partie du texte ou en la surlignant sur support informatiques peut mettre en jeu des compétences différentes. Enfin, l'objectif spécifique d'évaluer les compétences dans un monde dans lequel les nouvelles technologies de l'information et de la communication sont très développées ne s'applique *stricto sensu* qu'à l'interrogation informatique. Le rôle de l'interrogation sur papier a pour fonction principale l'ancrage avec les enquêtes plus traditionnelles. Cette hypothèse peut être introduite par des contraintes dans l'estimation.

On note aussi que PIAAC fait intervenir deux dimensions de compétences : littératie et numératie. Le découpage était différent dans les enquêtes précédentes, et en particulier dans IALS.

## 3.2. Les déterminants de la non réponse individuelle

On note dans l'interrogation une part assez importante de non réponses individuelles. C'est-à-dire que les personnes interrogées ne répondent pas à certaines questions. Cette non réponse peut-être liée à différents facteurs, dont deux semblent les plus importants :

- D'une part, il y a un effet de compétence qui amène les personnes peu compétentes à ne pas répondre à des questions perçues comme difficiles. En d'autres termes, les personnes interrogées préfèrent ne pas répondre plutôt que de donner une mauvaise réponse. Cependant, cet effet n'est pas univoque : les personnes ayant les compétences les plus élevées (et qui donc « répondent le mieux » aux exercices) ont tendance à préférer comparativement plus cet effet d'évitement que les personnes de plus faibles compétences (pour qui le nombre de questions « difficiles » est plus élevé et donc pour qui l'interrogation est plus uniformément complexe). Ceci accentue le différentiel de mauvaises réponses observées par rapport à la situation réelle. De plus, cela pourrait signifier que la non réponse n'est pas interprétable de manière identique pour tous les niveaux de compétences. De plus, notons que dans l'objectif de comparabilité internationale de l'enquête, ceci peut poser un problème dû aux différences d'habitudes scolaires entre les pays. Ceci avait d'ailleurs été relevé dans les études sur l'enquête IALS : en particulier, en comparaison à la France, le système anglo-saxon valorise beaucoup plus les réponses partielles et ces élèves sont beaucoup plus habitués au type d'interrogation « à choix multiples ».
- D'autre part, la non réponse est liée à un effet de lassitude, que l'on peut relier au trait plus général de « motivation » pour l'enquête. A cet égard, il faut noter que l'interrogation de l'enquête pilote était très longue, et pouvait entraîner une fatigue non négligeable. Deux

conséquences en résultent : une propension plus élevée à arrêter le questionnement d'une part et à répondre à chaque question particulière d'autre part, et une difficulté plus grande à répondre (un taux d'erreurs plus élevé). Le questionnaire de littératie peut aussi présenter une difficulté spécifique. En effet les exercices ont des longueurs très différentes, et certains exercices présentant un texte particulièrement long peuvent poser un problème : il faut en effet lire le texte en entier une fois, puis le parcourir une seconde fois pour en extraire des informations. Certains individus peuvent être découragés par ces textes. De plus, dans ces cas, la comparabilité entre présentation sur papier ou informatique est particulièrement délicate. On peut en effet douter de la comparabilité d'une lecture sur papier et d'une lecture sur un (petit) écran d'ordinateur, et à plus forte raison dans les conditions de passation d'une enquête statistique.

### 3.2.1. Non réponses et critères sociodémographiques

Tableau 6 : Déterminants de la non réponse

| | estimation | écart-type | t | p-value | signif. |
|---|---|---|---|---|---|
| *Constante* | 0.493458 | 0.102547 | 4.812 | 1.60e-06 | *** |
| *Sexe* | | | | | |
| homme | ref. | | | | |
| femme | 0.032354 | 0.020071 | 1.612 | 0.107131 | |
| *Nationalité* | | | | | |
| française | ref. | | | | |
| étrangère | 0.110076 | 0.019696 | 5.589 | 2.60e-08 | *** |
| *Années d'études* | | | | | |
| moins de 5 ans | ref. | | | | |
| entre 6 et 9 ans | -0.039582 | 0.086067 | -0.460 | 0.645643 | |
| entre 10 et 12 ans | -0.244546 | 0.084383 | -2.898 | 0.003795 | ** |
| entre 13 et 16 ans | -0.406290 | 0.084741 | -4.795 | 1.75e-06 | *** |
| plus de 16 ans | -0.465884 | 0.086089 | -5.412 | 6.98e-08 | *** |
| *Age* | | | | | |
| moins de 25 ans | ref. | | | | |
| de 26 à 35 ans | 0.081141 | 0.038983 | 2.081 | 0.037518 | * |
| de 36 à 45 ans | 0.209826 | 0.042582 | 4.928 | 9.00e-07 | *** |
| de 46 à 55 ans | 0.249894 | 0.045181 | 5.531 | 3.60e-08 | *** |
| plus de 55 ans | 0.293857 | 0.049976 | 5.880 | 4.79e-09 | *** |
| *Type d'emploi* | | | | | |
| ouvrier | ref. | | | | |
| employé | -0.129985 | 0.036576 | -3.554 | 0.000388 | *** |
| professions intermédiaires | -0.183466 | 0.041525 | -4.418 | 1.05e-05 | *** |
| cadres | -0.244534 | 0.042549 | -5.747 | 1.04e-08 | *** |
| dirigeants | -0.222049 | 0.042593 | -5.213 | 2.04e-07 | *** |
| inactifs | -0.148401 | 0.035675 | -4.160 | 3.32e-05 | *** |
| *Anneés de travail* | | | | | |
| moins de 10 ans | ref. | | | | |
| de 10 à 19 ans | -0.075936 | 0.034255 | -2.217 | 0.026747 | * |
| de 20 à 29 ans | -0.087417 | 0.039796 | -2.197 | 0.028159 | * |
| de 30 à 39 ans | -0.130087 | 0.044289 | -2.937 | 0.003349 | ** |
| plus de 40 ans | -0.082700 | 0.053271 | -1.552 | 0.120708 | |

Source : exploitation de l'enquête pilote.

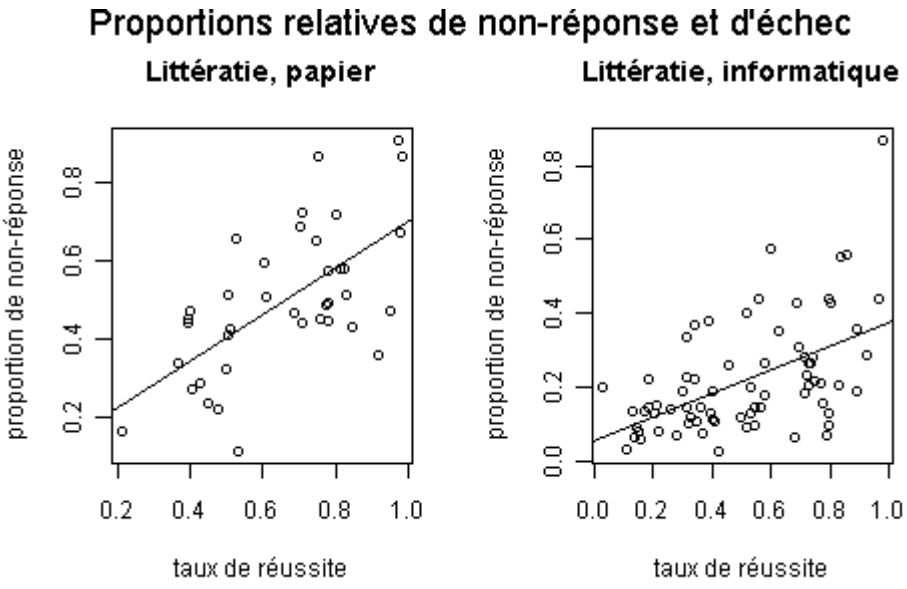
Lecture : Le modèle présenté ici est la régression de la variable indicatrice d'un taux de non réponse individuelle (pourcentage du questionnaire d'évaluation des compétences) supérieure à la moyenne observé. Les niveaux de signification sont ;*** : p-val < 0.001, ** : p-val < 0.01, * : p-val < 0.05

L'analyse de la non réponse en fonction des critères socio-démographiques fait apparaître plusieurs facteurs importants. On remarque tout d'abord qu'il n'y a pas d'effet de sexe. Au contraire, l'âge de la personne répondante est nettement significatif et a un effet qualitativement important. Un véritable gradient d'établit, qui voit la personne présenter d'autant plus de non réponses partielles qu'elle est plus âgée. La durée de la scolarité, qui est un indicateur de niveau scolaire atteint, et dont on peut penser qu'elle est corrélée à la compétence dans les traits mesurés par l'enquête voit elle aussi un gradient important. Plus on a eu une scolarité longue, plus la réponse à l'enquête est complète. La durée de l'expérience professionnelle, si elle apparaît comme statistiquement significative, a un effet beaucoup plus faible. On peut ajouter qu'être de nationalité étrangère amène une probabilité de non réponse importante plus élevée. Enfin, si on regarde le statut de l'emploi, on constate que des ouvriers aux employés puis aux professions intermédiaires, la probabilité de non-réponse importante diminue régulièrement. Elle diminue encore lorsque on passe aux cadres et aux dirigeants d'entreprise.

### 3.2.2. Non réponse et difficulté

Le tableau précédent montre déjà une relation nette entre non-réponse et indicateurs de compétence. On peut cependant nuancer ce jugement : dans les données de l'enquête, on peut observer que, si on considère l'ensemble des réponses « non justes », c'est-à-dire l'union des réponses fausses et des non-réponses, la part des réponses fausses diminue lorsque la compétence augmente. Cela indique un comportement différencié de non réponse entre différents niveaux de compétence.

Graphique 5 : Non réponses et réponses fausses par item

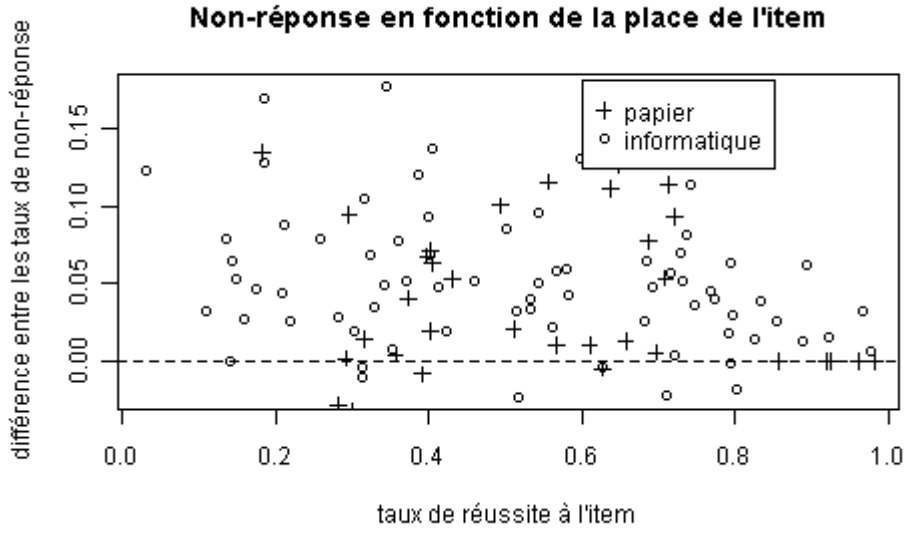



### 3.2.3. Non réponse et lassitude

La composante de motivation et de lassitude peut être approchée par la longueur de l'interrogation. Le questionnaire de l'enquête pilote était conçu de manière à ce que chaque item ait deux positions possibles, dans la première ou la seconde partie du questionnement (à l'exception de quelques questions d'orientation très faciles en numératie). La différence entre les scores obtenus dans chacune de ces deux positions permet d'en observer l'effet. On observe qu'en effet le taux de non réponse à l'item est plus élevé lorsque cet item fait partie de la seconde moitié de la passation des

exercices. De plus cette différence est indépendante de la difficulté de l'item. Ceci indique la présence d'un effet de lassitude lié à la longueur de l'interrogation.

Graphique 6 : Non réponse par item en fonction de la place dans le questionnaire




Il est aussi possible de mettre en évidence cet effet de lassitude en considérant pour chaque individu la suite des questions posées afin de repérer le rang des items laissés sans réponse. La difficulté est que, par construction du questionnaire, toutes les personnes ne sont pas interrogées sur le même nombre d'items.

### 3.3. Les traitements de la non réponse de l'enquête pilote

L'un des résultats les plus importants de l'enquête PIAAC sera la répartition de la population dans les différents groupes de compétences définis lors de l'enquête IALS. En particulier, la donnée la plus importante en terme de politique publique est la proportion de personnes souffrant de difficultés importantes (niveau 1) en littératie. Cette proportion avait été particulièrement élevée dans l'exploitation de l'enquête IALS, mais avait ensuite été rconnue comme fragile.

La non-réponse a un effet complexe sur l'estimation finale, et son traitement est donc important au double point de vue théorique et empirique. Cependant, l'architecture particulièrement complexe de l'enquête rend les choses difficiles. Il serait nécessaire d'enregistrer des variables de comportements permettant de modéliser et de mesurer des indicateurs de compétence et de motivation durant l'enquête. Par exemple, l'étude des temps de réponse sera l'une des pistes étudiées pour l'enquête finale. Cependant, ces temps ne permettent d'utiliser que l'interrogation informatique et non l'interrogation sur format papier.

L'une des possibilités pour estimer l'impact des non réponses est de comparer différents traitements possibles de la non réponse partielle. Le scénario central retenu par le Consortium PIAAC distingue deux types de non-réponses. La non-réponse « finale » est une non-réponse telle qu'aucune réponse n'est disponible dans une partie ultérieure de l'enquête. Ainsi, elle correspond aux cas pour lesquels l'individu interrogé décide d'arrêter l'enquête. Cette non-réponse est considérée comme un signe de démotivation indépendant de la compétence à mesurer, et est donc traitée comme une donnée manquante ne participant pas à l'estimation. Les autres non-réponses sont considérées comme le signe que la personne n'était pas capable de répondre et sont traitées comme des réponses fausses.

Toutefois, il est possible d'envisager un certain nombre d'autres traitements différenciés de la non-réponse. En particulier, il est possible de traiter en bloc les exercices comportant plusieurs questions. Ainsi, on peut considérer qu'un exercice pour lequel aucune réponse n'est disponible n'a pas été abordé, et est le signe que l'exercice n'a pas intéressé l'enquêté. Cette non-réponse est alors liée à la motivation et non à la compétence. Une autre variante consiste à considérer l'ensemble des bonnes

réponses, et à comparer les niveaux de difficultés des bonnes réponses avec le niveau de difficulté de la question pour laquelle une non réponse est enregistrée. On peut alors donner un critère de facilité à partir duquel on considère que ce n'est pas le manque de compétence mais plutôt le manque de motivation qui a amené la non-réponse.

Le tableau suivant présente le résultat du nombre de personnes de niveau 1 en littératie pour trois variantes de traitement de la non-réponse (ces estimations ont été faites par la DEPP). Deux variantes « extrêmes » sont ajoutées au scénario central présenté *supra*. Dans la variante « pessimiste », toutes les non-réponses sont considérées comme des erreurs pour l'estimation. Au contraire, dans la variante « optimiste », toutes les non-réponses sont considérées comme valeurs manquantes dans l'estimation. De plus, l'estimation peut être effectuée pour différentes populations de référence. Ainsi les estimations sont-elles faites séparément pour les interrogations sur chacun des supports, puis en prenant l'interrogation complète avec une contrainte sur les paramètres communs.

Tableau 7 : Pourcentage de personnes de niveau 1 en littératie (en %)

| Population de référence | support | variante « optimiste » | variante « centrale » | variante « pessimiste » |
|---|---|---|---|---|
| papier | papier | 44 | 49 | 52 |
| informatique | informatique | 34 | 35 | 35 |
| Ensemble | mixte | 39 | 41 | 42 |
| | papier | 48 | 49 | 52 |
| | informatique | 32 | 34 | 34 |

*Source : estimations DEPP*

Ce tableau montre que les différentes variantes de l'estimation ont un effet non négligeable mais qui reste cependant faible au regard de l'ordre de grandeur de ces résultats. On constate d'autre part que ces ordres de grandeur sont comparables à ceux qui avaient été observés pour l'enquête IALS.

Toutefois, il faut mettre l'accent sur la grande spécificité du cas de l'enquête pilote. En effet, l'ancrage effectué sur les items IALS pour se ramener à la même échelle de résultats a une grande importance. Pour cet ancrage, très peu d'estimations de paramètres d'item étaient disponibles, soit six pour l'interrogation informatique et seulement deux pour l'interrogation sur support papier. Dans ces conditions, de petites erreurs d'observation peuvent altérer les résultats en termes de proportions dans la population dans une grande mesure. Ceci indique que les résultats précédents sont intéressants par leur ordre de grandeur, mais ne seront sans doute pas comparables à ceux de l'enquête finale.

Deux facteurs, en plus des modifications de l'architecture de l'interrogation, seront différents et importants dans l'enquête finale. En premier lieu, l'ancrage sur l'échelle antérieure de compétence se fera sur un nombre beaucoup plus grand d'items. La plupart de ces items viennent de l'estimation de l'enquête ALLS. Cet ancrage sera alors beaucoup plus robuste. Ensuite, les exploitations de l'enquête pilote ont permis de raffiner le choix des items de l'enquête finale, et en particulier de supprimer un certain nombre d'items problématiques repérés durant l'exploitation de l'enquête pilote.

# Conclusion

La faiblesse des informations disponibles, à la fois sur les données des autres pays participants et sur les détails du modèle MRI utilisé par ETS, interdisent d'anticiper raisonnablement l'influence qu'aura le traitement des non-réponses partielles sur l'estimation de la répartition des répondants par niveau de compétence lors de l'enquête finale. Cette étude a cependant démontré qu'une prise en compte différenciée des différentes catégories de non réponse par des hypothèses un peu plus riches que celles utilisées habituellement dans ce genre d'enquête permet de diminuer, dans le cas le plus favorbale, de 8 point l'importance du groupe de niveau 1.

Le point le plus rassurant concerne cependant l'introduction d'exercices informatisés. Le recourt à la VM rend l'estimation des scores de compétence moins dépendante de la méthode de traitement de la

non réponse partielle finalement retenue. A défaut de corriger les faiblesses sous-jacentes aux modèles MRI dans le cadre d'une étude internationale, l'utilisation de la VM permet en tout cas de limiter partiellement le désavantage relatif de la France dû à la plus grande fréquence des données manquantes lors des évaluations psychométriques.

Pour autant, l'influence de la méthode de traitement retenue est sans commune mesure avec l'influence des défauts de structure de l'enquête elle-même. C'est en amont que se jouent réellement les possibilités d'agir sur les non-réponses en prenant en compte, dans le protocole d'enquête et dans la construction du questionnaire, les attitudes des enquêtés et les caractéristiques des interactions enquêteur/enquêté. Une enquête de mesure des compétences, surtout sur la population adulte, ne peut pas faire l'économie d'une réflexion en profondeur sur les éléments susceptibles de favoriser une attention et une motivation maximales pendant le totalité de la durée de l'interrogation. Si l'introduction de l'outil informatique est indéniablement un aspect positif au regard de cette préoccupation, elle n'épuise pas pour autant la totalité des solutions, certaines peu couteuses, qui pourraient encore être adoptées.

En ce sens, la longueur du questionnaire, le positionnement du module biographique et le choix discutable de certains items font craindre qu'une fois encore les problèmes de motivation non pris en compte (et donc de non réponses partielles) exerceront un effet néfaste sur la qualité et la fiabilité des résultats de l'enquête finale. Par contre, cette fois-ci, cet effet néfaste ne sera pas totalement spécifique aux résultats français.

# Bibliographie


Blum A. et Guérin-Pace F*., Des lettres et des chiffres. Des tests d'intelligence à l'évaluation du « savoir lire »*, Fayard, Paris, 2000.

Blum A. et Guérin-Pace F*.,* « L'illusion compararive. Les logiques d'élaboration et d'utilisation d'une enquête internationale sur l'illettrisme »*, Population*, 2, 1999.

Carey S. (Ed.)*, Measuring Adult Literacy. The International Adult Literacy Survey in the European Context*, Office for National Statistics, Londres 2000.

Galton F., *English Men of Science: their Nature and Nurture*, Mac-Millan & Co, Londres, 1874.

Herpin N. et Jonas N., *La sociologie Américaine. Controverses et innovation*, La Découverte, Paris, 2011.

Hochlehnert A., Brass K., Moeltner A. et Juenger J., « Does Medical Students' Preference of Test Format (Computer-based vs. Paper-based) have an Influence on Performance? », *BMC Medical Education*, 11-89, 2011.

Noyes J., Garland K. et Robbins L., « Paper-based versus computer-based assessment: is workload another test mode effect ? », *British Journal of Educational Technology*, 35- 1, pp. 111-113, 2004.

OCDE et Statistique Canada, *La littératie à l'ère de l'information. Rapport final de l'enquête internationale sur la littératie des adultes*, Paris, OCDE, 2000.

Walters P. B., « Betwixt and between discipline and profession: a history of sociology of education », in Calhoun C. (dir.), *Sociology in America*, University of Chicago Press, Chicago, 2007.

Wang, S., Jiao, H., Young, M. J., Brooks, T. E., et Olson, J., « Comparability of computer-based and paper-and-pencil testing in K-12 assessment: A meta-analysis of testing mode effects. », *Educational and Psychological Measurement*, 68, 5-24, 2008.