DE LA CONCEPTION A L'EXPLOITATION : LA QUALITÉ DANS LES ENQUÊTES AUPRÈS DES MÉNAGES

Daniel Verger

INSEE-Unité Méthodes Statistiques

***Précisions liminaires** :*

Ce texte développe une intervention faite au colloque francophone sur les sondages qui s'est tenu à Autrans les 17 et 18 octobre 2002.
Il doit servir de base à la rédaction d'un manuel des « bonnes pratiques » destiné aux nouveaux concepteurs d'enquête. La présente rédaction est incomplète et provisoire : en particulier certains exemples évoqués seront précisés, des exemples chiffrés provenant d'expertises et expériences internationales seront ajoutés, les références bibliographiques seront complétées. Le lecteur qui désirerait disposer dès maintenant de compléments peut consulter Froment(1994) et Platek(1984). Le manuel sera aussi complété par une partie consacrée aux enquêtes auprès des entreprises.

Plan :

3 parties principales dans ce texte. Après avoir indiqué comment, et avec quelles difficultés, on pouvait envisager de mesurer la qualité, on y développe les différents problèmes qui guettent le concepteur à chaque étape de la chaîne de production de l'enquête. La dernière partie propose quelques voies concrètes d'amélioration des protocoles existants.

Introduction

Les approches européennes habituelles (théorisées ? administratives ?) de la qualité d'une statistique se déclinent en diverses composantes réunies sous le vocable de RATACC : **R**(elevance) **A**(ccuracy of estimates) **T**(imeliness, punctuality) **A**(ccessibility and clarity) **C**(omparability) **C**(oherence) **C**(ompletness).

Dans le dispositif français, l'instance particulièrement chargée de veiller à la qualité des enquêtes statistiques mises en œuvre par le service public[1], à savoir le **Comité du Label**, dont le visa (i.e. le **label « d'intérêt général et de qualité statistique »**) est indispensable avant de pouvoir lancer une enquête « obligatoire » ou « d'intérêt général », cherche bien à juger les nouveaux projets à cette aune là, avec le souci d'en vérifier la « qualité ». Mais l'ensemble des composantes canoniques (Ratacc), n'est pas traité, de fait, par le Label -et par suite dans ce texte-. En ce qui concerne la pertinence, le projet est plutôt jugé en amont (dans le cadre du CNIS, qui décerne un «visa d'opportunité ») ; le Comité du Label se concentre sur l'adéquation entre la fin et les moyens et la qualité technique (on juge moins l'accessibilité, la ponctualité puisque ce sont plutôt des faits que l'on constate ex post alors que le Label est ex ante).

La présente contribution, forcément rapide, se nourrira en particulier de l'expérience acquise par l'auteur en 3 ans d'activité en tant qu'expert auprès du Comité du Label dans sa formation

[1] Insee, Services Statistiques des Ministères

« Ménages » (plus de 80 enquêtes analysées), ainsi que des expériences passées en tant que concepteur[2] ou maître d'ouvrage pour des projets d'enquête. Cette optique conduit à s'imposer de ne donner que des **exemples réellement rencontrés**, de ne travailler qu'avec un **« vrai » bêtisier**, pas avec des exemples artificiels construits pour l'occasion[3]. Grossir le trait, recourir à la caricature affaiblit le message : le concepteur sourit, mais avec l'arrière-pensée qu'il ne tomberait jamais dans un piège aussi grossier ! ici, le message est clair : ce sont des concepteurs en chair et en os qui ont proposé la formulation erronée…L'erreur, évidente quand on pointe dessus, ne l'était pas autant ex ante, puisque le questionnement est allé au moins au stade du test terrain, voire au stade de la présentation finale au Comité du Label. Plus subtil, le message a plus de chance d'être reçu[4].

Le présent texte ne doit ni démoraliser (et inciter à jeter le bébé avec l'eau du bain, et à douter de tout dans les enquêtes), ni être rejeté comme futile. Il doit inciter à améliorer la conception, sans perdre de vue qu'il restera forcément une portion d'incertitude, de flou irréductible, inhérente à l'instrument « enquête » : on peut réduire le flou, mais pas l'éradiquer complètement.

1. L' appréciation de la qualité : du rayon des instruments disponibles à l'éventail des difficultés rencontrées

Une enquête auprès des ménages est en fait une chaîne complexe de tâches (la programmation des tâches faite, dans le cadre de la méthode de conduite de projet en vigueur à l'INSEE, pour une édition passée de l'enquête Logement, avait identifié quelques 400 tâches élémentaires) avec pour maillon central l'interaction enquêteur-enquêté[5]. La difficulté à quantifier ce point a fait que l'on a souvent privilégié la partie trouvant son origine dans l'« aléa de sondage », la plus facile à mettre en formule. On va essayer d'avoir une présentation plus complète, plus globale, l'idée force étant que la qualité est une résultante d'ensemble, et qu'il est vain de raffiner sur un point, en laissant d'autres aspects à la dérive : c'est le **maillon faible qui définit la qualité**. Ce parti pris ne pourra être conduit avec un degré de rigueur scientifique uniforme, car il est exceptionnel que l'on dispose d'éléments quantitatifs fiables relatifs à des effets spécifiques bien identifiés.

De quoi dispose-t-on en effet pour juger de la qualité dans une enquête ? Sur quoi les membres du Comité du Label peuvent-ils asseoir leur expertise ? Le dossier qu'ils reçoivent comprend entre autres une fiche décrivant le plan de sondage, avec une estimation des taux de réponse, basée sur des tests ou des résultats d'enquêtes similaires, le questionnaire (le cas échéant en plusieurs versions selon l'évolution d'un test à l'autre) ainsi que le compte-rendu des tests. Pour une enquête complètement nouvelle, seuls ces éléments sont systématiquement disponibles. Dans de trop rares cas, on dispose d'opérations méthodologiques spéciales destinées à renseigner sur tel ou tel point délicat. Pour une enquête rééditant des opérations antérieures, on peut disposer d'éléments provenant de l'exploitation des éditions passées mais, comme on va le voir, il est difficile d'isoler les conséquences d'un défaut particulier. Lorsqu'un véritable **bilan de fin de projet** est disponible, on est évidemment dans une situation beaucoup plus favorable, ce qui fait regretter que ce type de document ne soit pas, actuellement, rédigé systématiquement.

[2]N'ayant, dans ce rôle, aucunement été à l'abri des difficultés évoquées, mon expérience sera sollicitée pour enrichir le « bêtisier » à éviter.

[3] Par exemple, dans le manuel canadien, on utilise une question fictive relative à la politique étrangère de San Marin, pour illustrer la recommandation de ne pas poser des questions hors du domaine de connaissance du répondant.

[4] Sauf à penser qu'à force de subtilité, on « coupe les cheveux en quatre », que ce qui est dit est vrai, mais n'a pas d'influence réelle.

[5] Il ne faut pas négliger non plus, comme source de difficulté potentielle, le rôle de la Direction régionale, qui intervient comme un agent supplémentaire ayant son propre comportement, par exemple au moment de la formation : n'oublions pas que le concepteur ne forme pas directement les enquêteurs, mais des formateurs DR « relais » qui ne sont pas toujours aussi neutres qu'on pourrait le souhaiter (aptitudes pédagogiques différentes, sensibilité spécifique au sujet de l'enquête….). Ce niveau de complexité ne sera plus évoqué dans la suite, même si on pourrait développer le thème « effets enquêteur » par des considérations sur « l'effet DR ».

1.1 Les instruments disponibles

1.1.1 Les tests

D'une façon standard, «ex ante », lors de la phase de mise au point de l'enquête, la qualité est appréciée, et améliorée, au cours d'une campagne de trois ou quatre tests. L'élément principal du diagnostic sur le questionnement est constitué des **remontées des enquêteurs, riches mais incomplètes** : **une question peut «bien passer» mais conduire à des résultats de mauvaise qualité**, sans que cela soit perçu des enquêteurs. On dispose ainsi d'une validation faible de la question, condition nécessaire pour qu'on puisse l'introduire, mais en aucun cas suffisante ; elle indique si, oui ou non, l'enquêté réagit contre la question, mais ni s'il la comprend réellement ni si elle se révèle utile.

Ainsi, lors de la préparation de l'enquête «Modes de vie-Production Domestique », à la Direction régionale de Bourgogne, lors du bilan du test, une première enquêtrice s'est plainte du questionnaire, qui, selon elle, contiendrait des questions inutiles, au premier rang desquelles une question sur la fabrication maison de linges, torchons … « plus personne ne fait cela de nos jours » ; ce à quoi une deuxième enquêtrice, tout aussi catégorique, a déclaré qu'elle trouvait aussi cette question inutile, mais parce que, selon elle, tout le monde faisait cette activité. Dans la discussion qui s'en est suivie, il est apparu qu'effectivement le taux de pratique était égal à 0 dans la zone de la première enquêtrice, à 100% dans la zone de la seconde ! et ce simplement parce que, dans le second cas, il y avait à proximité une usine textile, qui vendait à prix de gros du tissu au mètre…

Dans les anciennes enquêtes Epargne (1973-1975 et 1974-1976), on posait une question toute simple relative aux héritages « Avez-vous fait un héritage. Si oui, quel en était le montant » ; à l'exploitation on s'est aperçu que les taux de personnes concernées étaient faibles, même pour les personnes de plus de 75 ans, qui, selon toute vraisemblance, avaient généralement perdu leurs deux parents, et ce alors même que dans la distribution des montants renseignés figuraient des valeurs de l'ordre de 50 F ; il n'y a pas d'autre explication qu'une troncature effectuée par certains ménages, mais pas par tous, selon le montant, certains réservant le mot « héritage » aux seules transmissions à valeurs élevées. Un questionnement plus long, mis au point pour l'enquête Actifs financiers 1986, et conduisant à décrire l'ensemble du phénomène (mort des parents, point du patrimoine au décès -le « gâteau »-, point sur le nombre de cohéritiers -les « convives »-, recensement des biens transmis) aboutit à des pourcentages de ménages « héritiers » croissant avec l'âge et approchant les 100% chez les personnes ayant perdu leurs deux parents. Les tests de l'enquête Epargne n'avaient rien signalé.

Tous les exemples que l'on va citer de défauts qui apparaissent lors de l'exploitation statistique, principalement suite à des incohérences internes ou externes, proviennent en général d'enquêtes testées, mais pour lesquelles les enquêteurs n'ont rien fait remonter de particulier lors des réunions de bilan ! Il ne s'agit ici ni de nier l'intérêt des tests, qui contribuent grandement à l'amélioration des opérations, ni de critiquer le professionnalisme des enquêteurs : les problèmes n'apparaîtraient à leurs yeux que s'ils avaient une idée précise des ordres de grandeur des phénomènes visés par l'enquête, non pas au niveau national, mais au niveau de leur zone d'enquête spécifique, ce qui est évidemment hors de portée[6]. Si un test qui se passe mal indique clairement la marche à suivre, un test qui se passe bien est souvent trompeur, car il risque d'endormir la vigilance du concepteur par un illusoire bilan positif.

[6] Les tests ne fournissent pas non plus d'estimation fiable des taux de refus. Souvent, ils ne sont pas conçus pour cela (échantillon très particulier, pas « représentatif ») ; mais même dans le cas contraire, les conditions ne sont pas identiques à ce qui prévaut lors de l'enquête « en vraie grandeur » ; on demande la coopération du ménage pour aider à la mise au point, et souvent ceci est bien accueilli (même si on a pu observer des réponses du style «l'Etat ne fait rien pour moi, je ne vois pas pourquoi je ferais un effort pour l'aider », mais pour l'instant elles sont très minoritaires) ; pouvant jouer en sens contraire, il n'y a ni le visa d'obligation (cf infra) ni la publicité via la mairie qui est en général faite pour les enquêtes Insee.

1.1.2 Les exploitations de fichiers

Pour contourner ces limites qui viennent amoindrir l'efficacité du diagnostic des enquêteurs, c'est bien sûr vers l'exploitation des fichiers -définitifs ou de tests - qu'il faut se tourner, car c'est ainsi que peuvent apparaître des problèmes d'incohérence, révélateurs de difficultés.

- En premier lieu, il peut s'agir de **cohérence interne** des données.

Ainsi l'exploitation des résultats de l'enquête INED portant sur les pratiques sexuelles fait apparaître des différences inconciliables entre les nombres de partenaires décrits par les hommes et par les femmes, un peu comme si, interrogés au téléphone sur ce sujet intime, les hommes avaient tendance à flatter leur ego en surestimant le nombre de leurs conquêtes, alors que les femmes, plus discrètes, plus pudiques avaient une tendance inverse à la sous-estimation[7]…de quoi alerter sur la qualité des résultats obtenus !

Une **forme des liaisons entre variables sinon contraire du moins non conforme à ce que l'on attend**, à ce qui est connu par ailleurs, à la théorie, peut aussi alerter. Le fait que dans l'enquête IALS[8] destinée à mesurer les performances en littératie, les personnes les plus diplômées ne fassent pas de « sans faute » et donc ne révèlent pas de capacités supérieures en littératie est certainement symptomatique d'un problème au niveau du protocole d'observation[9]. On assimilera à ce cas de cohérence interne, ce qui se passe dans le cas de réinterrogations par entretiens « semi-directifs » (cf infra). L'enquête INSEE dite «Situations défavorisées » a ainsi été complétée par des entretiens enregistrés effectués non pas à partir d'un questionnaire fermé, mais à partir d'une grille semi-ouverte, par un chercheur (Bouchayer 1994) : ce mode de réinterrogation, où le chercheur prend le temps de stimuler la mémoire de l'enquêté, est à l'affut des incohérences, des zones d'ombre, des hiatus dans la continuité de la biographie, fait apparaître nombre de périodes courtes d'emploi ou de chômage, qui avaient été omises dans la réponse au questionnaire statistique fermé (cf Battagliola et alii dans Bouchayer 1994) et dont la révélation semble bien hors de portée de ce type d'investigation, quel que soit le soin apporté au questionnement.

Une forme plus riche d'analyse de la cohérence interne est disponible lorsque l'enquête en question comporte une dimension panel, et que le même ménage, le même individu est réinterrogé à plusieurs périodes : il s'agit de la **cohérence longitudinale.** Ainsi, dans le panel européen, les questions concernant la possession de biens durables cherchent à séparer les ménages non possesseurs en deux sous-groupes, selon que la non possession relèverait d'un choix ou au contraire serait le fait d'une contrainte d'ordre pécuniaire. L'instabilité de cette réponse dans le temps est certainement le signe du peu de fiabilité de la distinction : d'une année sur l'autre, si les pourcentages macroéconomiques restent remarquablement stables, ce n'est que le résultat de la compensation statistique entre deux mouvements en sens contraire d'ampleur non négligeable, puisque environ 20% des non possesseurs passent d'une non possession choisie à un manque contraint et réciproquement et ce pour tous les biens étudiés.

La chronique des revenus observés dans ce même panel, en l'absence d'entretien dépendant (avec contrôle de cohérence entre les déclarations successives) permet de calculer des taux d'entrées/sorties de pauvreté : selon que l'on se fie aux données brutes ou que l'on travaille sur des données apurées (à la main ou statistiquement, avec recours à des modèles économétriques), selon les hypothèses acceptées, on peut mettre en évidence des taux de sortie variant de 37 % à 12-15% (Lollivier-Verger

[7] D'autres interprétations sont envisageables : les hommes auraient un meilleur souvenir de leurs conquêtes passées …ou compteraient comme partenaires les prostituées, alors que celles-ci ne déclareraient pas comme tels leurs clients.
[8] International Adult Literacy survey ; cf infra
[9] Dans le cas de IALS, la difficulté est concentrée sur les très hauts diplômes, puisque la réussite aux épreuves commence bien par croître avec le diplôme, la décroissance, légère, se produisant après le bac. Reste qu'elle est surprenante et prouve bien que, pour cette sous-population, le problème est réel.

2002). Dans les 2/3 des cas, les « fiches Revenus » font apparaître une forte évolution, alors que la question directe sur l'évolution des revenus indiquerait plutôt qu'il ne s'est rien passé[10].

- L'analyse du **degré de cohérence avec des sources externes** est également riche d'enseignement. Considérons le domaine du patrimoine, en particulier celui des montants placés sur les divers actifs, financiers ou immobiliers ; le rapport entre les valeurs observées et les montants arbitrés par la Comptabilité Nationale, le «taux de couverture », permet de juger de la qualité des montants recueillis : les études conduites à partir de l'enquête Actifs financiers de 1992 (Arrondel, Guillaumat-Tailliet, Verger 1996) ont ainsi mis en évidence des taux de couverture allant de 97 % pour le logement, 98% pour le patrimoine professionnel à 11 % pour les comptes à terme et les bons en passant par 29 % pour les valeurs mobilières. Par rapport aux sources fiscales, l'enquête Emploi fournit des salaires sousestimés d'environ 20 % ; les dépenses de médicaments dans les enquêtes sur les dépenses de santé sont également sousestimées fortement (au moins 20 %).

Mais ces **exploitations directes ne fournissent pas une panacée**. En premier lieu, la cohérence externe peut être bonne sans que cela signifie grand chose : si comptabilité nationale et enquête Actifs donnent les mêmes valeurs pour l'immobilier, c'est sans doute parce que la fabrication des chiffres de la Comptabilité Nationale, dans ce registre, repose quasi exclusivement sur des enquêtes : on ne fait donc que vérifier que deux enquêtes donnent des résultats proches, pas qu'elles sont bonnes dans l'absolu ; inversement des chiffres discordants peuvent révéler un manque de qualité de la source externe, ou un simple problème de comparabilité ; c'est ce qui se passe avec l'évaluation Comptabilité Nationale des actions non cotées relatives à l'entreprise personnelle du ménage.

Les résultats des analyses directes des fichiers peuvent aussi être difficiles à interpréter. A titre d'exemple, on peut citer une étude qui avait été conçue pour mesurer économétriquement l'effet d'une interrogation par « proxy ». On avait travaillé sur l'enquête Emploi, plus précisément sur la variable donnant le logarithme du niveau de salaire ; aux variables classiquement introduites dans une équation de salaire (âge, diplôme, profession, secteur….), on avait rajouté une variable indiquant si c'était le salarié lui-même qui avait répondu, ou un proxy, en l'occurrence généralement son conjoint. Cette variable générait un effet significativement positif : toutes choses égales, le salaire est plus élevé lorsque le répondant est le proxy, ce qui semblerait prouver qu'interdire ou non le recours au proxy n'est pas sans conséquence. Une telle conclusion serait toutefois hâtive, car le phénomène peut n'être rien d'autre que la manifestation de problèmes de sélection, d'endogénéité, le mécanisme étant alors le suivant : si c'est un proxy qui répond, c'est que le salarié lui-même est absent, ce qui a d'autant plus de chances de se produire qu'il a des horaires de travail longs, ce qui peut être positivement corrélé avec le niveau de la rémunération[11]. L'effet du « proxy » ne serait alors que la manifestation de l'effet d'une variable omise, inobservée, la durée effective du travail. Mais les problèmes de déclaration de cette variable (trop de réponses conventionnelles, égales à la durée légale) empêchent un traitement correct du phénomène. Reste que le doute subsiste. Entre les deux premières vagues du Panel européen, on avait intercalé deux entretiens téléphoniques (4 et 8 mois après la première visite) destinés à maintenir le contact (et donc à réduire l'attrition) tout en renseignant sur les modifications vécues sur le plan démographique ou de l'emploi. Pour les chômeurs, on recensait en particulier les offres d'emploi reçues, acceptées ou refusées, et, le cas échéant, on s'efforçait de mesurer l'évolution du salaire de réserve. Ce dispositif, coûteux, a ensuite été abandonné, en particulier parce que l'évolution de ce salaire de réserve était rendue ininterprétable par un accroissement du taux de recours au proxy dans ces entretiens téléphoniques. Les réponses par proxy apparaissaient beaucoup

[10] A la fin de la deuxième vague du panel, on avait regardé tous les cas d'entrée-sortie de pauvreté ; dans la moitié des cas, il était certain qu'il s'agissait d'une erreur de mesure (confusion entre revenu mensuel et revenu annuel, erreur dans le passage de la valeur mensuelle à la valeur annuelle -une fois l'enquêteur avait posé la multiplication et la preuve de l'erreur de calcul était manifeste, oubli d'une composante voire de tous les revenus d'un individu une des deux années…). Quant aux 50 % restants, une partie apparaissait douteuse sans que l'on puisse conclure en toute certitude à l'erreur. Seul le tiers des mouvements semblait réel, soit un ordre de grandeur similaire à ce que l'on a observé à partir des sept premières vagues avec des méthodes plus statistiques.

[11] Ce peut aussi être le cas d'un ouvrier travaillant en 3/8, en horaires de nuit avec des primes d'astreinte spéciales…

moins dispersées, beaucoup plus « attirées » par le SMIC que les réponses fournies par l'intéressé lui-même (Canceill 199*)

De plus, ces **analyses ne peuvent en général indiquer la cause du problème** : on peut apprécier le rôle de la pondération, voire l'influence des corrections pour variables erronées (dans la mesure où l'on a bien gardé trace des variables brutes[12]) en faisant des variantes. Quantifier les éventuels effets d'une sélection endogène du répondant reste plus hasardeux pour ne pas dire impossible[13]. Enfin reste souvent quelque chose qui échappe au domaine de la preuve (« l'effet questionnaire ») à quoi on attribue ce que l'on n'a pas réussi à attribuer à autre chose.

Ainsi l'enquête Emploi, qui a une dimension panel (trois réinterrogations espacées d'un an dans le cas de l'enquête annuelle, six interrogations à un rythme trimestriel pour la nouvelle enquête en continu) fait apparaître ce que le jargon technique qualifie de «biais de rotation » : les différents fichiers correspondants aux différentes vagues d'interrogation, même pondérés de façon à être représentatifs de la population dans son ensemble, ne donnent pas la même valeur du taux de chômage, et ce systématiquement, pour toutes les années et dans la plupart des pays : les fluctuations d'échantillonnage d'une vague à l'autre, le refus différencié selon qu'en première visite la personne est au chômage ou non expliquent chacun une part de l'écart, mais une importante fraction du biais reste inexpliquée et ne peut provenir que du protocole d'enquête (effet d'habitude par rapport au questionnaire par exemple). Les calage et redressements effectués ne permettent pas de corriger sensiblement ce biais. Toutes causes confondues, les écarts (sur l'enquête annuelle) sont à l'origine d'une incertitude sur le nombre de chômeurs de l'ordre de 100 000 (sur un total d'environ 2,7 millions à l'époque).

Les écarts sur le nombre estimé de résidences principales entre l'enquête Emploi et l'enquête Logement, constatés après le recensement de 1990, n'ont jamais pu être expliqués, même si l'on soupçonne qu'une part importante pourrait provenir de différences dans les méthodes de repondération appliquées aux deux sources.

1.1.3 Les calculs de précision :

La mesure ex post de la qualité, au mieux, passe par les **estimations de variance**. Leur interprétation en est complexe, et leur obtention elle-même discutable. Alors même que l'on cherche à mesurer ainsi la précision du sondage (les formules utilisées ne sont d'ailleurs que des simplifications approximant les vraies formules et, de plus, présupposant l'absence d'erreurs de mesure), les chiffres obtenus dépendent aussi des erreurs de mesure, sans que l'on puisse quantifier séparément les deux effets. Or tout porte à croire que les erreurs de mesure, en l'absence de contrôles dûment introduits au sein du protocole de collecte, sont importantes[14].

1.2 Qualité « dans l'absolu » ou qualité « en vue d'un type d'exploitation bien particulier » ?

Reste, pour terminer cette partie sur la difficile mesure de la qualité, à insister sur un point trop rarement évoqué. Une difficulté ressentie fréquemment lors de l'examen des dossiers au Comité du Label provient du manque d'information sur la place que doit occuper une question particulière dans l'exploitation. Or ce n'est pas neutre. Prenons l'exemple des récentes enquêtes sur la santé (enquête Santé, Vespa -enquête sur les personnes infectées par le virus du VIH-) ; chacune contenait un grand nombre d'items destinés à cerner les symptômes de la dépression : chaque item est ambigu et

[12] un principe de base à respecter - comme en archéologie, toute restauration doit être réversible, voire décelable à l'œil nu- mais qui hélas souffre des exceptions…
[13] Cf infra détails sur l'origine de cette difficulté.
[14] D'où le débat pour ou contre l'introduction de contôles embarqués stricts -cf infra.

inexploitable en tant que tel, mais le score global peut avoir un sens (c'est avec le même type de présupposé que l'on avait construit le protocole d'observation des comportements face au risque et au temps) : on a été ainsi amené à recommander de ne mettre dans le fichier final mis à disposition que les scores globaux et pas les items élémentaires constitutifs. Plus généralement si l'on cherche seulement un ordre de grandeur, des défauts de qualité peuvent être sans importance alors que si l'on cherchait une évaluation précise ils seraient rédhibitoires[15] : or rien n'indique ce qui est visé (et rien n'assure qu'une fois le fichier livré, d'autres que le concepteur n'utilisent pas la variable dans un autre but que celui pour lequel elle a été créée). Faute de savoir qu'une variable a été introduite pour servir d'« instrument » dans le traitement de l'endogénéité, on peut en recommander imprudemment la suppression, comme inutile à la compréhension directe du phénomène d'intérêt.

1.3 Heurs et malheurs du benchmarking :

1.3.1 Le cas des audits sur les résultats de IALS ou « comment a-t-on pu obtenir 40% d'illettrés en France dans une enquête internationale » ?

L'histoire de cette opération est exemplaire pour plusieurs raisons : une conception et une exploitation parrainées par des spécialistes mondiaux reconnus, une collecte menée dans un grand nombre de pays selon des préceptes contraignants afin d'assurer des résultats comparables, et, en fin de course, un résultat « choc » : 41 % d'adultes en France auraient du mal à comprendre un texte simple -contre seulement 13 % en Suède ou aux Pays-Bas- et si les performances croissent avec le niveau de diplôme, elles saturent assez vite et certains très diplômés échouent. L'importance politique du sujet explique la richesse du matériel critique dont on dispose, puisque, suite à ces résultats, deux « audits qualité » officiels ont été commandités, enrichis d'une opération de « retest »[16] et de nombreuses études, dont la principale, conduite par l'INED, a fait l'objet d'un livre entier (Blum,Guérin 2000) . Quant au rapport initial de collecte, il fait environ 400 pages (T.S. Murray, I. Kirsch et L.B. Jenkins 1998). Malgré tout cela, il **est impossible de dégager les causes essentielles du dérapage** : les deux audits livrent ainsi deux diagnostics différents, surtout quant à l'importance relative d'une part des **problèmes de traduction et des biais culturels** (accoutumance à la lecture de graphiques au cours de la scolarité, cadre culturel des situations décrites trop anglo-saxon…), d'autre part des **problèmes d'échantillonnage ou de maîtrise de la collecte** (respect des consignes de la « méthode des itinéraires », choix du répondant vraiment aléatoire ou acceptation de la personne la plus motivée ; rôle de l'enquêteur pour soutenir cette motivation et pour veiller au respect du protocole -recadrer les gens qui ne cherchent pas la réponse dans le texte mais répondent d'après leurs connaissances de façon à ne pas classer les anarchistes en illettrés- ; traitement des personnes étrangères non francophones, distinction insuffisante entre non réponse par manque d'intérêt et non réponse suite à manque de compétence ou réponse fausse etc.), voire même les **problèmes de codage des réponses ou de spécification fine des modèles utilisés pour l'analyse**.

1.3.2 Le cas de l'audit sur les statistiques relatives à la détention de chiens

L'interrogation est venue des professionnels de la nourriture pour chiens, qui s'étaient émus de divergences assez fortes (de l'ordre de 3 %) sur le taux de ménages détenant des chiens, entre diverses enquêtes commandées à 3 sociétés. Un audit a alors été diligenté de façon à trouver des explications à cette différence qui représente quand même un certain nombre de gueules à nourrir. A nouveau, on a mis en évidence une série de défauts, à divers niveaux des 3 chaînes de production, qui, se cumulant, conduisaient à cet écart : pour l'une des sociétés, le problème majeur venait d'une erreur dans la publication de certaines statistiques qui étaient qualifiées de « % de ménages détenteurs » alors qu'il

[15] Si l'on mesure le revenu pour appréhender l'inégalité, on a besoin d'une qualité meilleure que si le revenu n'a d'autre rôle que celui d'un cofacteur destiné à expliquer une pratique de consommation ou l'existence d'un équipement.
[16] Réinterrogation des mêmes ménages avec un protocole verrouillé.

s'agissait de « % d'individus vivant dans un ménage doté d'un chien » ; pour une autre, il s'agissait d'un redressement insuffisamment précis sur le milieu social, le redressement habituel convenant à la plupart des biens usuellement étudiés, mais étant insuffisant dans le cas de ce bien « inférieur »[17] ; on ne pouvait non plus exclure un effet des dates de collecte, certaines enquêtes étant réalisées en janvier (juste après les fêtes, où s'offrent des chiots) alors que d'autres l'étaient en octobre (après les grandes vacances, période où l'on abandonne les animaux -effet « Brigitte Bardot ») ; les questionnaires différaient également légèrement, ce qui pouvait induire des traitements différents des cas marginaux comme les cas d'animaux mis en garde dans de la famille (écart entre le fait d'avoir chez soi un animal et en être propriétaire), ou d'animaux que l'on possède mais qui n'ont jamais le droit d'entrer dans le logement, qui restent à l'extérieur voire dans un chenil.

Tous ces exemples montrent que les différents problèmes de qualité peuvent avoir des effets qui sont loin d'être négligeables, même si la plupart du temps -mais pas toujours- ils ne suffisent pas à bouleverser complètement les ordres de grandeur des phénomènes. Reste à savoir si les conseils relatifs aux bonnes pratiques peuvent ou non réduire significativement cette marge d'incertitude. On espère prouver, dans le domaine de la littératie, grâce à l'enquête IVQ[18], que cela est possible [19].

2. Une approche globale, incluant tous les maillons de la chaîne des tâches

Les réflexions précédentes ont dicté l'organisation de la présente démarche : **une approche globale, incluant tous les maillons de la chaîne des tâches, mais qui doit parfois se contenter de suggérer faute de pouvoir toujours disposer de preuves.**

2.1 L'impossibilité d'avoir des résultats généraux

Même si elles font intervenir à diverses étapes des techniques statistiques éprouvées (échantillonnage, pondération, exploitation), pour lesquelles un traitement scientifique est possible (on peut calculer des écarts-types, effectuer des tests de significativité, juger de la rigueur des enchaînements voire de la rigueur des hypothèses), les enquêtes auprès des ménages se caractérisent par le fait qu'au cœur du processus productif se trouve l'interaction entre deux êtres humains, l'enquêteur et l'enquêté, alchimie subtile, complexe, malaisée à réduire en formules, mais essentielle dans la qualité finale. Il est facile d'imaginer, quand on a présente à l'esprit l'extrême diversité des personnes (genre, âge, niveau d'éducation, milieu social, histoire de vie…), l'irréductible complexité de cette phase. Si l'on considère que les « effets de moment » viennent rajouter une couche d'aléatoire, que telle personne va plus ou moins s'impliquer dans l'enquête, faire un effort plus ou moins intense pour répondre de façon précise selon qu'elle est fatiguée ou non, qu'elle est pressée ou détendue, qu'elle est ou non dans ses « bons jours », on ne s'étonnera pas que le **seul résultat établi soit la quasi absence de résultats généraux** et que ce résultat négatif semble bien caractériser, du moins dans l'état actuel des connaissances, ce domaine de réflexion. Concevoir une enquête ne relève pas (encore ?) d'une science, et c'est ce qui rend difficile la cristallisation des connaissances, la sédimentation des savoir-faire. Alors qu'un ménage a besoin de s'échauffer avant d'entrer dans le sujet, tel autre se fatigue vite : pour le premier la qualité a tendance à croître alors que c'est l'inverse qui se produit dans le second cas : on sait que les deux cas se rencontrent, mais, à l'heure actuelle, aucune mesure précise ne permet de savoir quelle sous-population est, statistiquement, la plus nombreuse. Rien ne prouve non plus que la façon de réagir de quelqu'un soit définie indépendamment du sujet : la lassitude n'apparaîtra sans

[17] Au sens de la théorie microéconomique, est qualifié d' « inférieur » tout bien dont la possession est d'autant moins fréquente que le revenu est plus élevé. Aucune connotation péjorative n'est associée à l'usage de ce qualificatif.
[18] Information et Vie Quotidienne
[19] Les premiers résultats ont été diffusés lors d'un Séminaire Recherche de l'Insee qui s'est tenu le 19/06/03. Le taux de personnes classées dans les plus bas niveaux de littératie serait de l'ordre de 13 % (entre 10 et 15 % selon les hypothèses faites sur les non-réponses totales et partielles) : on est loin des 41 % mesurés par IALS !

doute pas de la même façon selon que l'on parle d'un sujet d'intérêt pour la personne, voire d'une de ses passions, ou que l'on traite d'un domaine qui lui est indifférent[20]. Les mécanismes qui expliquent les fonctionnements de la mémoire sont encore largement obscurs ; or, dans une enquête, le recours à la mémoire est omniprésent. Face à l'énumération d'une longue liste d'items, pour un sujet donné, certains se rappelleront davantage le premier cité, d'autres le dernier et parfois c'est l'effet de « recensy » qui l'emporte et parfois celui de « primacy » ; pour se remémorer les événements qui ont ponctué leur vie, certains partiront du présent et remonteront le temps progressivement, chaque étape rafraichissant la mémoire sur les événements immédiatement antérieurs, d'autres procèderont de façon inverse, descendant systématiquement le cours de la chronologie, d'autres enfin se rappeleront, plus ou moins dans le désordre, de moments cruciaux qui formeront autant de pôles d'ancrage à partir desquels ils feront le travail de maïeutique nécessaire à la production du souvenir (moments importants de leur propre vie -un mariage, une naissance, un décès-, mais aussi événements plus généraux comme une catastrophe survenue dans leur voisinage, ou un événement politique national ou mondial…). Rédiger, dans ces conditions, un manuel de « bonnes pratiques » est une tâche particulièrement ardue, voire ingrate, l'auteur devant éviter à chaque instant deux écueils : du côté Charybde les généralités banales, du côté Scylla les anecdotes sans portée statistique. Conseiller une équipe sur la rédaction du questionnaire est sans doute la partie la plus difficile du travail de méthodologue. Non seulement il ne dispose pas de l'appui d'une science mathématisée, comme c'est le cas quand il parle d'échantillonnage, mais encore il touche au cœur du processus, là où fond (apanage du concepteur) et forme (où l'intervention du méthodologue est licite) sont difficiles à séparer. De plus il est rare que les contraintes matérielles permettent de monter les opérations méthodologiques qui permettraient de tester réellement l'efficacité d'une suggestion : délicates à concevoir et à réaliser, ces opérations coûtent cher, allongent le processus de préparation et sont donc en général sacrifiées ; dans la dernière partie de ce texte, diverses propositions sont faites de façon à progresser dans cette voie. Actuellement; on en est en général réduit à convaincre, -et puissent les anecdotes citées y aider- ou à constater ex post qu'il aurait mieux valu procéder autrement si on avait voulu pouvoir exploiter telle ou telle question. Il est en effet fréquent de déplorer que les enquêtes soient sous-exploitées, ou que l'exploitation prenne trop de temps (cf.: le rapport INSEE dit « rapport Rempp-Faucheux ») ; or souvent ces difficultés cachent des problèmes de qualité : telle question n'est pas exploitée parce que de fait elle s'est révélée inexploitable[21] ; telle étude prend du retard parce qu'un travail de bénédictin a dû être fait pour apurer les données et éradiquer les incohérences faute d'avoir eu un contrôle efficace à la collecte. L'absence, sinon de publicité, du moins de simple information sur ce point dans les bilans d'opération font que l'on fait perdurer les formulations erronées. Animé par un louable souci de comparabilité (au cours du temps et d'une opération à l'autre), le concepteur recherche en effet souvent dans les opérations passées l'exemple d'une formulation spécifique ; outre l'aspect « habit d'Arlequin » des questionnaires obtenus par ces techniques de « couper/coller », on peut déplorer que ceci conduise à reprendre des questions imparfaites : le vrai test de qualité n'est pas, pour une question, d'avoir figuré dans un questionnaire passé, c'est d'avoir été exploitée avec succès. Une des ambitions de ce texte est d'inciter les équipes conceptrices à mettre en doute les certitudes trop rapidement acceptées de façon à améliorer la qualité des projets.

[20] Il est ainsi arrivé qu'un enquêté féru de produits financiers ait prolongé l'entretien pendant 6 heures, car il a tenu à donner à l'enquêtrice un véritable cours pour compléter ses instructions, insuffisantes d'après lui. Une autre fois, un enquêté passionné de bricolage a d'abord testé les connaissances de l'enquêtrice (en lui répondant n'importe quoi) et n'a répondu correctement qu'une fois rassuré par la réaction de l'enquêtrice qui ne s'en était pas laissé compter.

[21] Soit parce que les résultats ont fait apparaître des ambiguïtés, soit parce que les effectifs concernés s'étaient révélés trop faibles : c'est en effet un travers souvent rencontré que de définir le questionnaire indépendamment de la taille de l'échantillon et de demander des détails intéressants mais qui nécessiteraient des échantillons beaucoup plus importants pour être observés de façon fiable. Ainsi la partie biographique de l'enquête IVQ détaille le passé migratoire des individus avec une finesse qui mériterait pour pouvoir être exploitée un échantillon de 20 000 ménages, pas de 3000. En l'occurrence, on a choisi de ne pas réduire le questionnement car le détail inutile non seulement ne compliquait pas la collecte mais semblait même généralement la faciliter.

2.2 La chaîne des tâches

Schématisée à grands traits, une **enquête ménage c'est successivement un échantillon, un questionnaire, une collecte terrain, des opérations de contrôle et d'apurement et une exploitation**. Chaque maillon de cette chaîne est susceptible de se révéler le « maillon faible » mettant en péril l'édifice tout entier, à cause de défauts spécifiques de qualité, défauts dont on va essayer maintenant de dresser la liste :

2.2.1 Un échantillonnage :

La première condition d'une bonne enquête, c'est de pouvoir disposer d'une bonne base de sondage correspondant au champ concerné. Les **méthodes par quotas** peuvent, bien appliquées, présenter des propriétés satisfaisantes pour des opérations de petite taille portant sur des pratiques ne présentant pas une concentration forte selon des critères non aisément observables. Mais les enquêtes présentées au Comité du Label recourent plutôt à des **sondages aléatoires**, soit parmi des listes d'individus bénéficiaires de telle ou telle mesure, soit parmi des bases de logement (Echantillon-maître, Echantillon Emploi…), soit enfin dans des bases de numéro de téléphone. Les qualités et défauts de ces trois types de bases sont bien connus :

2.2.1.1 fichiers « administratifs » :

Dans ce cas, ce sont surtout des problèmes dus à l'**obsolescence** ou à l'incomplétude. Il se peut aussi que le recours à un tel échantillonnage interdise quasiment certaines exploitations. Deux écueils dans ce registre. Le premier est connu sous le nom de « **sélection endogène des fichiers de stock** » (ou « stock sampling ») ; il se produit quand on veut mesurer des durées, par exemple de chômage, à partir d'un fichier de personnes présentant la caractéristique d'intérêt -par exemple un fichier de chômeurs à un moment donné. Les durées moyennes estimées sont biaisées par rapport à ce que l'on obtiendrait si on disposait pour l'ensemble des individus de l'histoire des phases de chômage, c'est à dire la durée moyenne du chômage pour toutes les personnes étant passées par le chômage au cours d'une période donnée : le fichier contient bien toute la population concernée pour les chômeurs récents, alors que, pour les chômeurs qui ont été concernés par le phénomène à une date plus ancienne, seuls y subsistent les chômeurs de longue durée. Si l'on rajoute à cela, l'existence évidente de troncatures à droite (on ne peut estimer la durée totale de chômage pour un chômeur en cours), on perçoit que l'on ne peut produire l'estimation souhaitée à partir de la source disponible, sans recours à des hypothèses très fortes -et non testables- sur la loi, inconnue, du phénomène. Or ce biais peut être important, puisque, dans un régime stationnaire avec taux de sortie de chômage constant, on double la vraie valeur. La seule façon de s'en sortir «proprement » est de bâtir un panel[22]. La deuxième difficulté à éviter concerne le cas où l'on désire faire de **l'évaluation de mesures** d'aides à partir de fichiers de bénéficiaires de cette aide, à cause de phénomènes d'endogénéité de la sélection des bénéficiaires : si les personnes qui se dirigent vers l'aide ont des caractéristiques individuelles supérieures (resp. inférieures) à celles de la population générale, on surestime (resp. sous-estime) l'efficacité de la mesure. Pour résoudre cette difficulté, deux conseils : rajouter à l'échantillon de bénéficiaires un échantillon témoin, tiré dans la population générale des non bénéficiaires et faire attention lors de l'exploitation !.

[22] Ce type de phénomène peut aussi se rencontrer dans des enquêtes utilisant un échantillon de logements habituel. On l'observera, par exemple, si on cherche à estimer la durée de vie d'un appareil ménager à partir de l'âge du parc des appareils en service. L'âge moyen est un estimateur biaisé vers le haut de la durée de vie, car à une date donnée, on n'observe, pour les générations d'appareils les plus anciennes, que les appareils les plus robustes, les seuls à être encore en service..

2.2.1.2 bases de logement :

Parmi les défauts usuels, on peut citer la **couverture plus ou moins complète** de certaines zones d'accès difficile (zones de montagne, îles…), ainsi que la **perte de précision créée par le souci de concentrer** la collecte, afin de faciliter le recrutement des enquêteurs et leur investissement dans la formation. Ces défauts, pour les bases INSEE, sont assez marginaux, voire tout à fait négligeables pour la plupart des sujets[23]. Il faut toutefois souligner que le champ retenu est celui des **ménages habitant dans des logements ordinaires** ; les personnes habitant dans des communautés (foyers de jeunes travailleurs, résidences universitaires, prisons, centres médicaux de longs séjours, asiles psychiatriques, hospices et maisons de retraite..) ne sont pas couvertes d'ordinaire ; pour certains sujets, ce défaut de couverture est rédhibitoire ; il faut alors envisager une collecte spécifique auprès des communautés (type enquête HID-prison[24] ou certaines opérations complémentaires à l'enquête Situations Défavorisées auprès des maisons de retraite, des asiles psychiatriques) ; même pour des sujets « classiques » comme l'étude du chômage, une extension serait sans doute un plus du point de vue de la qualité de l'observation de certaines populations spécifiques, les jeunes en particulier. Les personnes sans domicile échappent aussi aux investigations conduites à partir de ces bases. Dans le cas d'opérations spécifiques destinées à renseigner sur cette population, il faut monter des protocoles d'échantillonnage ad hoc (à base d'usage de fichiers administratifs et de tirages directs sur le terrain - cf opération Sans domicile) dont l'étude dépasse le cadre de ce travail.

L'absence de « trous » n'est pas la seule qualité souhaitable pour un échantillon. Même quand on échappe à ce défaut extrême, il se peut que l'échantillon ne fournisse pas une maquette fidèle de l'univers d'intérêt : les méthodes dites d'« équilibrage », mises en œuvre pour assurer les bonnes propriétés de « représentativité » de l'échantillon[25] ne peuvent en effet être mises en œuvre que sous contrainte (le nombre maximum de variables d'équilibrage, par exemple, est assez faible quand on veut tirer un nombre d'unités primaires conduisant à un réseau d'enquêteurs de taille gérable). Il faut aussi, bien entendu, que les variables objectives qui permettent de sélectionner l'éventuel sous-champ d'intérêt (par exemple les moins de 65 ans, les plus de 15 ans…) soient disponibles dans la base[26].

Une dernière étape, dont la qualité peut se révéler cruciale, se rajoute quand on cherche à tirer non pas des logements (i.e. des ménages au sens conventionnellement donné à ce vocable à l'Insee) mais que l'on cherche à **sélectionner des individus**. Traditionnellement la façon de faire était de recourir à la méthode de Kish[27] : l'enquêteur dispose d'un numéro, imprimé sur la fiche adresse du logement, et d'un tableau prérempli qu'il doit compléter à partir de la liste des individus ; le principe est de fournir, à l'intersection d'une ligne et d'une colonne spécifiées par la méthode, le numéro d'ordre de la personne à enquêter ; en théorie la méthode fournit des probabilités de tirage à peu près égales pour les divers individus. En pratique, on observe des défauts, avec une sous représentation des jeunes (Berthier, Caron, Néros 1999), ceci étant dû à des comportements « déviants » de certains enquêteurs qui renoncent à mettre dans le tableau de composition du ménage des individus pour lesquels ils anticipent que, s'ils étaient tirés, ils auraient du mal à réaliser l'entretien. Ce problème peut entacher aussi les autres méthodes classiquement disponibles pour réaliser ce tirage, la méthode des anniversaires (on choisit l'individu dont la date anniversaire est la plus proche, dans le futur ou le passé, de la date d'entretien) ou la méthode des prénoms (on prend l'individu dont l'initiale du prénom

[23] Compte tenu de deux effets en sens contraire, une augmentation de variance due à l'effet de grappe résultant de la concentration et un gain induit par les pratiques de calage a posteriori, pour l'échantillon-maître Insee, le « design effect » résultant est proche de 1.

[24] HID : Handicap, Incapacité, Dépendance.

[25] Un échantillon est dit « équilibré » vis à vis de certaines variables s'il permet d'estimer exactement (par l'estimateur d'Horvitz-Thompson) le total de cette variable dans la population. Deville et Tillé ont récemment développé l'algorithme du « cube » pour permettre la réalisation pratique d'échantillons équilibrés. Ces techniques ont été mises en œuvre récemment pour la constitution de l'échantillon-maître de l'Insee et pour la mise au point des échantillonnages du Recensement Rénové de la Population.

[26] Faute de quoi, il faut se résigner à une coûteuse opération en deux phases, la première servant de filtre pour sélectionner la sous-population concernée.

[27] Le succès de cette appellation, dont peu savent qu'elle renvoie au statisticien récemment décédé Leslie Kish a été tel que désormais on parle d'individu Kish pour désigner l'individu retenu quelle que soit la méthode employée !

est la première dans l'ordre alphabétique ou la plus proche de l'initiale du prénom de l'enquêteur …). Les performances de ces méthodes dépendent principalement de la qualité de remplissage de la liste des individus du ménage. Elles se distinguent sur le plan de la facilité de mise en œuvre (un argument étant la possibilité de mettre en œuvre la méthode au téléphone -donc sans avoir à ouvrir l'ordinateur- lors de la prise de contact, ce qui permet d'éviter des déplacements inutiles quand on doit interroger l'individu lui-même sans proxy), et par des caractéristiques relatives au degré d'aléatoire garanti : la méthode des anniversaires ne donne pas une représentation correcte des mois de naissance (ce qui peut être rédhibitoire pour certains sujets comme l'étude des cursus scolaires[28]) ; la méthode des prénoms ne conduit pas à une bonne représentation des genres et des générations, les initiales (voire la seconde lettre) n'étant pas uniformément réparties entre hommes et femmes, et selon les années de naissance (fluctuations de la cote des prénoms).

2.2.1.3 bases téléphoniques :

Dans l'état actuel de la législation, l'INSEE ne fournit pas d'échantillons à des sociétés de service, même quand celles-ci travaillent pour une administration ; rares étant les sociétés de sondage à disposer d'une base de ménages (comme le panel de la Sofres), voire d'un réseau d'enquêteurs compétents dans les entretiens en face à face, un grand nombre d'enquêtes sont réalisées au téléphone, sur les bases d'annuaires vendues par France Télécom[29]. Les qualités et défauts de ces bases sont bien connus et les évolutions récentes, qui auraient tendance à les aggraver, nécessitent des investissements techniques pour essayer d'y remédier : le problème le plus crucial est sans doute celui de la liste rouge[30], mais on peut aussi citer les problèmes de double lignes, et de portables, tous problèmes qui induisent des écarts part rapport à l'équiprobabilité de tirage[31]

Un échantillonnage ne se réduit pas à la disponibilité d'une base de sondage. Sa qualité dépend aussi du tirage, des éventuelles stratifications, différences de taux. Si la réflexion théorique peut guider efficacement à ce stade (par exemple consigne de tirer avec des taux plus forts là où le phénomène d'intérêt est dispersé), elle n'est pas toujours conclusive, en particulier pour répondre à la question lancinante qui oppose, en particulier, sociologues et statisticiens : faut-il mieux avoir un grand nombre de réponses de qualité médiocre ou un nombre plus réduit d'informations de haute qualité ? Quelques monographies fouillées ne seraient-elles pas préférables à un large échantillon de réponses truffées

[28] De plus, la CNIL s'est, dans le passé, montrée réticente à la collecte de l'information relative aux jours et mois de naissance, pour des raisons liées à la confidentialité : quand on connaît la date de naissance complète, l'identification indirecte des personnes est en effet grandement facilitée. Récemment, cette réticence semble s'estomper. Les ménages, eux, ont tellement l'habitude de décliner leur date de naissance dans les formulaires administratifs, qu'ils donnent le jour et le mois même quand on ne leur demande pas ; ne noter que l'année complique plutôt la tâche de l'enquêteur, d'où des risques d'erreur. Dans un panel, l'identification précise des individus, indispensable à un suivi longitudinal sans erreur, requiert la connaissance de la date de naissance complète. Elle figurera donc dans la nouvelle version du Tableau de composition des ménages.

[29] Parmi les avantages des enquêtes téléphoniques, le premier est leur coût, plus faible. Les inconvénients seront discutés au fil de ces réflexions : pas de support visuel pour les questions aux libellés longs ou complexes ; absence de maîtrise du contexte, risque que le ménage raccroche au premier « blanc »…

[30] Le problème est rendu ardu par l'attitude de France Télécom qui est extrêmement discret quant au nombre de personnes sur liste rouge (et sa version soft, la ligne orange) et à leurs caractéristiques. On soupçonne bien l'existence d'un biais difficile à corriger mais sans pouvoir en quantifier l'ordre de grandeur et la nature exacte. De plus la CNIL a tendance à ne pas admettre systématiquement l'emploi de méthodes correctrices, comme celle qui conduit à générer des numéros aléatoires, de façon à respecter ce qu'elle considère comme une manifestation explicite de la part du ménage de la volonté de ne pas être dérangé. Ce n'est que si le concepteur peut prouver l'existence d'un biais particulier au sujet étudié, les personnes d'intérêt ayant une bonne raison d'être particulièrement nombreuses à recourir à la liste rouge, que l'on peut obtenir des dérogations (exemple : l'enquête sur les violences envers les femmes, avec la corrélation forte entre le fait d'avoir subi des violences et le fait de s'être inscrite sur liste rouge).

[31] L'assimilation entre base téléphonique et enquête téléphonique n'est d'ailleurs pas inévitable. Il se peut que l'on soit conduit à partir d'une base téléphonique, à réaliser des enquêtes en face à face. La base n'est là que pour pallier l'absence d'une liste de logements. Elle sert à déterminer un point de départ : on peut alors choisir d'interroger le ménage habitant le logement correspondant au numéro de téléphone, mais on peut aussi faire accomplir par l'enquêteur un circuit -défini par des règles précises- pour définir le logement à enquêter à partir de ce point de départ (méthode dite « des itinéraires »), ce qui a l'avantage d'éradiquer presque parfaitement le problème de liste rouge.

d'erreurs de mesure ? Une réponse catégorique dans un sens ou dans l'autre serait caricaturale : enquêtes statistiques et monographies devraient être plus souvent utilisées conjointement pour éclairer des phénomènes complexes (d'où l'intérêt d'une réinterrogation par un chercheur d'un sous-échantillon de répondants à une enquête statistique) ; et, au sein des enquêtes statistiques, l'arbitrage quantité-qualité devrait être moins systématiquement qu'actuellement en faveur de la quantité (défaut induit par le fait que l'on sait mesurer l'impact en variance d'un accroissement d'échantillon alors que l'on ne sait pas quantifier celui d'un gain de qualité !).

2.2.2 un questionnaire :

Un « bon » questionnaire doit résoudre une sorte de quadrature du cercle : les questionnaires doivent être précis tout en restant courts et intelligibles par toutes les strates de la population, sachant qu'au sein de celle-ci certaines personnes immigrées n'ont à leur vocabulaire qu'une petite centaine de mots de français[32]. Au palmarès des défauts les plus fréquemment relevés, on évoquera les **concepts mal définis**, **les tâches impossibles** les **formulations complexes mal comprises,** les **descripteurs importants manquants**; ainsi que le **choix mal maîtrisé des périodes de référence.**

2.2.2.1 les concepts mal définis :

Ils s'avèrent d'autant plus pernicieux qu'ils se cachent derrière des mots courants que tout le monde croit comprendre mais derrière lesquels chacun met un contenu différent. La liste en est si longue qu'il est impossible d'être exhaustif : ami, famille, patrimoine, revenu, salaire, démarche, emploi, âge au premier emploi, sport, maladie, fièvre, livre ne sont que quelques exemples parmi d'autres de ces mots d'apparence anodine dont on ne saurait trop se méfier. Afin d'expliciter un peu la nature des difficultés rencontrées, on peut citer quelques travaux de terrain qui ont été conduits et qui illustrent cet aspect.

Pour la formation des enquêteurs à l'enquête sur les Actifs financiers, on a réalisé quelques « micros-trottoir » afin de faire expliciter ce que les gens considéraient comme leur patrimoine : tel homme, jeune, pense immédiatement patrimoine financier et c'est tout ; telle femme d'artisan englobe dans son patrimoine, son mari, sa camionnette et son chien alors qu'un troisième considère que son patrimoine c'est son métier, sa santé. Interrogés quant au sens du mot démarche, dans une opération complémentaire à l'enquête Emploi, certains considèrent que lire les offres d'emploi dans les journaux, ou mettre, dans un commerce de quartier, un avis de recherche de bébés à garder ne constituent pas des démarches faites pour trouver un emploi (alors que les consignes d'Eurostat et du BIT les considèrent comme telles), car pour eux faire une démarche signifie se déplacer auprès d'une administration. La plupart des enquêtes peuvent servir à alimenter la liste des exemples dans ce registre : interrogés sur leur pratique du sport, certains oublient les marches lors de randonnées si elles sont faites dans l'esprit de « se balader » et on ne sait pas très bien à partir de quand les baigneurs deviennent des nageurs. Certaines affections chroniques comme l'asthme, les lombalgies, les allergies sont parfois perçues comme des maladies, parfois non. Le port de lunettes génère le même type d'ambiguïtés. La nouvelle enquête sur les pratiques culturelles et sportives présente un grand nombre de cas où cette difficulté est susceptible de se manifester. En marge de la préparation de cette opération, on avait préparé un petit questionnement de façon à faire préciser, une fois l'enquête remplie, ce qui avait été pris en compte par la personne dans l'évaluation faite. On avait centré l'opération sur les mots ou les expressions qui semblaient a priori particulièrement polysémiques : émissions (de télévision) sur les arts et la culture ; visite de monuments historiques, festival ; lire un

[32] En cas de non maîtrise du français, l'enquêteur essaie de trouver dans l'environnement de l'enquêté quelqu'un susceptible de faire la traduction, en général un enfant, parfois un voisin. Lors de l'opération méthodologique faite autour du concept d'emploi et de chômage (cf infra), la personne ayant conduit les entretiens a constaté qu'une personne d'origine étrangère ne savait pas ce que signifiait le mot « courrier » et l'écoute de l'enregistrement d'un entretien avec un chômeur maghrébin a révélé qu'il confondait sans cesse « chercher » et « trouver ».

livre. Malgré la petitesse de l'échantillon testé (une dizaine de personnes), les enseignements ont été très éclairants[33]. A titre d'exemple, citons seulement que si certains considéraient que la Fête de la Musique ou les Fêtes de la bière à Munich étaient des festivals, d'autres étaient d'un avis contraire. Pour certains visiter un monument nécessitait qu'on soit entré dans une construction bâtie par l'homme alors que d'autres avaient une vision plus extensive, incluant par exemple des jardins ou des lieux historiques (comme les plages du débarquement…). A la question sur les fréquences de visite, certains comptaient pour 1 l'ensemble des visites réalisées un jour donné dans une ville donnée alors que d'autres comptaient chaque monument…Un documentaire sur la vie animale, une biographie d'un chanteur de variétés ou une émission sur le cancer étaient classés par certains dans les émissions sur l'art ou la culture, alors que d'autres avaient une vision restrictive incluant les arts plastiques, le cinéma, la littérature ou l'histoire mais écartant la géographie, l'ethnologie ou les sciences physiques.
Pour réduire ce flou qui rend les interprétations des résultats hasardeuses, puisqu'on ne sait si en présence d'une disparité, on commente une **différence réelle de pratique** ou une **différence dans les contours du concept**, plusieurs solutions se présentent, aucune n'étant une panacée et toutes étant coûteuses, en temps de questionnement et en formation des enquêteurs en particulier. On peut renoncer à imposer à l'enquêté un contenu prédéfini et préférer garder la définition indigène propre à la personne ; dans ce cas il importe de lui faire préciser le sens donné. Dans l'enquête Contacts, c'est ce qui avait été choisi pour dénombrer le nombre d'amis : dans une première phase, on demandait à l'enquêté de définir ce qu'il considérait comme un ami, avant de les lui faire dénombrer. Ceci conduit à un nombre d'amis très différents de ce qu'on obtient d'une simple demande de quantification, sans l'étape préalable de définition (Godechot). Dans l'enquête réalisée en Pologne sur les conditions de vie, une procédure en deux étapes a été employée pour recenser le nombre de voisins. C'est certainement la solution à adopter lorsqu'il n'existe pas de définition « estampillée » du concept. Dans le cas de nomenclatures utilisant des postes susceptibles d'être définis par énumération des éléments constitutifs, la solution passe par un usage judicieux des exemples concrets. Il faut **éviter les mots généraux, abstraits désignant l'ensemble du poste de façon synthétique**, solution sans doute adaptée pour un nomenclaturiste mais hasardeuse pour l'enquêté « lambda ». Plutôt que de formuler la question sous la forme « lisez vous des quotidiens nationaux d'information générale » voire sous la forme « lisez vous des quotidiens nationaux d'information générale (ex : Le Monde, le Figaro) » -avec le risque que la parenthèse ne soit pas lue par l'enquêteur-, on préfèrera la formulation « lisez vous le Monde, ou le Figaro, ou un autre quotidien national d'information générale », avec des exemples choisis de façon à dessiner les contours du poste de façon dénuée de toute ambiguïté, en faisant aussi en sorte de citer les cas les plus fréquents[34][35]. En tout état de cause, éviter au maximum les filtres qui conditionnent l'entrée dans un questionnement détaillé à une réponse à une question générale utilisant un concept ambigu. Par exemple, si l'on demande aux personnes si elles ont souffert d'une maladie lors des 12 derniers mois, et que c'est seulement en cas de réponse positive que l'on détaille s'il s'agissait de telle ou telle maladie, le recensement étant dès lors conduit à partir d'une longue liste d'affections, on obtient un résultat très différent de ce que l'on obtient par interrogation détaillée non filtrée. C'est ainsi que dans l'enquête Emploi en continu, on a remplacé la question « Avez-vous fait des démarches, si oui lesquelles » par « avez-vous fait…suivi de la liste exhaustive des démarches.. », et ceci a réduit le nombre de chômeurs qui déclaraient ne pas avoir fait de démarches, et qui, à ce titre, étaient radiés du concept de chômage au sens du BIT.

On peut rapprocher de cette difficulté l'ambiguïté que l'on relève souvent dans les questionnaires autour de l'utilisation du mot « vous ». En quelque sorte « vous » est le premier concept mal défini, celui que l'on rencontre dans tous les questionnaires, qui surgit quelle que soit la thématique particulière de l'opération. La première difficulté, banale quand on l'exprime, est encore source de

[33] Cet approfondissement méthodologique a été posé à l'ensemble des enquêtés de l'Ile de France en mai-juin 2003 ; actuellement seuls les résultats du test préparatoire ont été exploités.
[34] Les tirages étant connus, dans le cas des journaux, c'est particulièrement facile.
[35] Il faut toutefois veiller à éviter que l'enquêté, ne trouvant pas son cas dans la liste des exemples, se réfugie dans l'éventuelle modalité « autres ». Le danger vient de ce que l'enquêté ne comprenne pas qu'il s'agit d'exemples emblématiques et croie qu'il s'agit d'une liste fermée excluant tous les cas non cités. Ce type de comportement a pu être constaté dans des tests du recensement. Il est moins à craindre dans une enquête en face à face, l'enquêteur étant là pour redresser le tir en cas de besoin.

problèmes dans de nombreux questionnements : il s'agit seulement d'éviter la confusion entre le « vous » personnel qui désigne l'individu auquel on s'adresse et le «vous » collectif qui désigne l'ensemble du ménage auquel il appartient. Un exemple parmi d'autres de formulation ambiguë : « possédez-vous un ordinateur ? »[36]. Le travail de terrain a fait émerger des niveaux plus subtils de problèmes. Dans une enquête expérimentale sur les comportements face au risque et au temps, une question demandait à l'enquêté si, suite au problème de la vache folle, il avait modifié sa consommation de viande. Une femme a répondu qu'elle ne savait pas quoi répondre : en effet, en tant qu'individu, elle n'était pas concernée, car végétarienne de longue date, elle ne consommait jamais de viande et donc la crise n'avait rien changé, mais, en tant que maîtresse de maison, il lui arrivait de servir de la viande à ses invités, et dans ce rôle, elle avait changé de comportement car elle avait remplacé le bœuf par d'autres viandes. Vu le caractère expérimental de l'opération, tout ce raisonnement a été noté par l'enquêteur et sauvegardé ; ce n'aurait pas été le cas dans une enquête plus industrielle, ce qui aurait posé un problème de qualité. A une question similaire sur les changements de pratique suite aux problèmes de Sida, une enquêtée d'origine maghrébine, de milieu social fort modeste, a fait preuve d'une subtilité de même nature : en tant que femme, a-t-elle dit, elle est fidèle et n'a donc pas changé ses habitudes, mais en tant que mère elle a pris sur elle de parler à ses enfants de sujets que culturellement il ne lui était pas permis d'aborder[37]. L'usage du « vous » présuppose l'unité de l'individu alors que les sociologues insistent sur la pluralité des identités de chaque être humain (Grumbach 1982), et pas seulement dans le cas de schizophrénie. Trop souvent, on postule l'unicité de la réponse possible, ce qui force le sujet à choisir une de ses identités, sans que l'on sache les critères mobilisés pour effectuer cette sélection …Ces deux exemples confirment que la réalité est plus complexe, en même temps qu'ils montrent la difficulté à prévoir ce que les enquêtés sont capables de concevoir, les nuances qu'ils peuvent percevoir. S'il est fréquent d'épingler le concepteur qui, du fond de la tour d'ivoire constituée par son bureau parisien, aurait tendance à penser que chacun est capable de maîtriser un langage ardu, il est plus rare de lui conseiller de ne pas trop vite conclure à son incapacité à comprendre un discours subtil ; ce serait pourtant aussi judicieux !

Si le flou dans les concepts est si répandu, c'est qu'en général autour d'un «noyau dur » aisé à identifier et ce de façon consensuelle, s'agrègent des situations proches que l'on rattache conventionnellement à ce noyau dur. C'est au niveau de cette assimilation que l'on observe des différences entre ménages. Prenons l'exemple de l'opposition entre les concepts de salarié et d'indépendant. Si les noyaux durs, en quelque sorte les archétypes des deux catégories, sont faciles à reconnaître, la limite qui sépare les cas intermédiaires que l'on va assimiler aux salariés et ceux que l'on va considérer comme des indépendants est plus difficile à définir et par là-même à respecter par tous (récemment, dans le commerce, se sont développées diverses formes de gérance, de franchises qui font que la personne a certaines caractéristiques des « vrais » indépendants tout en ayant aussi des caractéristiques de salariés)[38].

[36] Derrière ce problème, se cache en réalité tout un monde de complexité, celui de l'agrégation des préférences individuelles en préférences d'un collectif, le «ménage ». La théorie microéconomique s'est penchée sur la question, concluant à l'impossibilité d'agréger les préférences en toute généralité (théorème dit « d'impossibilité d'Arrow »). Obtenir un ménage qui ait les « bonnes » propriétés (transitivité en particulier) suppose que soient réunies certaines conditions (existence d'un « dictateur » -éventuellement «tournant »…). En général donc l'agrégation est impossible, et recenser les pratiques, les opinions au niveau du ménage n'a pas de sens : il faut individualiser. Les résultats du panel européen montrent que les réponses aux questions portant sur la satisfaction (emploi, ressources, santé…) sont souvent hétérogènes au sein d'un même ménage, même pour des sujets, comme le niveau de vie, ayant une composante collective indéniable

[37] Elle a même déclaré qu'elle avait enfreint le Coran, qui enjoignait de ne pas parler de ces sujets.

[38] Un exemple sera ultérieurement développé, autour de la collecte des variables utiles au chiffrement de la PCS. On dispose en effet sur cette variable d'un matériau riche, provenant du rapprochement des sources Emploi et Recensement (Guglielmetti 2002).Les zones de flou sont bien identifiées. Elles proviennent de la difficulté à obtenir de l'information fiable sur certains concepts délicats comme la fonction ou la position professionnelle. Dès la mise en place du questionnement, certaines difficultés avaient été signalées par les nomenclaturistes ; mais on est aussi certainement dans une situation où le flou a augmenté au cours du temps, avec le changement dans les procès de production, dans la force régulatrice des conventions collectives, voire avec la tertiarisation de l'emploi.

2.2.2.2 les tâches impossibles :

La plupart des enquêtes ne se contentent pas de recenser l'existence de pratiques ; elles cherchent à préciser une intensité, une fréquence. C'est dans ce registre que l'on trouve la plupart des exemples que l'on peut ranger dans cette catégorie des tâches « impossibles », de ces tâches que le concepteur statisticien demande à l'enquêté alors que souvent il serait bien en peine de le faire pour son propre cas. En général, on demande à l'enquêté de sommer sans aide et sans calcul intermédiaire des grandeurs qu'il n'a pas l'habitude de calculer, comme les «revenus du patrimoine », le « revenu superbrut », le montant «les dépenses de vacances », le nombre de maladies qu'il a eues …ce qui combine en général problèmes de mémoire et incertitude sur les frontières du concept[39].

Un remède semble bien s'imposer : **décomposer**. Il faut procéder par étapes, travailler au niveau de chaque composante homogène plutôt qu'au niveau du concept global, et se donner les moyens de raviver la mémoire au moyen d'un calendrier précis ; l'ordinateur est beaucoup plus performant pour faire additions et multiplications que l'enquêté qui n'est pas forcément un génie du calcul mental[40]. Si l'on ne veut pas passer ce temps minimal qu'il faut pour obtenir une valeur relativement précise [41], on a toujours la solution de proposer au ménage un système de tranches, ce qui a l'avantage de lui indiquer l'ordre de grandeur des approximations admissibles. D'autres solutions ont été essayées : ainsi, dans l'enquête Modes de Vie-Production domestique, on avait introduit l'usage de fourchettes pour mesurer le temps passé aux diverses activités de bricolage, de jardinage ou autres travaux d'aiguille. Au lieu de réduire la difficulté, on l'avait de fait multipliée par deux, certains enquêtés (enquêteurs ?) ayant compris qu'on leur demandait entre quel maximum et minimum précis variaient les temps passés[42]. On rencontre aussi des formulations comme «environ » ou «en moyenne », la première étant préférable à la seconde car elle affiche clairement l'intention de se contenter d'une évaluation grossière alors que la seconde pourrait être perçue comme le souci de mesurer une moyenne au sens mathématique du terme, ce qui a peu à voir avec ce que l'on recherche ; mais la tâche du ménage n'en est pas vraiment simplifiée, car il n'est en rien guidé en matière d'ordre de grandeur de l'approximation acceptable[43].

Les travaux conduits en marge de la préparation de l'enquête sur les pratiques culturelles et sportives peuvent à nouveau fournir un exemple. Dans l'enquête, une cinquantaine de pratiques étaient étudiées, avec, pour chacune, une demande d'évaluation du nombre de fois où la pratique avait été faite sur douze mois. On a comparé la réponse spontanée, avec ce qu'on obtenait, dans une deuxième phase, en remontant le temps, mois par mois, à partir de la date de l'enquête, en demandant pour chaque mois le nombre de fois où la pratique avait été faite (on avait choisit la visite de monuments historiques) et le

[39] On pourrait aussi classer dans cette rubrique les questions demandant de se remémorer les opinions, les anticipations que l'on avait à certaines époques passées, par exemple pendant l'enfance, avec tous les risques de rationalisation a posteriori, de reconstruction que cela entraîne, ainsi que celles visant à établir des calendriers rétrospectifs sans que soient prévues les étapes nécessaires au « rafraîchissement » de la mémoire.

[40] Comme toute médaille, celle-ci a son revers : si le répondant arrondit par trop chacune de ses réponses élémentaires, on peut arriver à des sur(sous)-estimations de forte ampleur au niveau de l'évaluation globale, les arrondis n'ayant pas toujours la bonne idée de se compenser. Dans l'enquête Modes de Vie déjà évoquée, on avait des évaluations de temps passé par grandes familles d'activité, puis dans un deuxième temps des évaluations détaillées activité fine par activité fine. Les incohérences entre les premières et la somme des secondes étaient nombreuses, sans que l'on puisse conclure à la supériorité de l'une ou de l'autre façon de procéder ; mais il faut dire que l'on était dans un domaine particulièrement complexe, celui de l'observation des temps passés par l'intermédiaire de questionnements rétrospectifs. Quand on enquête sur la durée du travail, on observe aussi que les déclarations de durée globale diffèrent de la somme des déclarations de durées élémentaires.

[41] Sans avoir la naïveté de croire que l'on peut atteindre une précision extrême par enquête !

[42] Par contre l'usage des fourchettes marche quand on essaie d'appréhender les valeurs de marché de l'appartement ou du patrimoine immobilier possédé, sans doute parce que le ménage, sur ce thème, est habitué à l'usage d'évaluations plancher et plafond.

[43] La façon dont les gens arrondissent leurs réponses peut être décelée à partir des déclarations de revenus ou de salaires, quand elles sont demandées en clair. A l'enquête Emploi, on a pu mesurer qu'environ 40 % des montants étaient arrondis aux 100 f les plus proches -désolé, mais l'étude date de l'ère préeuro- et aussi 40% aux 500 f les plus proches, la façon d'arrondir étant assez peu dépendante du montant déclaré, ce qui conduit à une erreur relative très importante dans la zone des bas revenus : n'oublions pas que 500 f c'est presque 10% du SMIC ! Une autre donnée présentant des phénomènes d'arrondis importants concerne l'âge des appareils ménagers en service, avec des pics très marqués aux multiples de 5. Il est souvent illusoire de croire que l'on obtient davantage de précision avec un questionnement en clair qu'avec l'usage de tranches.

détail des monuments visités. Les personnes interrogées n'ont jamais pu remonter à 12 mois dans le passé ; certains avaient oublié les visites -privées- lors de déplacements professionnels à l'étranger, d'autres avaient oublié les visites faites dans la ville ou la région habitée, ne comptant que ce qui s'était passé au loin, lors des vacances. Personne n'est retombé (sauf un cas où deux erreurs s'étaient compensées) sur l'évaluation initiale et tous ont conclu au fait que la déclaration spontanée n'avait aucune valeur.

On peut aussi trouver dans la littérature sur le sujet des exemples parlants montrant l'augmentation du pourcentage des gens malades en fonction de la longueur des listes d'exemples cités et du détail des évaluations élémentaires prévues.

Dans les enquêtes sur les vacances, on rencontre fréquemment une question qui, sous des dehors anodins, cache un véritable chef d'œuvre de tâche impossible ; c'est la question « Quelles ont été les dépenses de vacances de votre ménage ?». Tout y est : les problèmes conceptuels (les cadeaux souvenirs sont-ils ou non des dépenses de vacances ?…), les problèmes d'agrégation sur plusieurs registres simultanément (sommation de dépenses d'alimentation, de vêtements, d'hébergement, de transports, de visites faites par plusieurs individus), chacun mettant sans doute en œuvre un processus de remémoration différent, problèmes d'agrégation rendus encore plus délicats par le fait que ces débours peuvent avoir été effectués à des moments différents (arrhes versés des mois avant le déplacement, tickets achetés sur place, paiements par cartes de crédit débités quelques semaines après..). Et pour faire tout cela le répondant a au plus une minute ou deux ! La quantification des jours d'absence au travail sur une année, du nombre de périodes de chômage ou de maladie sur l'ensemble de la carrière professionnelle fournissent aussi des exemples de totaux difficiles à produire.

Une forme légèrement différente de tâche impossible apparaît quand on essaie de forcer les enquêtés à réfléchir dans une unité autre que celle à laquelle il se réfère spontanément. Lors de Modes de Vie - Production domestique (enquête déjà citée), il s'agissait, dans une partie du questionnaire, de mesurer en kilogrammes la production familiale des potagers, légume par légume, fruit par fruit, ou celle des activités de la chasse et de la pêche. Les tests ont montré que l'on ne pouvait obtenir cette évaluation. Les enquêtés étaient incapables de faire la conversion : dans le meilleur des cas, on obtenait une évaluation de la production dans les unités spontanément utilisées : on a récolté une brouette de carottes, un saladier de fraises ; on a élevé trois lapins, pêché trois bars et deux soles…On a donc accepté ceci, et effectué, ex post, au sein d'une unité de chiffrement, le transcodage à partir d'équivalences obtenues par avis d'expert ou mesures moyennes…Il faut **toujours aller dans le sens du ménage** ; c'est lui l'élément incontournable de l'opération et **tout ce qui lui simplifie la tâche doit être mise en œuvre**, et ce **même si dans ce jeu « à somme nulle » ce que le ménage gagne, le concepteur le perd**[44]. Une stratégie du concepteur qui serait de rejeter sur le ménage la charge des tâches qu'il souhaite lui-même éviter -soit pour se faciliter l'exploitation, soit pour raccourcir le questionnaire- serait forcément une stratégie perdante. Un exemple éclairant a été observé lors de la mise en place du panel européen, à propos de la collecte des revenus du patrimoine. Dans le questionnaire européen, il y avait une seule question, posée à chaque individu «quel a été le montant de vos revenus du patrimoine au cours des douze derniers mois ?». Le patrimoine n'était ni défini, ni décomposé entre patrimoine mobilier et immobilier ; rien n'était prévu pour traiter les possessions jointes par plusieurs individus. Les tests ayant montré la mauvaise qualité de ce qui serait ainsi collecté, on a conçu deux pages de questionnaire, détaillant en particulier tous les éléments du patrimoine, demandant outre les revenus, les montants possédés pour pouvoir réaliser des imputations dans le cas de non réponses sur les revenus. La première vague du panel a utilisé cette version pour les revenus 1993 et les revenus des 9 premiers mois de 1994. Afin de simplifier, de gagner du temps, le questionnement a été réduit de moitié l'année suivante, pour la seconde vague. Les montants observés sur 1994 (année complète) ont été plus faibles que les montants observés l'année précédente sur les 9 premiers mois.

[44] Ceci est d'ailleurs aussi valable pour les développements précédents autour du concept flou : on laisse au ménage le soin de trancher par défaut là où l'on n'a pas su (ou voulu) décider ex ante.

2.2.2.3 les formulations complexes mal comprises :

Le débat est animé sur ce point, et sur son corollaire, à savoir la politique en matière de **reformulation**. Deux écoles s'affrontent. L'une, radicale, condamne toute reformulation, l'autre, plus nuancée, en reconnaît le caractère inévitable, tout en essayant de définir le cadre de ce qui est admissible et de ce qui ne l'est pas. Il est, me semble-t-il, unanimement reconnu qu'aucune formulation ne saurait convenir à l'ensemble de la population ; tout au plus peut-on espérer que la formulation retenue soit comprise d'emblée par une forte proportion (voire la majorité) des enquêtés. L'enquêteur aurait donc en général, sinon à reformuler, du moins à expliquer ce qui est attendu. C'est là que les avis divergent : les « radicaux » pensent que l'enquêteur ne peut reformuler de façon neutre, et qu'il ne peut que biaiser le sens de la question. Mieux vaudrait un « ne sait pas » qu'une réponse ininterprétable faute de connaître la question à laquelle elle correspond réellement. A l'opposé, d'autres pensent que la reformulation n'ajoute pas vraiment de bruit supplémentaire, et que de toutes façons on n'est jamais sûr de ce qui a été compris de la question. Une voie médiane, inexplorée à ce jour, me semble ouverte par les possibilités techniques modernes liées à l'usage de CAPI : le concepteur pourrait prévoir tout une série de reformulations « licites » classées par ordre de simplification croissante : on aurait ainsi, à défaut de compromis idéal, une solution de «moindre mal » entre la nécessité de reformuler pour être compris par tous et le danger de modifier le registre de la réponse : reste à faire preuve d'imagination pour trouver les reformulations adaptées. Un exemple de reformulation parmi d'autres m'avait naguère frappé, car je l'avais entendu lors d'une visite à la division Enquête Ménages d'une Direction régionale : ayant à poser dans l'enquête mensuelle de conjoncture (enquête téléphonique) la question « à votre avis, compte tenu de la situation économique, est-il opportun d'épargner » , face à un enquêté qui ne comprenait pas, l'enquêteur (en l'occurrence un agent de la DEM) a fini par demander si la personne avait épargné…ce qui est indéniablement plus simple mais n'a pas le même sens. On comprend alors que la série obtenue, pour cette question, ait toujours été considérée par les conjoncturistes comme difficile à interpréter.

Le degré de complexité acceptable dépend du mode d'entretien, le téléphone étant particulièrement peu apte à l'usage de formulations longues ou à tiroir. La solution est à nouveau dans le recours à une procédure par étapes, décomposant la difficulté en une série de difficultés moindres ; en aucun cas, la solution ne saurait résider dans une simplification abusive de la question. Si la question initialement prévue se révèle complexe, c'est que ce que l'on veut mesurer nécessite que l'on apporte des précisions. Les supprimer ne peut conduire qu'à un abâtardissement de la réponse. Dans les situations d'entretien en face à face, l'usage de supports (cartes codes) permet d'aider à la compréhension de questions (ou de modalités de réponses) complexes[45]. Aux concepteurs d'une enquête récente destinée à mesurer les impacts en matière de santé de l'explosion de l'usine toulousaine AZF, il était aisé de recommander de scinder en plusieurs sous-questions la question proposée «Dans les mois qui ont suivi l'explosion, avez-vous éprouvé au moins une des difficultés suivantes : des images ou des souvenirs de l'explosion répétitifs, des troubles du sommeil, des difficultés de concentration, une irritabilité, des trous de mémoire concernant l'explosion ? »

[45] Une telle complexité dans les modalités de réponse apparaît souvent dans les questions destinées à cerner les raisons d'une pratique. On énumère en général une longue liste de raisons, en demandant à l'enquêté de choisir la (ou les) raison(s) principale(s). C'est alors que l'on peut observer les effets de «recensy » ou «primacy » évoqués supra. Il est de loin préférable de demander après chaque item si cet item a concerné ou non le ménage, puis lui faire choisir, dans un deuxième temps, une fois tous les items passés en revue, le plus important. Evidemment cette façon de procéder est plus chronophage. Pour les questions de moindre importance, qui ne sont pas centrales dans l'enquête, il est préférable de recourir à une question « semi-ouverte » -innovation sémantique et pratique faite lors de la mise au point de l'enquête Emploi en continu- : l'enquêteur ne lit pas les modallités de réponse, mais les utilise pour coder instantanément la réponse «en clair » de l'enquêté. Solution rapide, mais le risque de ne récolter que les raisons les plus évidentes, et par là les plus banales, est augmenté.

2.2.2.4 les descripteurs importants manquants :

Un autre élément de qualité d'un questionnement réside dans la richesse du corpus des descripteurs sociodémographiques collectés. Or il arrive encore que l'on découvre tardivement des cas d'enquêtes où le facteur explicatif principal est omis : c'est en général du revenu qu'il s'agit, car souvent perdure l'idée que le sujet, tabou, ne peut être abordé dans une enquête, ce qui n'est plus vrai depuis plusieurs décennies[46]. Pour éviter ce type de mauvaise surprise, un effort de rationalisation a été effectué, par le biais de la réflexion sur le «**tronc commun aux enquêtes ménages»** ou **« tableau de composition du ménage »**[47]. Des sociologues, des économistes et des démographes se sont réunis pour fixer l'ensemble des descripteurs à introduire, ainsi que la (ou les) forme(s) recommandée(s), ceci ayant aussi un effet bénéfique en matière de comparabilité. Notons enfin que certaines variables sont importantes non pas tant pour leur force explicative directe mais parce qu'elles sont susceptibles de fournir de bons « instruments » pour le traitement économétrique des problèmes d'endogénéité (Robin)[48]. Il faut être particulièrement vigilant à ne pas sacrifier dans le processus de calibrage du questionnement ces variables dont l'utilité peut ne pas apparaître à première vue, mais qui sont néanmoins indispensables à une exploitation correcte des résultats.

2.2.2.5 le choix mal maîtrisé des périodes de référence :

Fréquemment, la collecte des pratiques se fait sur un laps de temps dit « période de référence » ; le choix judicieux de cette période est aussi un facteur de qualité. Le bon choix doit constituer un juste milieu entre le souhait de ratisser large pour récupérer le maximum de pratiques rares, les limites de la mémoire, le risque de confusion lié à l'usage de périodes différentes d'une question à l'autre, sans oublier le calage vis à vis d'éventuels repères dont disposerait le répondant. Actuellement, le choix n'est que trop rarement soigneusement justifié. Des pratiques opposées coexistent. Ainsi, dans l'enquête sur les Budgets des ménages, les dépenses de consommation sont collectées sur une période de quinze jours (carnet), de un mois, deux mois, six mois, voire un an (dans le questionnaire rétrospectif) selon le rythme d'achat supposé prévaloir dans l'ensemble de la population. La crainte des erreurs est passée au second plan par rapport au souci de caler sur les possibilités de la mémoire. Dans l'enquête sur les pratiques culturelles et sportives, c'est l'option inverse qui a prévalu, toutes les pratiques étant recensées sur 12 mois quel que soit leur rythme standard. Les conséquences du choix d'une période de référence ne sont pas uniquement de nature technique, car il se peut que ce choix ait une incidence directe sur le concept effectivement mesuré. Prenons le cas d'une interrogation sur le revenu. Ce que l'on mesure par une question sur le revenu mensuel -ou plutôt sur le revenu du dernier mois- ne renvoie pas à ce que l'on mesure avec une question sur le revenu annuel (le premier n'est pas 1/12 du second) : les primes irrégulières ne sont pas incluses de la même façon, le recours à une définition fiscale du revenu est plus probable quand on se réfère à l'année (à cause de la prégnance de la déclaration annuelle de revenus[49]). Se restreindre aux

[46] Ce qui ne veut pas dire que la qualité de la variable obtenue soit irréprochable. Elle est entachée d'un grand nombre d'erreurs. Néanmoins, telle qu'elle est, pour l'étude de toutes les pratiques liées à la consommation, à l'épargne, elle se révèle, dans un modèle économétrique toutes choses égales d'ailleurs, un facteur explicatif beaucoup plus discriminant que la catégorie sociale, son substitut habituel.

[47] Dit TCM.

[48] Attention, ceci ne signifie pas que n'importe quelle question peut fournir un instrument valable ! Trouver des « bons » instruments, est même la partie la plus délicate d'un traitement de l'endogénéité, puisqu'il s'agit de trouver une dimension qui, alors que l'on est en présence de deux phénomènes si liés qu'ils en apparaissent simultanés, joue un rôle dans l'explication de l'un sans être pertinent pour le second. C'est bien l'exploitation des tests -et elle seule- qui peut renseigner sur la capacité d'une variable à jouer le rôle d'instrument, en passant avec succès les tests de qualité prescrits par la théorie.

[49] Interrogé lors du panel européen, un retraité a répondu spontanément aux deux questions sur la montant de ses revenus, la première portant sur l'ensemble de l'année précédente, la seconde sur les neuf premiers mois de l'année en cours. Il a été alors frappé par l'évolution que ses réponses traduisaient alors qu'il savait pertinemment que sa retraite n'avait pas évolué. Avec l'enquêteur, ils se sont acharnés, plusieurs minutes durant, à essayer de comprendre ce qui s'était passé ; ils y sont parvenus et la raison était la suivante : pour la déclaration annuelle, l'enquêté s'était spontanément référé à la déclaration fiscale (en l'occurrence après déduction de la réduction que lui valait le fait d'avoir eu trois enfants) alors que, cette référence étant indisponible pour l'année en cours, il s'était basé sur le versement perçu (donc sans tenir compte de la réduction fiscale). Mais pour un cas où l'erreur a été découverte, combien sont passés inaperçus ?

pratiques réalisées sur le laps de temps réduit que représente la période de référence est aussi un moyen commode -et correct- de limiter le nombre des pratiques à étudier de façon détaillée et donc d'éviter une explosion du temps de questionnement. Cependant, il peut arriver qu'une telle restriction ne suffise pas à assurer que l'on reste dans les limites du temps imparti.[50] Pour sélectionner davantage, il est fréquent de donner comme consigne de ne s'intéresser qu'aux dernières occurrences (la dernière, les trois dernières[51]) ; cette pratique ne garantit pas l'absence de biais. Les défauts sont encore plus flagrants si on demande une sélection de l'item le plus important, ou de plus longue durée. La seule méthode à recommander est une sélection aléatoire dans l'ensemble des pratiques ayant eu lieu pendant la période de référence.

Ces quelques grandes rubriques sont loin d'épuiser tout ce qui pourrait être évoqué dans un manuel exhaustif des bonnes pratiques en matière de questionnement. Ce sont seulement les points essentiels. Brièvement, on peut citer d'autres points qui mériteraient attention et qui sont, aussi, souvent **sujet à querelles d'experts**. Une première réflexion serait relative à la consigne que **toute question doit avoir sa place dans le plan d'exploitation.** Certes c'est un critère pertinent pour juger de l'opportunité d'introduire une question sensible dans un questionnaire (et c'est dans ce sens que la CNIL se réfère à un « principe de finalité »). Néanmoins **une application trop rigoureuse pourrait se révéler pernicieuse**. On doit pouvoir introduire dans un questionnement des questions qui n'ont d'autre utilité que de servir de «liant », qui servent à contenter l'enquêté -qui a ainsi l'impression d'avoir pu expliquer son cas, faire valoir son opinion. Dans l'enquête Actifs, les questions d'opinion sur les qualités trouvées aux divers produits avaient certes une utilité en elles-même, mais leur mérite principal était de briser l'impression pesante que créerait une enquête entièrement centrée sur la mesure des montants. Certaines questions ouvertes n'ont d'autre but que de servir d'exutoire à l'enquêté qui ne doit pas, après l'entretien, rester sur l'impression que l'enquête ne lui a pas permis de s'exprimer. Enfin, certaines questions n'ont d'intérêt que de permettre la reconstitution d'un concept complexe à partir d'éléments épars (par exemple, construire le chômage BIT à partir de l'absence d'emploi, du fait d'être disponible pour en prendre un, d'avoir fait des recherches pendant la période de référence, ou construire la PCS à partir du métier, du grade, de la position hiérarchique, de la fonction et de l'activité de l'employeur..) : les questions élémentaires n'ont pas de vocation à être publiées en tant que telles. Elles n'en demeurent pas moins indispensables[52].

On pourrait aussi développer les réflexions autour des essais **d'inclusion d'items fictifs** dans une série d'items réels, afin de mesurer les erreurs -de remplissage, de déclaration ou de saisie[53], sur **l'influence de l'ordre des questions**, domaine rarement étudié mais qui peut se révéler, dans certaines

[50] Le cas problématique pourrait survenir dans le cas de pratiques très inégalement réparties au sein de la population -comme certaines pratiques culturelles ou sportives-, avec une grande majorité de personnes pratiquant rarement et une faible minorité d'individus pratiquant très souvent. Si l'on réduit trop la longueur de la période de référence, on se retrouve avec une proportion énorme de questionnaires vides ; avec une période de référence plus longue, on récolte davantage de questionnaires informatifs, mais au prix d'une minorité de questionnaires potentiellement très longs.

[51] Ceci a été fait, par exemple, pour l'enquête sur les déplacements.

[52] Un cas semblable s'observe dans de nombreuses enquêtes touchant au domaine de la santé. Les aspects de santé mentale (dépression…) sont traditionnellement abordés par les chercheurs travaillant sur des échantillons de patients par le biais d'échelles, dont la construction nécessite d'obtenir la réponse à un grand nombre d'items élémentaires, qui, individuellement, n'ont que peu de sens et n'offrent pas de garantie de qualité de collecte dans une enquête en population générale (2 exemples choisis prmi quelques dizaines d'items de même type : au cours des quatre dernières semaines, y at-il eu des moments où vous vous êtes senti débordant d'énergie… ; indiquez l'intensité du plaisir -de aucun plaisir à plaisir extrême et ultime en passant par plaisir léger, moyen et majeur- suscité par le fait, assis de regarder un magnifique coucher de soleil dans une région sauvage du monde). Les échelles obtenues par agrégation seraient plus interprétables et plus robustes. La bonne pratique consiste à ne publier, et même à ne laisser figurer dans le fichier mis à disposition des utilisateurs que les échelles synthétiques, l'accès aux variables élémentaires étant réservé aux seuls concepteurs aux fins de validation méthodologique de la démarche.

[53] Deux exemples de cette pratique : dans une enquête Biens Durables et Ameublement, dans la kyrielle des biens listés dans un document autoadministré, on avait glissé un «essuie-verre électrique ». Dans l'enquête réalisée par l'OFT sur les représenttions en matière de drogue, on avait cherché à savoir ce que les gens pensaient du «MOP ». Dans les deux cas, les enquêtés qui sont tombés dans le piège étaient rares, ce qui est plutôt rassurant !

circonstances, crucial[54], ou sur **l'importance du choix du système de tranches proposé sur les ordres de grandeur obtenus**[55].

2.3 une collecte terrain :

Deux grands registres dans cette rubrique, ce qui concerne la **non réponse globale** (incluant les cas d'impossibles à joindre, d'absents de longue durée et les refus d'emblée) ; et ce qui concerne la **qualité de cette réponse** (incluant la non réponse partielle, la précision et la sincérité des réponses…). Sur le premier point, pour les enquêtes en face à face, on observe une **tendance structurelle à la détérioration** (doublement des impossibles à joindre et des absents de longue durée en 10 ans) surtout dans les grandes villes. Mais la France reste plutôt mieux lotie que d'autres pays comme l'Allemagne -échec du panel européen- ou certains pays de l'Est: les taux de refus à Varsovie, Budapest atteignent ou dépassent 50 % même pour des enquêtes assez simples. En France, ce n'est que pour les enquêtes lourdes à carnet (du style Budget des ménages), que l'on atteint des taux de cet ordre dans les zones les plus urbanisées, en Ile de France particulièrement. Par ailleurs, de plus en plus, les enquêteurs manifestent de la réticence à exercer leur métier dans les quartiers dits « sensibles » ; ceci est un enjeu qualité fort pour les années à venir : assurer la couverture effective de toutes les catégories d'habitat. Les réels incidents sont encore rares, mais il faut envisager des mesures pour éviter une détérioration (accompagnement de l'enquêtrice par du personnel masculin de la DR, recrutement d'enquêteur appartenant à ces quartiers, version allégée, sur papier, du questionnaire pour éviter à l'enquêteur de se déplacer avec un micro-ordinateur susceptible de susciter des convoitises…). Un deuxième souci, pour les zones les plus urbanisées, trouve son origine dans la prolifération des digicodes et autres systèmes de protection, sans aucun doute responsables d'une partie des impossibles à joindre. Pour les enquêtes téléphoniques, on note une augmentation du nombre d'appels nécessaires pour un joindre un ménage, d'où un biais difficile à corriger dans les enquêtes par quotas ou avec fichier d'adresses de remplacement, biais en faveur des personnes les plus faciles à joindre. Pour comprendre le mécanisme, il faut s'arrêter un instant sur le fonctionnement des centres automatiques d'appels téléphoniques qui gèrent les enquêtes en CATI. C'est une machine qui compose les numéros, et qui bascule l'appel vers l'enquêteur au moment où le contact est établi. Les numéros occupés ou qui sonnent dans le vide sont remis ultérieurement dans le circuit, avec une gestion des heures d'appel, de façon à ce qu'un même numéro soit appelé à des heures différentes, sur plusieurs jours. En attendant le réappel, d'autres numéros sont essayés. Ces enquêtes étant réalisées par quotas, quand une strate est complète, on cesse de réappeler les numéros jusqu'alors infructueux. Si, pour une strate, on dispose d'un réservoir de numéros potentiels trop grand, avec un nombre de numéros de beaucoup supérieur au nombre d'enquêtes à réaliser, on aura en fin de compte une forte surreprésentation des personnes les plus faciles à joindre. Quand le Ministère de la Jeunesse et des Sports a commandité récemment une enquête sur la pratique sportive, les concepteurs étaient conscients du danger et avaient dans le contrat mis une clause sur l'obligation d'avoir un ensemble de répondants présentant une proportion équilibrée d'hommes et de femmes. Au premier tiers de la collecte, un contrôle a donné une forte surreprésentation des femmes ; au second tiers même constatation. A la fin, le contrat avait été respecté et la bonne proportion d'hommes assurée. Hélas, quand on a croisé les variables âge et sexe, on s'est aperçu que la répartition était totalement distordue par rapport à la réalité : toutes les tranches d'âge moyennes étaient fortement féminisées alors que les jeunes étaient tous ou presque masculins. Ce qui est arrivé était clair : au début de la collecte, les femmes au foyer -d'âge mur- ont été contactées avec succès dès le premier essai d'appel. Quand, en fin de collecte il a fallu respecter le quota d'hommes, on est allé les chercher là où les tranches d'âge n'étaient pas saturées, c'est à dire chez les jeunes. Pour rendre représentatif un tel fichier, il faut des pondérations très dispersées, d'où une moindre qualité (sans compter que le redressement, dans un tel cas, a peu de chance de restaurer des variables cachées comme le goût pour les activités physiques, de par la corrélation entre cette variable et le fait d'être souvent hors du logement et donc difficile à joindre au téléphone). Une seule

[54] En témoigne le célèbre exemple du journaliste américain en URSS et du journaliste soviétique aux Etats-Unis.

[55] On dispose d'exemples quantifiés montrant que l'estimation du temps moyen passé devant la télévision est très différente selon que l'on demande une estimation en clair, selon que l'on propose un système de tranches très détaillé vers le bas de la distribution des durées ou un système de tranches, au contraire, détaillé vers le haut (Froment 1994).

solution, ne pas fournir dès le départ un volume trop important d'adresses et de numéros de réserve, pour forcer la société à s'acharner sur les premiers numéros fournis jusqu'à obtenir une liaison : on entend dire de plus en plus qu'il faut fréquemment de l'ordre de 17/18 appels pour joindre un individu. En cours d'enquête, les problèmes rencontrés imputables à l'enquêteur sont surtout dus à des raccourcis (l'enquêteur ne lit pas exhaustivement toute la question, tous les libellés de réponse[56]…), à des **reformulations abusives**, faussant le sens de la question ou conduisant à **« suggérer » la réponse**[57] (cf supra) ; les erreurs de compréhension sont plus rares (ex : les comptes à terme confondus avec les compte-chèques), la formation étant en général soignée. Le non respect des filtres était également un défaut fréquent avant la capisation[58] ; désormais ceci ne peut plus se produire et c'est sans doute un des avantages concrets majeurs de Capi. Restent quand même des « effets enquêteurs », surtout dans la collecte et la retranscription des libellés «en clair ». Il est vain d'espérer faire une analyse approfondie de ces libellés en absence d'enregistrement de la réponse ; lors de l'opération sur les comportements relatifs au risque, des grappes de «non concerné » ou de « je suis fidèle » se retrouvaient, souvent exclusivement parmi les réponses fournies par un seul enquêteur. Du côté de l'enquêté, le principal problème est lié au fait que le répondant n'est pas forcément **l'informateur optimal** -d'où de nombreux cas de « proxy » ayant à répondre sur des domaines qu'il connaît mal (en particulier les opinions d'autrui)-. Notons à ce sujet que l'informateur optimal peut ne pas être l'individu concerné lui-même : la femme est ainsi meilleure informatrice sur les dépenses d'habillement de son mari que celui-ci ; dans le domaine de la santé, il se peut que l'entourage soit plus au courant de l'état réel de santé du malade que le malade lui-même. Mais il y a aussi des **problèmes de « franchise »,** souci de se montrer sous un jour favorable, ou volonté de (trop) bien faire. Un exemple de ce type de **biais « bien intentionné »** a été observé lors d'une enquête Emploi du temps ; l'enquêté avait indiqué qu'il avait changé la journée qu'il devait décrire, car il la jugeait non « représentative » (il s'agissait du jour du mariage de sa fille, événement sinon unique du moins suffisamment rare pour n'être pas emblématique du quotidien), détruisant une représentativité « macro » au nom d'une représentativité « micro ». Dans ce registre aussi quelques points importants sont contestés. Ainsi, **est-on plus sincère en face à face ou au téléphone, avec un enquêteur que l'on connaît[59] ou un parfait inconnu** ? La question se pose de façon particulièrement cruciale lorsque l'enquête traite de sujets sensibles, pratiques frauduleuses voire illégales, ou simplement touchant des domaines intimes comme les pratiques sexuelles. Il existe des méthodes pour obtenir les taux réels au sein d'une population ; elles sont basées sur l'exploitation des propriétés statistiques des mélanges de distributions : l'enquêté se voit présenter un jeu de cartes ; il tire une carte sans la montrer à l'enquêteur : si c'est une carte noire, il répond à la question sensible (par exemple, «avez-vous déjà pris une drogue dure »)., si c'est une carte rouge, il répond à une question anodine dont on connaît (et c'est là le point clef de la méthode) la répartition dans la population initiale (par exemple : « possédez-vous un lave-vaisselle ») ; à aucun moment l'enquêteur ne sait à quelle question correspond le « oui » qu'il a enregistré. La confidentialité est totale ; l'analyse des résultats permet d'obtenir le bon taux de pratique dans la population. Jusqu'à présent cette méthode, qui crée un surcroît de variance en échange d'un gain en justesse, n'a pas été pratiquée en France à l'échelle d'une grande enquête statistique, par crainte que l'enquêté soupçonnant un « coup fourré » ne joue pas le jeu. Comme les enquêtes n'abordent pas le domaine des pratiques réellement condamnables, on se contente de précautions simples (formulation euphémisée, possibilité pour l'enquêté de répondre par un numéro d'item sans

[56] On peut agir à ce niveau en précisant clairement le protocole qui doit être suivi, distinguant bien les questions fermées pour lesquelles toutes les modalités doivent être lues avant de demander à l'enquêté de choisir la ou les modalités qui conviennent à son cas, les questions fermées pour lesquelles on peut s'arrêter de lire les modalités dès que l'enquêté s'est reconnu dans l'une d'elles, les questions semi-ouvertes pour lesquelles on ne lit aucune modalité, où l'on demande une réponse spontanée en clair, mais pour laquelle l'enquêteur dispose d'une liste de modalités qui lui servent à effectuer un codage à chaud, et les vraies questions ouvertes où l'on note sans la coder la réponse spontanée. Souvent le questionnaire ne précise pas la démarche à suivre. Plus de rigueur à ce niveau faciliterait le travail de l'enquêteur et le contrôle du respect du processus.

[57] Lors d'un accompagnement, la conceptrice d'une enquête Emploi du temps avait constaté qu'un enquêteur réputé « bon » posait systématiquement la question sur le lavage de la vaisselle, quand il se trouvait face à un homme, sous la forme « en ce qui concerne la vaisselle, je pense que c'est madame qui la fait ? ». Et ensuite, on commente la pérennité des stéréotypes !

[58] avec parfois des résultats surprenants : à l'exploitation de l'enquête Actifs financiers on a observé un ménage qui avait répondu à la fois à la question sur les raisons de possession d'un produit et sur les raisons de non-possession du même produit.

[59] par exemple quelqu'un du même village, du même immeuble

qu'il soit forcé d'énoncer la réponse…) en comptant sur **l'anonymat de la relation enquêteur-enquêté** pour obtenir une déclaration sincère[60] ; c'est dans ce contexte que se situe la discussion sur les performances relatives du téléphone et du face à face. Usuellement, on prétend que les gens sont plus à l'aise -et donc plus sincères-pour parler de sujets délicats au téléphone ; c'est oublier un peu vite que le téléphone peut aussi être le lieu d'expression privilégié des hâbleurs de tout poil (cf. différence déjà citée entre les nombres de partenaires décrits par les hommes et les femmes). Ceci permet d'émettre un certain doute sur la capacité du téléphone à faire surgir la vérité de la bouche des enquêtés. Autre question controversée : quelle est **l'influence de la durée d'entretien sur la qualité**, sachant qu'elle est la résultante de deux phénomènes agissant en sens inverse, l'apprentissage et la lassitude[61]. Il est arrivé que l'enquêteur ait eu à recommencer une enquête, l'enquêté ayant dit au bout d'un moment qu'il avait jusque là répondu n'importe quoi, mais qu'il était désormais convaincu du sérieux de l'enquête et était prêt à répondre scupuleusement[62]. A l'inverse, l'enquêté à qui l'on assène une longue liste de pratiques a vite fait de repérer que s'il répond «non », on lui pose moins de questions complémentaires que s'il répond «oui » (par exemple, en cas de réponse positive, on lui demande une fréquence, un montant, alors qu'on ne demande rien en cas de réponse négative) : il peut alors être tenté au bout d'un certain nombre d'items de répondre systématiquement non pour abréger. Le concepteur a à sa disposition diverses stratégies : il peut recourir à un ordre aléatoire des éléments[63] (il répartit alors la sous-estimation sur tous les éléments) ; il peut aussi commencer par un tableau synoptique des pratiques, où seule la question de l'existence est posée. L'effet redouté n'existe pas, puisque c'est seulement une fois collecté l'ensemble des réponses que l'on complète la description des pratiques déclarées : on a un peu « piégé » l'enquêté, mais on diminue les biais de déclaration. L'usage des **périodes de référence** bornant le laps de temps sur lequel on relève les pratiques est aussi sujet à des difficultés spécifiques avec des effets qui peuvent aller en sens contraire. Ainsi, **face à un carnet** de consommation[64], **certains ont des stratégies d'évitement** -on consomme moins pendant la période de relevé- pour gagner du temps, s'épargner de la peine de remplissage, alors que d'autres **modifient leur calendrier de consommation** pour faire pendant la quinzaine de relevés la totalité de leurs achats mensuels, pour éviter d'avoir à montrer la modestie de leur niveau habituel de consommation. De

[60] Dans les enquêtes abordant les sujets d'argent -revenus, patrimoine- on a plutôt tendance à penser qu'il vaut mieux un enquêteur inconnu que l'on ne reverra jamais. Dans un Comité du Label récent, lors de l'analyse de l'enquête VESPA destinée à étudier les personnes infectées par le virus VIH, la question s'est posée de savoir si on déclarait plus facilement des pratiques sexuelles à risque (nombre de partenaires différents par semaine, rapports non protégés, sodomie avec de nombreux partenaires...) à son médecin, quelqu'un du corps médical, ou à un enquêteur. L'équipe conceptrice a cité des études conclusives montrant que pour l'usage de la drogue, on déclarait plus sincèrement la consommation à l'enquêteur qu'à son médecin : des analyses ont été faites sur d'anciens drogués qui affirmaient à leur médecin qu'ils suivaient ses recommandations et ne prenaient plus rien alors qu'ils disaient le contraire à un enquêteur ; c'est cette dernière version des faits que l'analyse a confirmée.

[61] La difficulté d'avoir une vision claire sur ce point provient en partie d'une collusion objective d'intérêts qui ne pousse pas à relater de façon neutre les expériences ni à chercher à approfondir la question. Pour les enquêteurs, une enquête courte est souvent aussi une enquête facile, ne nécessitant pas un investissement en formation trop lourd ; c'est aussi une enquête pour laquelle la gestion des rendez-vous est plus simple, qu'il est plus facile à conduire dans les interstices d'une autre activité…pour le concepteur c'est une enquête moins chère à l'unité qui lui permet d'avoir à budget constant un plus gros échantillon. Tout le monde a un certain intérêt à vouloir que l'entretien aille vite ; mais c'est la qualité qui pâtit ; bousculer l'enquêté n'est pas la meilleure façon d'obtenir une réponse valable. Dans l'opération de réinterrogation des chômeurs ne faisant pas de démarches, évoquée ailleurs dans ce texte, une enquêtée a avoué avoir acquiescé à tout ce que disait l'enquêtrice car celle ci semblait très pressée et qu'elle n'a pas voulu lui faire perdre son temps en réfléchissant ! Une autre preuve du peu de fondement des discours d'enquêteurs relatifs à la fatigue des enquêtés au-delà de 45 minutes se manifeste lorsque l'on prévoit une deuxième visite pour les entretiens longs ; la plupart du temps l'enquêteur enchaîne les deux visites sous prétexte que l'enquêté le souhaitait, avait le temps…et alors, par miracle, il n'est plus question ni de fatigue ni de baisse d'attention. Ceci ne veut pas dire que l'on peut impunément rallonger les durées, mais seulement qu'il faudrait investir davantage pour connaître précisément les véritables limites, éventuellement en fonction du type d'enquêté.

[62] Lors du test du questionnaire 2ème phase de l'enquête Structure des salaires (un recto-verso postal adressé au salarié), un enquêté s'est étonné de la brièveté du questionnement, se demandant ce que l'Insee pourrait faire d'un ensemble aussi mince d'informations. Il semble qu'il y ait davantage d'unanimité quant aux problèmes posés, pour l'enquêteur aussi bien que pour l'enquêté, par les dépassements importants par rapport à la durée annoncée, d'où la nécessité d'avertir honnêtement les enquêtés sur la durée probable, éventuellement en leur indiquant le risque de variabilité forte en fonction de leur situation (i.e. dans l'enquête Patrimoine, on annonce une durée plus longue en cas de patrimoine diversifié).

[63] Mais certains critiquent cette méthode, car elle complique l'analyse : il faudrait introduire la place de l'item dans le questionnement pour l'individu concerné pour mesurer correctement l'effet des diverses caractéristiques sociodémographiques sur la pratique.

[64] Mais aussi face à un questionnement rétrospectif avec période de référence

même, volontairement ou inconsciemment, certains ont **tendance à rapatrier dans la période de référence des consommations de peu antérieures** parce qu'elles sont importantes pour eux et qu'ils ont envie d'en faire état, que ce soit dans les carnets ou lors des entretiens rétrospectifs[65]. Ainsi, on observe systématiquement dans les enquêtes sur les biens durables plus de voitures de moins d'un an que de voitures mises en service sur cette période. Quel est l'effet qui l'emporte ? A nouveau, tout semble pouvoir arriver selon les personnes, les sujets…et l'on manque d'expériences suffisamment précises et variées pour conclure.

2.4 La saisie et le chiffrement :

La mise en place de la collecte assistée par ordinateur (CAPI, CATI, CAWI) ou de la lecture optique (recensement) et la suppression consécutive des ateliers de saisie a complètement modifié la façon dont se présentaient les problèmes de qualité de la saisie -la saisie de masse (avec son taux de 1 caractère sur 1000 erroné) ne porte plus que sur quelques documents papier autorenseignés et est remplacée par la saisie directe par l'enquêteur -qui n'est pas forcément plus à l'abri des fautes de frappe[66]-. Les considérations autour de l'efficacité en termes de qualité de la double saisie ne sont plus vraiment centrales. Apparaissent par contre les réflexions autour des **« contrôles embarqués »,** qui tendent à intégrer dès la conception du questionnaire les contrôles correspondants, et autour de la **codification automatique.** Sur ces deux points, les avis sont contrastés. Les contrôles (par exemple contrôles longitudinaux dans le cas des panels) qui éradiquent les fausses transitions ne lissent-ils pas trop les données ? la qualité « Sicore »[67] avec son arbitrage entre taux de codage et justesse du codage (Deschamps, Destandau, Dumontier 2000 ; Bulot 2002), est-elle indubitablement meilleure que celle du codage manuel ? n'aurait-on pas moins de flou mais plus de risque de biais ?

Le fonctionnement du logiciel de codification automatique a été optimisé pour gagner du temps dans la reconnaissance d'une masse très nombreuse de libellés ; il s'efforce de reconnaître les mots en n'en lisant qu'une partie, en l'occurrence quelques bigrammes sélectionnés en fonction de leur pouvoir informatif (en français, pour les libellés de profession, c'est d'abord le second bigramme, puis le premier) ; contrairement à ce qui se passe en codage manuel, la surabondance de précisions peut s'avérer nuisible. Sans entrer dans le détail, il faut retenir que la qualité du processus dépend de la façon dont les libellés sont reportés, mais qu'il n'est pas simple, quand on ne baigne pas dans la mécanique Sicore -et donc pour l'enquêté comme pour l'enquêteur-, de savoir reconnaître le « bon » libellé. Quelques exemples récents montrent les erreurs qui peuvent se produire (heureusement en général détectées lors des phases de contrôle et donc corrigées, les fichiers d'apprentissage utilisés par Sicore faisant l'objet d'un suivi continu et étant régulièrement mis à jour) : pour le codage de la profession, on a trouvé « joBS d'été[68] » codé en « suBStitut de la République », « inGEnieur coMMercial » confondu avec « saGE feMMe », « taXI » avec « taXIdermiste ». Pour le codage des libellés d'activité d'emploi du temps « je console ma fille » s'est retrouvé en « jeux vidéo »…En matière de produits de consommation, « Elle et Vire » s'est retrouvé en magazine (féminin)….sans parler des lunettes et autres lentilles dont on ne sait s'il faut les classer en alimentaire ou en optique.

Mais il semble que **dans les deux cas, les effets positifs l'emportent** sur les quelques inconvénients (et ce même sans introduire dans la balance les importants gains sur le volume de travail réalisés)[69].

[65] Dans l'enquête Habillement, où l'on recensait les achats sur une période de trois mois, tous ces phénomènes ont été observés.

[66] Une preuve en est fournie par la découverte dans l'enquête Emploi en continu, de 80 000 militaires du contingent en 2002, la seule explication possible étant dans l'existence de fautes de frappe.

[67] Sicore -Système Informatique de COdage de Réponse aux Enquêtes » est le logiciel de codification automatique actuellement utilisé à l'Insee.

[68] En majuscule, les bigrammes pris en compte par SICORE.

[69] On a évoqué en première partie la difficulté à tirer du signal du bruit qui l'entoure et à discerner un vrai mouvement d'entrée (sortie) de pauvreté à partir des variations de revenus telles qu'elles sont observées dans un panel sans « interview dépendant », sans contrôle longitudinal lors de la collecte. Les experts internationaux -les anglais en particulier pour leur panel, mais aussi les américains avec le PSID- insistent sur l'importance décisive qu'ont eu les contrôles de ce type sur l'interprétabilité des évolutions. Un dernier exemple, dans le cas français, peut être cité ; il concerne l'enquête Emploi, et les

Restent que les contrôles doivent être spécifiés par des experts conscients des effets qu'ils peuvent introduire et qu'il faut s'efforcer de garder trace du processus correctif mis en œuvre, en particulier sauvegarder la réponse originelle[70].*

2.5 Des imputations et redressements :

Deux phénomènes dans cette rubrique, la **correction de la non-réponse totale** et celle de la **non-réponse partielle**. En ce qui concerne le premier registre, les méthodes font apparaître un arbitrage entre dispersion des poids (d'où risque de non stabilité des résultats) et précision du calage. Le choix des variables de calage et du niveau de finesse peut s'avérer sensible (lors de l'audit susmentionné sur la mesure du taux de possession de chiens, on avait observé qu'un calage sur la catégorie sociale donnait des résultats différant de 1% selon que l'on utilisait une version du code agrégée -8 positions- ou détaillée -12 positions-). Rappelons aussi que l'importance revêtue par ce point dépend fortement des modes d'exploitation envisagés : alors que pour l'établissement de statistiques descriptives simples, cette pondération peut jouer un rôle sensible, pour les économètres elle ne s'impose pas dès lors que le conditionnement adéquat a été introduit dans le modèle[71]. L'opportunité d'un calage sur des statistiques externes portant sur les variables d'intérêt elles-mêmes et non plus sur les seuls cofacteurs peut aussi être discutée, même si usuellement la réponse qui l'emporte est plutôt négative.

Dans le second registre, les principales questions opposent les méthodes d'**imputation déterministes** ou **aléatoires** (effondrement du Gini en cas d'imputation déterministe…). Les autres différences, comme celles qui séparent les méthodes par «Hot-Deck» (ou plus sophistiqué par plus proche donneur), des approches par régression[72] sont davantage des nuances de spécification que des oppositions de fond. Ces méthodes utilisées pour combler des lacunes, sont aussi utiles dans des cas de nature voisine, en particulier la transformation des variables en tranches en variables continues par les résidus simulés (ou généralisés) et le traitement des arrondis (salaires, âge des biens durables, durée du travail…) par méthodes de lissage ou non-paramétriques[73].

2.6 Une exploitation :

Dernière étape avant l'obtention des résultats, l'exploitation introduit un nouveau niveau de qualité :du choix de l'unité statistique (individu ou ménage) au choix des méthodes statistiques mises en œuvre, l'éventail des éléments pouvant agir sur la qualité finale est large (mais traiter de ce point dépasse le cadre limité de ce travail).

---

calendriers rétrospectifs d'activité qui recélaient, faute de contrôle, de nombreuses incohérences génératrices de fausses transitions. Dans ce cas, l'exploitation a pu séparer le signal du bruit, mais ce n'est pas toujours le cas et il arrive que le bruit rende une question inexploitable.

[70] Parmi les exemples de contrôles abusifs, deux contrôles qui avaient été introduits pour éviter les erreurs de frappe lors de la saisie des informations démographiques dans le Tronc commun : le premier stipulait que les conjoints devaient être de sexes différents et avoir des âges dont la différence n'excédait pas vingt ans. Pour éviter une situation où les erreurs généraient de faux couples homosexuels, on a rendu impossible la description de la réalité, ce qui a d'ailleurs valu au Directeur Général de l'Insee de recevoir un courrier critique de SOS Homophobie, qui a accéléré la suppression du contrôle incriminé.

[71] Ceci du moins lorsque la sélection est exogène. Les problèmes de sélection endogène des répondants sont à la fois plus graves dans leurs conséquences et plus difficiles à pallier(cf. infra).

[72] Les méthodes économétriques peuvent s'appliquer d'une façon déterministe ou aléatoire selon que l'on ajoute ou non un résidu à l'estimation obtenue ; le «Hot Deck» est fondamentalement de nature aléatoire. Ce qui diffère entre les deux méthodes, c'est le caractère plus ou moins additif de la spécification.

[73] Sans entrer dans les détails, mentionnons que des travaux d'imputation de ce type ont été réalisés sur la dernière enquête sur les budgets des ménages ; les principales sources de revenu ont fait l'objet d'imputations économétriques, à la fois pour compléter en cas de données manquantes et pour transformer les valeurs déclarées en tranches en valeurs en clair. Les données de consommation ont été complétées soit par imputations économétriques soit par méthode du plus proche donneur (Chopin, Massé, Rouquette 2002).

3. Quelques voies pour améliorer le processus actuel

Les réflexions précédentes ne doivent pas être prises comme un réquisitoire contre les pratiques actuelles de mise en place des enquêtes nouvelles ; le souci de qualité est présent chez la plupart des concepteurs. Les tests prévus sont toujours réalisés, et, à l'Insee, ils sont plutôt en plus grand nombre que dans les autres instituts de statistique européens. Elles ont pour but d'alerter contre la subtilité de la plupart de ces défauts qui guettent à chaque étape du processus et réussissent encore parfois (souvent ?) à passer au travers des mailles du filet tendu pour les repérer[74]. Afin d'être plus efficace dans nos techniques de repérage et d'éradication des défauts, on peut **recommander un certain nombre d'aménagements aux protocoles actuels.**

3.1 Les tests

On remarque souvent que manquent, en début de processus, de véritables **« tests en bureau »** : il est frappant de voir à quel point on aurait pu éviter de recourir au terrain par une simple réflexion préalable. Quand, face à une remarque d'expert, le concepteur acquiesce en soulignant « les enquêteurs ont fait remonter, après un passage terrain, le même problème », loin d'être satisfait de cette convergence d'opinion, il faut plutôt la considérer comme preuve de l'inutilité du recours au terrain : dans le meilleur des cas, on n'a fait que découvrir ce que l'on est capable de concevoir en bureau ; au pire, on a peut-être ainsi manqué une opportunité de discerner un problème plus subtil, l'arbre, en quelque sorte, cachant la forêt. Mais c'est peut-être le prix à payer pour que les concepteurs (un peu Saint Thomas) soient convaincus : on a souvent observé que les experts (y compris les concepteurs des éditions antérieures de la même enquête) n'arrivaient pas à convaincre de la pertinence de leur avis (souvent à cause d'une croyance que c'était peut-être ainsi avant, mais que maintenant les choses avaient évolué et que la difficulté ne pouvait qu'avoir disparu) alors que les enquêteurs étaient crus sur parole ! Cette « naïveté » face à ce que disent les enquêteurs, qui, même lorsqu'ils n'ont pas d'intérêt stratégique à mettre en lumière tel ou tel aspect, n'ont qu'une vision partielle de la réalité et une capacité limitée à juger du degré auquel l'objectif est atteint, faute d'avoir toujours assimilé en profondeur les tenants et aboutissants de l'opération, diminuerait sans doute si les **concepteurs accompagnaient plus systématiquement les enquêteurs sur le terrain**[75]. C'est un blocage dont il ne faut pas sousestimer l'importance ; néanmoins il ne faut pas désespérer et **commencer par des tests en bureau** (l'enquête internationale Share, sur le vieillissement, prévoit des « mock tests ») est un conseil à appliquer systématiquement[76]. Soulignons que le tout premier test est d'essayer de répondre honnêtement soi-même à l'enquête : si on n'y arrive pas , qu'est-ce que ce sera pour des gens moins familiers avec le questionnaire ! Par rapport aux tests actuels, qui se limitent à administrer la questionnaire lui-même, on devrait systématiquement rajouter des compléments, par exemple des questions, qui ne seraient pas forcément reprises dans le questionnaire final, mais qui seraient conçues pour cerner les ambiguïtés de concept, ou l'ampleur des approximations faites dans une réponse rapide par rapport à ce que donnerait un questionnement plus approfondi[77]. Dans ce registre, le test apporterait vraiment un éclairage spécifique : le recours au terrain, seul, peut permettre de mesurer l'ampleur réelle des ambiguïtés que le raisonnement discerne[78]. On peut aussi recommander une exploitation statistique de ces tests, éventuellement -mais ce n'est pas toujours

[74] Ce qui a d'autant plus de chance d'arriver que le concepteur a un ego moins tourné vers l'autocritique.

[75] Une consigne que l'on devrait rendre obligatoire pour chaque concepteur, même si ce n'est pas toujours simple à organiser.

[76] Les membres de la famille, les collègues de bureau sont des cobayes tout désignés pour une première étape. On peut organiser ensuite des réunions de présentation des enquêteurs pour recueillir leur avis, mais sans qu'ils aillent sur le terrain.

[77] L'exemple cité à propos de l'enquête Pratiques culturelles et sportives pourrait servir de modèle.

[78] Ce problème s'insert dans la question plus vaste de la validation scientifique des opérations d'enquêtes. Dans sa réflexion sur MDA, le Comité des investissements avait mis en lumière une moindre validation de l'aspect scientifique des projets par rapport à la validation de l'ingénierie. Des solutions sont à l'étude ; mais d'ores et déjà le principe d'une consultation précoce (en tout état de cause avant la « blaisification » des questionnaires) dans le cadre du Comité du Label est acté. Reste à en déterminer les modalités pratiques.

indispensable - en en gonflant l'échantillon. Des «tests d'exploitation », sur gros échantillon[79] ont l'intérêt de permettre une meilleure estimation des taux de réponse ainsi que des exploitations statistiques : il s'agirait de les généraliser pour toute enquête nouvelle pour lesquelles les exploitations d'éditions antérieures ne peuvent être utilisées pour vérifier la qualité[80].

3.2 Les appariements entre sources

Les appariements individuels de sources pourraient à la fois améliorer la précision et alléger les temps d'entretien : ainsi un appariement systématique avec les données fiscales permettrait de concentrer la collecte sur les composantes non imposables du revenu (les prestations sociales), ou sur les cas mal observés par le fisc (travailleurs frontaliers, foyers très en deça de la limite dimposition…). Un appariement avec les données de la CNAM pourrait alléger l'enquête Santé de tout ce qui concerne l'observation des médicaments achetés (partie, de plus, de qualité médiocre dans une enquête). Reste à monter un dossier convaincant à l'adresse de la CNIL, toujours vigilante face à des projets d'appariements, dossier prouvant que, même pour le ménage, les avantages l'emportent sur les inconvénients.

3.3 Les opérations méthodologiques spéciales

Une mesure précise de la qualité nécessiterait de construire des opérations méthodologiques spéciales[81], chacune adaptée à la mesure des conséquences d'un effet particulier, soupçonné de grever la qualité. Elles seraient conçues spécialement comme des « plans d'expérience » destinés à contrôler les facteurs autres que le facteur d'intérêt. Coûteuses en temps et en argent, elles sont rarement mises en œuvre en France ; un objectif est d'associer à chaque enquête INSEE une opération méthodologique particulière, par exemple une réinterrogation des non-répondants (Enquête Santé), ou des **entretiens semi-directifs associés** (enquête Histoire de vie, Patrimoine) pour n'en citer que deux types.

Les réinterrogations par entretiens semidirectifs déjà évoquées en première partie indiquent quelques voies possibles : on peut faire expliciter les concepts (dans l'opération complémentaire à l'enquête Emploi, on a fait parler autour des concepts d'emploi, de démarches afin de comprendre pourquoi dans l'enquête certains chômeurs disent rechercher du travail mais déclarent ne faire aucune démarche) ou, ce qui est proche, travailler sur les contenus indigènes des rubriques des nomenclatures (dans le complément à Pratiques culturelles et sportives, outre les exemples déjà cités, on a fait réfléchir aussi sur ce que sont des magazines littéraires, des romans historiques, des sports de combat ou des arts martiaux[82]) ; on peut aussi utiliser la mise en cohérence subtile des diverses parties pour faire ressurgir des événements omis plus ou moins volontairement (ce qui a été fait dans le complément à l'enquête Situations défavorisées, afin de creuser le problème du nombre des plages d'emploi de courte durée ou de la multiactivité) ; on peut essayer de mesurer les effets mémoire en se donnant le temps et les moyens de faire le travail de maïeutique nécessaire à la production d'un souvenir précis et complet (les petites périodes de chômage ou de petits boulots, les activités

[79] Quelques centaines, voire un millier de questionnaires peuvent être nécessaires, pour les opérations nouvelles complexes. C'est ce qui a été fait récemment pour Histoire de Vie.

[80] Encore faut-il que les délais de réaction soient brefs, en particulier que les équipes informatiques livrent rapidement un fichier exploitable.

[81] Ce que l'on vient d'évoquer à propos des tests en est une forme, modeste quant aux effectifs observés, mais similaire quant à la forme.

[82] Dans le cadre de la préparation de l'enquête Modes de Vie, afin de voir si le concept de «production domestique » recouvrait un concept indigène, on avait donné aux ménages des cartes représentant des activités et demandé de les regrouper, puis d'expliquer les regroupements faits : aucun regroupement ne recoupait les nomenclatures «savantes » des experts travaillant sur les emplois du temps (loisir, travail domestique.. ;) ; on observait plutôt des tas du type « ce que je fais souvent, rarement, jamais » ou «ce que j'aime faire, ce qui m'indiffère et ce que je déteste. ». Cette expérience prouvait clairement que les catégories générales ne sont pas des outils utilisés par les gens et qu'elles n'ont donc aucune raison d'être définies dans leur tête

secondaires, le nombre de livres lus,). Pour Histoire de vie, on prévoit de faire parler les gens autour de la discrimination, pour voir si la technique d'entretien fait apparaître plus ou moins de situations discriminantes ; pour Patrimoine, on va essayer de faire préciser comment les enquêtés font l'évaluation de leurs actifs patrimoniaux afin de comprendre l'origine des fortes sous-estimations observées partout et de tous temps : dissimulation volontaire, oubli, valorisation au prix d'achat et non à la valeur actuelle ; prise en compte particulièrement prudente des plus (ou moins-)values non réalisées.

Même si on ne peut multiplier les opérations qualité, un objectif d'une opération par enquête serait un objectif prioritaire.

Un danger guette cependant, que l'on pourrait rapprocher d'une variante du principe d'Eisenberg (perturbation du phénomène par la mesure). L'enquêteur se comporterait différemment lorsqu'il y a une opération qualité (ou lorsqu'il est accompagné) : c'est en partie vrai, mais il faut relativiser ; il est difficile d'éviter que le «naturel, chassé, ne revienne au galop » tout au long d'une enquête d'une heure ! Néanmoins, il est possible que les enquêteurs associent enquêtes méthodologiques et contrôle de leur travail et donc aient des arrières-pensées plus ou moins stratégiques aboutissant en fin de compte à faire échouer l'opération. Lors de l'enquête Patrimoine 1997, on avait conçu un recto-verso postal complémentaire ; or il s'est avéré ex post qu'il avait souvent été présenté de façon désinvolte à l'enquêté (d'où une baisse des taux de réenvoi) parce que certains enquêteurs soupçonnaient le test d'être un ballon d'essai destiné à vérifier dans quelle mesure on pourrait remplacer les enquêtes en face à face par des enquêtes postales moins chères, avec tous les risques générés au niveau de l'emploi et du revenu des enquêteurs. Une opération spécifique menée pour tester l'ampleur de l'effet induit par l'obligation sur le taux de réponse semble avoir aussi été l'objet de problèmes de cet ordre, les enquêteurs -ayant intérêt à maintenir l'obligation- se seraient montré particulièrement peu insistants avec les ménages faisant partie de l'échantillon pour lequel l'obligation avait été levée, ce qui pourrait expliquer que les effets mesurés alors aient été plus massifs que ceux que l'on a pu mesurer par la suite, lors d'enquêtes qui n'avaient pas obtenu le caractère d'obligation.

Quant à la réinterrogation des non répondants, elle semble moins riche, dans la mesure où la non réponse est souvent due à des facteurs conjoncturels (je n'ai pas le temps à cette période de l'année, je suis malade, j'ai des problèmes familiaux) peu (ou pas du tout) liés au sujet de l'enquête. Le déchet (refus..) crée de la variance, puisqu'il diminue le nombre d'observations, mais pas de biais (du moins pas de biais impossible à redresser par calage sur des variables observables[83]). L'opération méthodologique n'est en fait là que pour s'en assurer. On la fera donc principalement lorsque l'on anticipe un lien possible entre les causes de refus, ou d'absence et le sujet de l'enquête : Santé, Transports et Emploi du Temps sont trois cas où ce type de problème est potentiellement important, dans la mesure où l'on aurait du mal à toucher les malades, les grands voyageurs et les personnes « surbookées ». Jusqu'à présent, les rares fois où l'on a fait une telle opération, on a pu conclure à l'absence de biais et donc à la possibilité de corriger les non réponses par simple calage -sur la pyramide des âges en particulier-[84]. Reste que toutes ces opérations sur la réinterrogation de non répondants ne sont que partiellement concluantes, car il reste toujours un noyau incompressible de non-répondants irréductibles, dont on peut toujours postuler qu'ils ont un comportement spécifique, même s'il n'en est rien pour la frange des non répondants initiaux ayant accepté la réinterrogation. C'est pourquoi rien ne vaut une action destinée à contenir les non réponses à un niveau le plus faible possible et ce grâce à un effort de suivi constant de la collecte.

[83] La non-réponse n'est pas homogène par âge, type de ménage, type d'habitat. Néanmoins ceci peut se corriger par repondération, pourvu que ces critères soient introduits, comme le veut la théorie, dans le procès de redressement. Ce qui est plus grave c'est le refus lié au thème de l'enquête, car si les malades, par exemple, répondent moins à une enquête Santé, le redressement va, en quelque sorte remplacer des vieux malades par des vieux en bonne santé, au grand dommage de la pertinence des résultats.

[84] Deux opérations ont été menées à ma connaissance, l'une sur l'enquête Santé 1980, l'autre sur Structure des salaires 1992, mais c'était plutôt, dans ce dernier cas, sur de la non réponse de la part des entreprises.

Le suivi de la qualité pour la phase de collecte passe par un certain nombre d'actions (**relance des non-répondants, contrôle a posteriori auprès du ménage** qui peuvent faire l'objet d'un **PAQ** (Plan Assurance Qualité) allant même jusqu'à se traduire par l'obtention d'une norme ISO.

Conclusion

Cas particulier : de IALS à IVQ ou « comment faire mieux la prochaine fois » ?

Ce bilan peut paraître négatif : à quoi pourraient bien servir des enquêtes ménages entachées de tant de défauts ? Et ces principes théoriques peuvent-ils réellement améliorer la situation actuelle ou ne peut-on, compte tenu de la nature des enquêtés et des enquêteurs, infléchir notablement le processus de production des enquêtes ? Certes le spectre d'une opération inexploitable n'est pas une vaine menace, comme l'expérience IALS l'a prouvé. Néanmoins, il est rare que l'on en arrive à une telle extrémité : le progrès technique, le cumul des expériences, l'instauration du Comité du Label et celle du Comité des Investissements ont eu des résultats positifs et, généralement on réussit à s'apercevoir à temps des diverses erreurs liées à la conception. La qualité moyenne des opérations soumises a tendance à augmenter, même si des exceptions demeurent. Plus précisément, dans le passé, on a pu constater un certain nombre de fois une amélioration notable de la qualité. Outre les cas déjà évoqués dans le cours du texte, on peut rappeler les effets bénéfiques de la « capisation » d'enquêtes : dans la partie de l'enquête Patrimoine consacrée à la description du cursus professionnel, on avait observé, lors de la première édition, de nombreux problèmes de raccord entre les périodes décrites; après capisation, les incohérences ont disparu. Le traitement des libellés d'activités par SICORE a permis une codification mieux maîtrisée et surtout beaucoup moins coûteuse que la codification manuelle passée. Tout le soin apporté à la conception de l'enquête Histoires de Vie a permis la mise au point d'un questionnement sur un sujet novateur, a priori difficile à traiter par une approche statistique à partir d'un questionnaire fermé, qui a convaincu même les détracteurs de la première heure. Le bilan de fin de collecte réalisé sur l'enquête Coûts de la main d'œuvre a permis d'améliorer l'ergonomie de l'édition suivante. Les travaux méthodologiques réalisés lors de la préparation de la future enquête sur la Structure des salaires (Ardilly, Koubi 2002) ont permis, avec une perte de précision négligeable, de réduire fortement (d'environ 30%) le coût de l'opération dans sa première phase (auprès des employeurs) en répartissant autrement les échantillons d'établissements et de salariés à interroger -légère augmentation du nombre d'établissements et baisse de 300 000 à 200 000 du nombre de salariés- et d'utiliser les gains réalisés pour étendre la seconde phase (interrogation directe des salariés) à l'ensemble de l'échantillon alors qu'auparavant cette opération n'avait pu être conduite que sur une petite partie de la population concernée.

Dans le futur proche, une nouvelle opération devrait apporter la confirmation du rôle positif des travaux méthodologiques. En effet, le Ministère de l'Education et l'INSEE rouvrent le dossier IALS en se lançant dans l'enquête IVQ, enquête conçue sur des principes un peu similaires (on cherche à apprécier, par une enquête auprès des ménages, les performances en littératie et numératie de la population adulte, à partir de documents liés à la vie quotidienne). Le pari « qualité » repose sur un protocole décrit minutieusement, faisant usage de CAPI, avec une mesure précise des temps de réflexion et une grille d'observation où, exercice par exercice, l'enquêteur doit noter l'attitude du sujet (signes extérieurs de stress, d'énervement, réponse au hasard…) afin d'estimer économétriquement à la fois un niveau de compétence et un comportement d'implication, de motivation : réussir nécessite d'être compétent et motivé, avec une place laissée au hasard (on peut avoir juste en choisissant la modalité de réponse au hasard, surtout quand l'exercice a la forme d'un QCM). Afin de s'adapter rapidement au niveau de la personne, l'enquête commence par un module d'orientation utilisant un support convivial et non scolaire (un programme de télévision) ; les personnes en grande difficulté se voient très rapidement proposer des exercices simples (compréhension écrite et orale, production écrite…) conçus pour diagnostiquer la nature et l'origine des difficultés. Les autres se voient dirigés vers un module haut, 1/3 de l'échantillon se voyant proposer des exercices tirés de l'enquête IALS, les 2/3 des exercices entièrement nouveaux. On espère ainsi à la fois permettre un lien avec IALS et tester

la robustesse des conclusions au choix précis des exercices. Un module biographique renseigne sur toutes les étapes de la vie du répondant, surtout pour ce qui a trait à la formation ; il est complété par un module « débrouille » pour les personnes en difficulté (comment compensent-elles leur handicap) et par un module « pratique de lecture (au foyer et au travail) » pour les autres. Les enseignements des tests sont globalement positifs : on trouve un taux de personnes en grande difficulté plus proche de 10% que de 40%, ce qui nous rapproche des résultats des tests effectués lors des Journées d'Appel de préparation à la Défense -JAPD- ; les enquêtés ne refusent pas massivement le protocole et sont plutôt impliqués, même si une minorité manifeste de l'énervement à partir du deuxième texte (15%) et si quelques-uns se trompent par étourderie. Néanmoins les constatations faites sur IALS se retrouvent ; ce ne sont pas les plus diplômés qui réussissent le mieux (erreurs liées à un mode de lecture trop souvent « en diagonale »..) et il suffit d'un rien dans la formulation des questions pour que l'enquêté oublie la consigne de chercher les éléments de réponse dans le texte exclusivement et réponde en fonction de sa connaissance du sujet[85]. Le processus de mise au point des textes a mis en lumière l'extrême difficulté à concevoir des énoncés dénués d'ambiguïté pour lesquels on peut coder sans problème ce qui est juste et ce qui est faux[86]. Seule ombre au tableau, il semble encore subsister quelques effets «enquêteur ». L'opération en vraie grandeur est actuellement terminée sur le terrain. La publication des résultats est prévue pour juin…en espérant que l'on ne trouve pas à nouveau 40 % d'illétrés en France[87].

[85] Un des textes donnait des tableaux statistiques sur les accidents de la route, à partir desquels il s'agissait de dire quel était le type de voie le plus dangereux, de valider ou d'invalider l'affirmation selon laquelle plus il y avait de trafic, plus il y avait de victimes. On demandait, pour détailler la réponse à cette dernière question «pourquoi », dans l'espoir d'obtenir le processus utilisé pour tirer l'information des tableaux. La quasi-totalité des répondants a produit des réponses du type « parce que ce sont les jeunes (resp. les vieux) qui conduisent mal » « parce que je le sais.. ; », « parce que je travaille à la DDE et donc je le sais. »

[86] Un exercice a été tiré d'un texte issu d'un guide touristique -en l'occurrence le guide du routard-. La mise au point en a été difficile : en particulier il n'était pas précisé si une église était ouverte entre 12 et 14 heures. On avait considéré que, faute d'indication ; elle était évidemment ouverte. Or, dans les tests, on a vu que les répondants se comportaient différemment selon que dans leur région les églises étaient fermées ou non au moment de la pause méridienne. Une question portait sur ce que l'on pouvait visiter le mercredi -i.e. un jour où aucune fermeture spéciale n'était mentionnée ; une enquêtée a été perturbée car elle avait interprété la question comme « que peut-on visiter spécialement le mercredi » et que ne voyant rien à ce sujet dans le texte, elle se déclarait dans l'incapacité de répondre.

[87] Depuis la tenue des JMS, les choses ont évolué : le Séminaire Recherche prévu a bien eu lieu et les premières exploitations confirment la réussite globale de l'opération, même si une erreur dans la programmation informatique Capi a causé la perte de certains résultats lorsque l'enquêté avait interrompu l'enquête avant terme. Le pourcentage de personnes de niveau 1, incapables de comprendre un message fût-il simple, selon les hypothèses de correction des données manquantes, varie de 1 ou 2 % autour de 15 % (on est loin des 40 %) et même certains défauts constatés au test semblent avoir disparu : si les plus diplômés ne font pas toujours le sans faute qu'ils seraient capables de faire, globalement, les performances moyennes ne cessent jamais de croître avec le niveau de diplôme.

Références bibliographiques :

P. Ardilly et M. Koubi (2002) : « Utilisation d'information auxiliaire et optimisation d'échantillon : le cas de l'enquête sur la structure des salaires » Journées de Méthodologie Statistique 16 et 17 décembre 2002

L. Arrondel, F. Guillaumat-Tailliet et D. Verger (1996) : « Montants du patrimoine et des actifs :qualité et représentativité des déclarations des ménages » Economie et Statistique

F. Battagliola ,I. Berteaux-Wiame, M. Ferrand et F. Imbert (1994) : « Dire sa vie : entre travail et famille » in Trajectoires sociales et inégalités, coordonné par F. Bouchayer Erès INSEE

A. Blum, F. Guérin-Pace (2000) : Des lettres et des chiffres Fayard

F. Bulot (2002) : « La codification de la profession à l'Insee : historique et dernières avancées »,Journées de Méthodologie Statistique 16 et 17 décembre 2002

C. Berthier, N. Caron, et B. Néros (1998) : « Le kish, les problèmes de réalisation du tirage et de son extrapolation » Document de travail INSEE, série Méthodologie Statistique N°9810

G. Canceill :

N. Chopin, E. Massé et C. Rouquette (2002) : « Imputations dans l'enquête Budget de familles », Journées de Méthodologie Statistique 16 et 17 décembre 2002

F. Deschamps, S. Destandau et F. Dumontier (2000) : « le codage automatique d'un carnet de dépense est-il plus complexe que celui d'un carnet d'activité ? » Actes des Journées de Méthodologie statistique 4 et 5 décembre 2000

J.C. Deville et Y. Tillé (2000) « Echantillonage équilibré par la méthode du Cube, variance et estimation de variance », Actes des Journées de Méthodologie statistique 4 et 5 décembre 2000

M. Froment (1994) : « L'erreur de mesure : approche cognitive des interactions entre le questionnaire, l'enquêté et l'enquêteur » Mémoire de DEA (Sciences de gestion) Université Paris-Val de Marne Paris XII

O. Godechot (2000) : « Plus d'amis, plus proches ? Essai de comparaison de deux enquêtes peu comparables » Document de travail INSEE, série Méthodologie Statistique N°0004

M. Grumbach (1982) : « Idéal de l'individu : individu statistique, individu social. Remarques sur l'individuation des pratiques et des opinions dans la méthodologie du questionnaire » Actes de la Journée d'étude Sociologie et Statistique Insee-Société française de sociologie 1982

F. Guglielmetti (2002) : «Autoclassement versus classement objectivé : petit exercice sur la robustesse d'une classification socio-professionnelle», Journées de Méthodologie Statistique 16 et 17 décembre 2002

S. Lollivier et D. Verger (2002) : «Erreurs de mesure et entrées-sorties de pauvreté », Séminaire Recherche 13 juin 2002

R. Platek, F.K. Pierre-Pierre et P. Stevens (1985) «Elaboration et conception des questionnaires d'enquête » Statistique Canada, Division des méthodes de recensement et d'enquêtes -ménages

J.M. Robin (1999) : « Modèles structurels et variables explicatives endogènes » Document de travail INSEE, série Méthodologie Statistique N°2002

T. Scott Murray, I. Kirsch, L. B Jenkins (1998) Adult Literacy in OECD Countries (technical report on the first International Adult Literacy Survey) National Center for Education Statistics