LA COLLECTE PAR INTERNET EST-ELLE L'AVENIR DES ENQUÊTES GÉNÉRATION DU CÉREQ ?

Christophe Barret, Christophe Dzikowski[1]

Céreq, Département Entrées et Evolutions dans la vie active

Résumé

Le dispositif Génération a pour objectif l'étude des parcours d'insertion professionnelle des jeunes débutants. À cet effet, un panel de jeunes sortants du système éducatif est interrogé par téléphone à 3, 5 et 7 ans après leur entrée sur le marché du travail. Une expérimentation de collecte par internet (Cawi : Computer-Assisted Web Interviewing) a été menée en parallèle de l'enquête 2013 auprès de la Génération 2010, sur un échantillon disjoint de celui de la collecte principale par téléphone (Cati : Computer-AssistedTelephoneInterviewing).

Le questionnaire Cawi est, à quelques modifications marginales près, identique au questionnaire Cati. Les abandons en cours de collecte plus importants pour la collecte internet sont vraisemblablement liés au manque d'ergonomie du questionnaire web, à sa longueur mais aussi de manière plus marquée à certaines questions difficiles, telles que celles comportant des consignes longues.

Une comparaison des structures des réponses Cawi et Cati sur un champ comparable nous permet de conclure à une égalité statistique sur 40% des variables du questionnaire. En revanche, 30% des variables ne présentent aucune égalité statistique sur aucune des modalités. Les différences observées sont vraisemblablement liées à un effet de sélection et/ou à un effet de mesure.

Pour essayer de distinguer ces deux effets, une méthode de matching,habituellement utilisée dans l'évaluation de dispositifs publics,a été mobilisée. Ces méthodes permettent d'estimer les effets d'un dispositif, dans le cas d'expérimentations non préalablement contrôlées, en comparant les mesures faites sur les individus traités et leurs contrefactuels non traitésayant des caractéristiques similaires avant l'accès au dispositif. Dans le cas présent, le dispositif étudié est la passation par internet. Les écarts constatés entre le groupe traité (Cawi) et celui de leur contrefactuel (Cati) sont interprétés ici comme de l'effet de mesure. Une réflexion sur les stratégies d'agrégation des données Cati et Cawi en présence de biais de mesure est présentée. Enfin, pour améliorer la distinction des effets de sélection et de mesure, le protocole de la prochaine expérimentation multimode détaille notamment la mise en place d'un échantillon embarqué avec affectation aléatoire du mode de réponse.

Abstract

The Generation survey studies young people's transitions from school to work. For this a cohort is interviewed by phone (Cati: Computer-Assisted Telephone Interviewing) 3 years, 5 years and 7 years after exiting the education system. An experiment of collection through internet has been carried out in 2013 along with the survey of the 2010 generation, over a separate sample.

The Cawi questionnaire is roughly comparable with the Cati questionnaire. Drop-outs during the survey, more frequent for the internet survey, appear to be linked to the web questionnaire design, to its length but also to some difficult questions, namely those including long indications.

A comparison of the responses structure from both surveys concludes to a statistical equality for 40% of the variables. Nevertheless, 30% of the variables don't show any equality on any of the modalities. Observed differences are apparently linked to a selection and/or measurement effect.

In order to distinguish those two effects a matching method has been used. Those matching methods allow estimating the effects of a program (policy measure), for non-controlled experiments, with the comparison of "treated" and "non-treated" groups. In the present case, the program under study is the internet survey. Observed differences between the treated group (Cawi) and the counter factual (Cati) are used as a proxy for the measurement effect. Finally, in order to improve the distinction between measurement and selection effects, the protocol from the forthcoming mixed mode survey is presented. In particular, the use of embedded experiment with random assignment of the response mode is described.

Mots-clés

Collecte, multimode, sélection, mesure, matching

[1] barret@cereq.fr ; dzikowski@cereq.fr. Nous remercions également Pascale Rouaud pour sa relecture attentive.

Introduction

A l'initiative du Céreq, une expérimentation de collecte par internet a été menée en parallèle de l'enquête par téléphone Génération 2010 courant 2013. Les intérêts d'une enquête reposant pour tout ou partie sur une collecte internet seraient nombreux : une diminution des coûts du fait d'une passation en auto-administré pour une partie des enquêtés, un gain de temps du fait que potentiellement de nombreux questionnaires puissent être réalisés dès le début de l'enquête, une augmentation de la couverture en atteignant des répondants aux caractéristiques spécifiques les rendant difficiles à joindre habituellement par téléphone.

Cependant une revue de littérature sur les collectes par internet et sur les enquêtes multimodes informe des nombreuses difficultés méthodologiques qui se posent.

Ainsi la passation d'un questionnaire par internet, du fait de l'absence d'enquêteur prenant en charge la complexité du questionnaire en expliquant ou reformulant les énoncés, peut donner lieu à des réponses de moins bonne qualité. D'une part les risques d'abandon et d'incompréhension sont plus importants. D'autre part, dans ces conditions, l'enquêté peut avoir une implication moindre et par conséquent avoir une tendance plus importante à expédier le questionnaire en répondant de façon aléatoire. Dans la littérature, ces comportements sont regroupés sous le terme « satisficing ».

Une autre difficulté des interrogations multimodes est que les données recueillies selon des modes de collectes différents peuvent être affectées par des effets de sélection, c'est-à-dire que les personnes interrogées par un mode de collecte le sont parce qu'elles ont une propension particulière à répondre avec ce mode de collecte. Ainsi les populations répondant par des modes distincts peuvent avoir des caractéristiques différentes, ce qui est d'une part recherché mais pose d'autre part des problèmes quant à l'agrégation des données.

A cet effet de sélection, s'ajoute un effet de mesure. Ce dernier est défini comme la différence entre la vraie valeur du répondant que l'on souhaite recueillir et la valeur effectivement déclarée. Ces effets de mesure ne sont pas les mêmes selon le mode de collecte. Un exemple connu est celui du biais de désirabilité sociale. Lorsqu'une question sensible est posée par un enquêteur, l'enquêté peut dissimuler sa vraie réponse et donner une réponse qu'il pense correspondre à une norme sociale. Ce biais serait moins présent dans le cadre d'une enquête auto-administrée comme internet ou l'absence d'une personne tiers permettrait de répondre plus sincèrement. Ce biais ne serait cependant pas totalement absent en auto administré mais d'une ampleur moindre.

Enfin, la conception des questionnaires pose aussi question. Certains concepteurs préconisent d'optimiser chaque questionnaire selon le mode de collecte ce qui diminuerait les erreurs de mesure. Cependant, dans ce cas, il y a un risque que les questions optimisées pour chaque mode soient comprises de manière différente et d'ajouter au contraire des erreurs de mesure. De ce fait, certains préconisent de concevoir un questionnaire unique qui réponde aux attentes des différents modes. Cependant des questions écrites exactement de la même manière ne donnent pas nécessairement lieu à une compréhension et une perception identique selon les modes de collecte. Pour résoudre ces contradictions, on peut être amené à construire les questions et les interfaces pour donner des « stimuli », c'est-à-dire une perception de l'information, qui soient a priori les mêmes. On parle alors de « generalized design ».

Il n'y a pas forcément de consensus pour l'ensemble des points présentés ici. Par exemple selon la population d'intérêt ou la qualité du questionnaire, le phénomène de satisficing qui voudrait que les données par internet soient nécessairement de moins bonne qualité n'est pas forcément vérifié.

Après une courte présentation du Céreq et du dispositif des enquêtes Génération, l'article expose les principales difficultés de collecte lors de la première expérimentation internet réalisée en 2013. Une comparaison entre l'enquête principale (Cati) et l'enquête expérimentale (Cawi) est ensuite présentée en tenant compte de la précision des estimateurs. Pour certaines variables, aucune différence significative n'est observée. Pour les autres variables, les différences significatives sont liées à des effets de sélection et/ou de mesure. Une méthode de matching est mise en œuvre pour tenter de distinguer ces deux effets. La fin de l'article propose des pistes de réflexion pour agréger les données dans un contexte d'enquête multimode en présence d'effet de mesure et propose un nouveau protocole expérimental, multimode cette fois, permettant de mieux discerner les effets de sélection et de mesure.

1. Le Céreq et le dispositif des enquêtes Génération

1.1. Le Centre d'Etudes et de Recherches sur les Qualifications

Le Céreq est un établissement public sous tutelle du Ministère de l'Éducation nationale, de l'Enseignement Supérieur et de la Recherche et du Ministère du Travail, de l'Emploi, de la Formation professionnelle et du Dialogue social. La relation formation-emploi est au cœur de beaucoup d'enjeux de la société française. Le Céreq est un expert majeur dans ce domaine depuis quarante ans. A travers ses études et recherches, le Céreq analyse les conditions de l'acquisition des qualification, par la formation initiale et continue, par l'exercice d'une activité professionnelle ; l'évolution des qualifications liées aux transformations des technologies, de l'organisation du travail et de l'emploi ; les conditions d'accès aux emplois ; les conditions de la mobilité professionnelle et sociale, en fonction de la formation reçue et de la gestion de la main-œuvre par les entreprises. Le Céreq évalue également les formations, les dispositifs et politiques publiques mis en œuvre. Enfin, il est chargé de formuler des avis et propositions en matière de politiques de formation et d'enseignement.

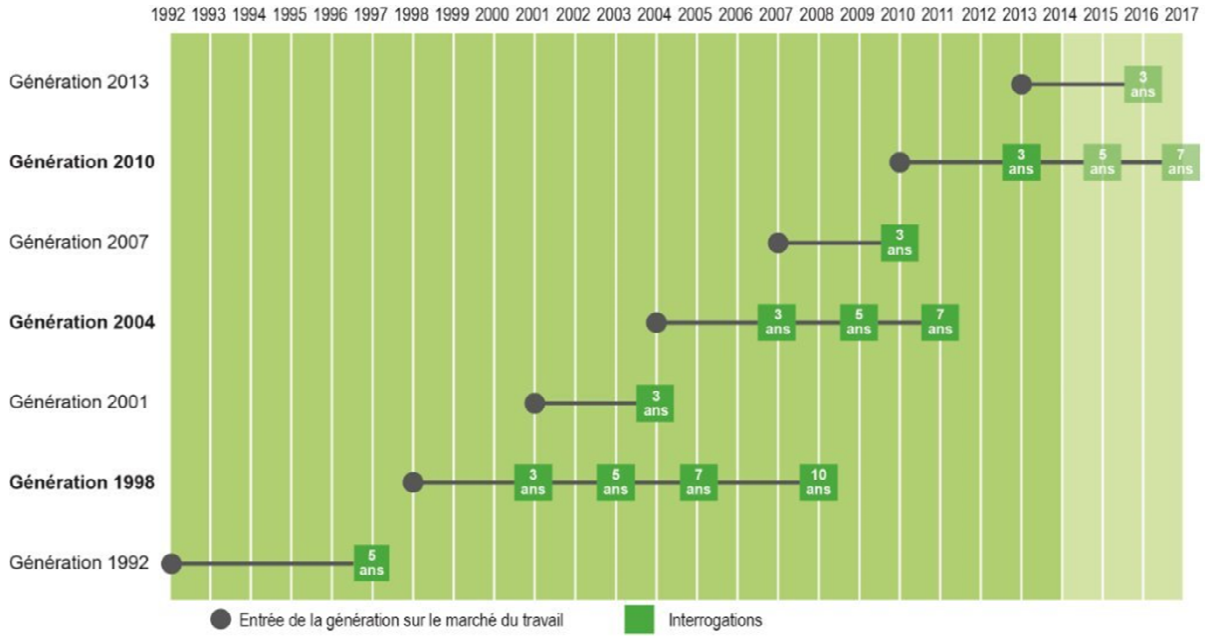
1.2. Un dispositif unique pour étudier les premiers pas dans la vie active, selon la formation initiale suivie

A la fin des années quatre-vingt-dix, le Céreq a mis en place un dispositif d'enquêtes original qui permet d'étudier l'accès à l'emploi des jeunes à l'issue de leur formation initiale. Tous les trois ans, une nouvelle enquête est réalisée auprès de jeunes qui ont en commun d'être sortis du système éducatif la même année quel que soit le niveau ou le domaine de formation atteint, d'où la notion de "Génération". Il s'agit d'une enquête téléphonique (Cati : Computer-AssistedTelephoneInterviewing)

L'enquête permet de reconstituer les parcours des jeunes au cours de leurs trois premières années de vie active et d'analyser ces parcours au regard notamment du parcours scolaire et des diplômes obtenus. Certaines cohortes sont interrogées plusieurs fois pour suivre les débuts de carrière.

En s'appuyant sur un calendrier qui décrit mois par mois la situation des jeunes et sur des informations plus précises concernant le premier emploi et l'emploi occupé au bout de trois années passées sur le marché du travail, ce dispositif permet non seulement d'analyser les trajectoires d'entrée dans la vie active mais aussi de distinguer, d'une génération à l'autre, les aspects structurels et conjoncturels de l'insertion.

Calendrier du dispositif des enquêtes Génération



1.3. L'avenir du dispositif

Le Céreq s'est engagé depuis 2012 dans un vaste projet de rénovation du dispositif des enquêtes Génération. En effet, le financement de ce dispositif au-delà du cycle actuellement engagé (échéance 2017 lors de la troisième interrogation de la Génération 2010) ne pourra plus être assuré en totalité par le Céreq. Il apparaît donc nécessaire de rénover le dispositif afin d'en réduire le coût, de trouver de nouvelles sources de financement et éventuellement de réinterroger la notion même d'insertion professionnelle des jeunes. A cet effet un groupe de travail a été mis en place dont l'objectif est d'aboutir en 2018-2019 à un dispositif rénové techniquement, scientifiquement et budgétairement soutenable.
Dans le cadre de cette réflexion, l'introduction du mode de collecte par internet est l'une des pistes techniques majeures envisagées. Ainsi, le Céreq prévoit de profiter de l'ensemble des enquêtes Génération 2010 ou 2013 (réalisées en 2013, 2015, 2016, 2017) pour mener des expérimentations reposant pour tout ou partie sur une collecte par internet.

1.4. Le protocolede l'enquête monomode internet

En parallèle de l'enquête principale téléphonique (d'avril à juillet 2013) auprès de la Génération 2010, une collecte exclusivement par internet a été mise en place à titre expérimentale (de mi-mai à fin juillet 2013).Ces deux collectes ont été effectuées à partir d'échantillons disjoints.

L'échantillon Cati a été tiré dans *U*qui est la base de sondage des jeunes présumés sortants du système éducatif construite par le Céreq en collectant l'information auprès de l'ensemble des établissements dispensant de la formation initiale. Cette base de sondage a été stratifiée par le croisement de la variable région et d'une variable catégorisant les sortants sur trois niveaux : enseignement secondaire, supérieur court, supérieur long. Sur chaque strate, le tirage est équilibré sur les niveaux de formations avec la macro cube, soit un peu moins d'une vingtaine de niveaux par strate. Au total, 279 000 individus ont été tirés et constituent l'échantillon S_1 de l'enquête principale.
L'échantillon Cawi a été tiré dans le complémentaire de cette même base de sondage, $\{U \backslash S_1\}$. La base de sondage a été stratifiée par le croisement de la variable région et d'une indicatrice portant sur l'appartenance au secondaire. Sur chaque strate l'échantillon est équilibré sur la taille de la strate. L'échantillon Cawi, notéS_2, contient 264 000 individus.

Pour les informer de l'enquête, les individus présents dans l'échantillon Cawi ont reçu une lettre avis électronique uniquement, lorsque l'adresse email était disponible. Les mails envoyés aux enquêtés, contenant le lien vers le questionnaire Cawi, étaient personnalisés avec le nom et prénom de l'individu. Pour les individus de l'échantillon qui ne disposaient pas d'email une recherche d'adresse électronique a été effectuée par un prestataire privé (noté M) .

Outre les mails (mail avis et relances), deux moyens de contact supplémentaires ont été mis à la disposition des enquêtés : une adresse email technique renvoyant vers l'équipe gestion d'enquête du Céreq et enfin l'adresse de la page Facebook dédiée à l'enquête (outil de communication particulièrement adapté à la cible de l'enquête à savoir les jeunes).

Le questionnaire Cawi est quasiment identique au questionnaire Cati. Les aménagements concernent l'introduction de consignes pour certaines questions afin de pallier l'absence d'enquêteur. De légères reformulations de questions ont été nécessaires pour prendre en compte le passage d'un entretien téléphonique à une enquête autoadministrée. Les modalités des questions sont identiquesà l'exceptiondes précisions (écrites entre parenthèses) non citées par l'enquêteur au téléphone qui étaient affichés sur l'enquête Web.
Le Cawi a été développé en faisant apparaitre une question par page. Comme pour le Cati, il est nécessaire de répondre à la question pour passer à la suivante.
Dix relances par email ont été réalisées (soit 11 envois), envoyés à différents moments de la journée, espacés d'une semaine à quinze jours. Les relances et les connexions aux questionnaires envoyés ont été suivies par le prestataire (noté L) qui gérait également l'enquête principale téléphonique. Les emails envoyés par le prestataire M n'ont pas été suivis au niveau individuel mais n'ont donné lieu qu'à des statistiques sur l'ensemble des enquêtés.

Chaque répondant ne pouvant répondre que par un seul mode de collecte, il ne s'agit pas d'une expérimentation de collecte multimode, mais bien de deux enquêtes en parallèle.

2. Bilan de collecte : Une déperdition importante à chaque phase

2.1. Les limites de la base de sondage

La base de sondage des jeunes présumés sortants du système éducatif a été construite par le Céreq en collectant l'information auprès de l'ensemble des établissements dispensant de la formation initiale.

Sur les 264 000 individus de l'échantillon Cawi, 64 000 adresses mails étaient disponibles dans la base de sondage, soit un quart des observations.

La disponibilité des adresses emails était hétérogène selon les types d'établissements de formation. Ainsi, 40% des jeunes sortants des universités, des écoles de commerce ou des écoles d'ingénieur étaient possédaient une adresse mail, alors que les jeunes sortants des rectorats (soit l'essentiel des sortants du secondaire) n'étaient pas couverts par ce mode de contact.

Au-delà de l'indisponibilité du mail, les coordonnées obtenues lors de cette phase datent de la sortie du système éducatif en 2010 pour une enquête réalisée en 2013. Dans certains cas, le mail collecté est donc obsolète.

Sur les 200 000 autres individus de l'échantillon, une tentative d'enrichissement des adresses mails par le prestataire M a permis de retrouver et de « louer » 32 000 mails. Cette opération s'est révélée décevante (seuls 160 questionnaires réalisés).

La base de sondage évolue à chaque vague de collecte. Le taux de remplissage des adresses mails de 25% pour Génération 2010 sur la totalité de la base pourra être amélioré dans le futur grâce à différentes opérations.Les bases d'élèves académiques (BEA) gérées par les rectorats intégreront à terme cette information (les BEA constituent 50% de la base de sondage). Des démarches auprès des académies gérant les bases SIFA (apprentis) sont menées en vue d'uneintégrationdu mail dans leur système d'information. De plus, le Céreqdemande désormais systématiquement cette variable de contact lors de chaque collecte de la base de sondage. Cela encouragera les établissements à récupérer à terme cette information pour mieux nous la communiquer à l'avenir.

2.2. Un envoi de mail délicat

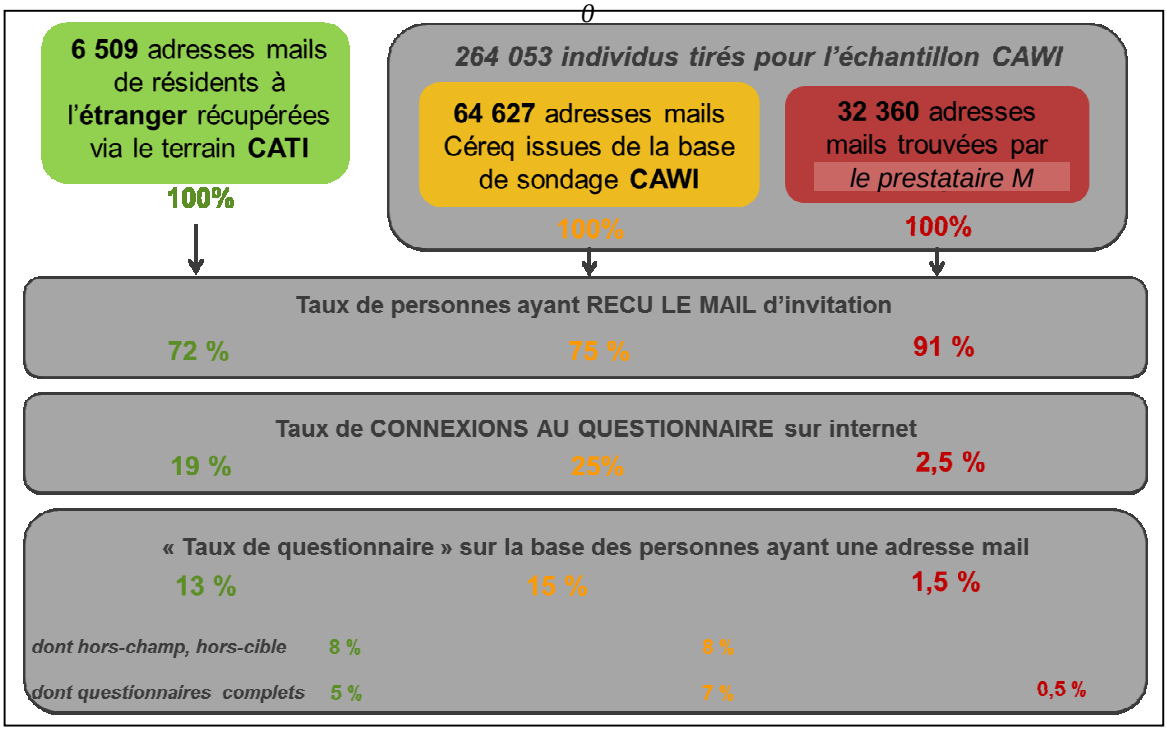
En parallèle de l'échantillon *S2* de l'enquête expérimentale (Cawi), les jeunes de l'échantillon *S1*de l'enquête principale (Cati) repérés comme résidant à l'étranger ont eu la possibilité de répondre (également à titre expérimental) à ce questionnaire Cawi. Ce sont ainsi 6 500 adresses mails qui ont été récupérées auprès de ces jeunes.

Au final, plus de 100 000 individus ont été contactés par mail pour répondre au questionnaire Cawi aboutissant à environ 5 000 questionnaires terminés et dans le champ Céreq (dont 334 questionnaires de résidants à l'étranger et 160 issus de la phase d'enrichissement).

Sur les 64 000 adresses mails disponibles dans la base de sondage, 48 000 mails ont effectivement été reçus, 16 000 connexions sur le lien du questionnaire ont été enregistrées et 12 000 répondants ont terminé la première partie du questionnaire qui permet de déterminer les individus appartenant au champ de l'enquête (questionnaire filtre). Parmi eux, 5 000 individus sont hors champ, plus de 2 500 sont dans le champ mais ont abandonné en cours d'enquête et moins de 4 500 individus sont dans le champ et ont terminé le questionnaire.

Le schéma ci-après présente les ratios de déperdition aux différentes étapes selon l'origine du mail.

Les différentes étapes de la collecte

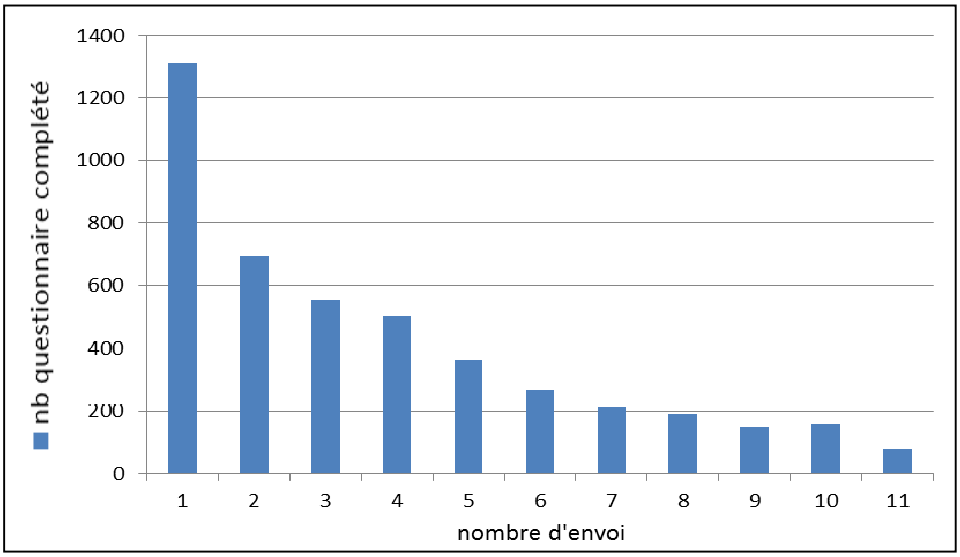


Plusieurs facteurs expliquent ces déperditions aux différentes étapes :

- L'expérience de l'emailing en masse du prestataire L n'était pas suffisante pour éviter tous les blocages.
- Le libellé de l'expéditeur du mail ne paraissait pas suffisamment officielavec les deux prestataires L (enquête) et M (enrichissement).
- La qualité de l'adresse mail retrouvée par le prestataire chargé de l'enrichissement parait discutable (suspicion de récupération de boites « poubelles » non consultées et utilisés par l'individu pour des inscriptions sur les sites commerciaux)
- La qualité du mail-avis était insuffisante. De petits détails ont provoqué des craintes (de phishing notamment) chez certains utilisateurs (retours messagerie) : logo de mauvaise qualité, lien de connexion au questionnaire n'appartenant pas au domaine Céreq.fr.

A l'inverse, les relances jouent positivement sur les taux de connexion et in fine sur le nombre de questionnaires réalisés (voir graphique suivant). En effet les 64000 individus disposant d'une adresse mail ont été relancé jusqu'à 10 fois. On constate que parmi les enquêtés, 3 sur 10 répondent dès le premier envoi, mais il y a également 3 enquêtés sur 10 qui répondent après le 5e envoi. Pour le prestataire M (enrichissement), seuls 3 envois ont été réalisés.

L'effet des relances



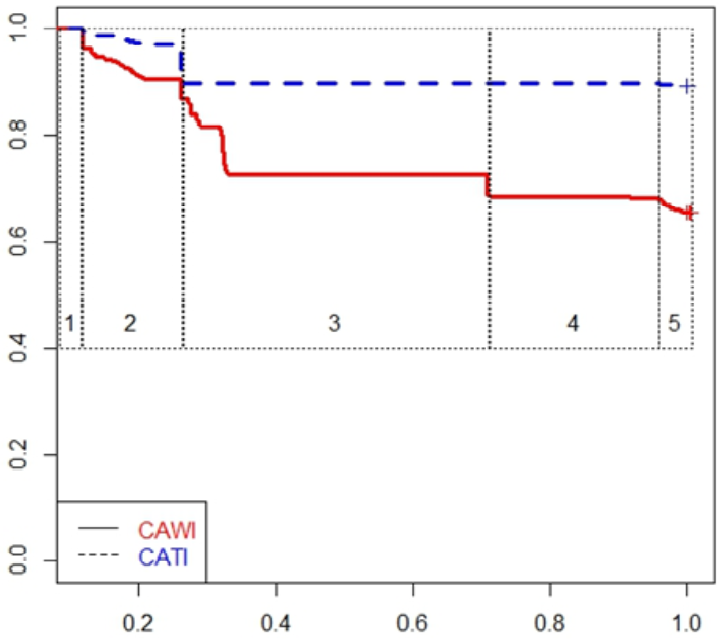
Au final sur les 64000 mails disponibles dans la base de sondage, 15% des jeunes ont accepté de répondre. Ce taux est deux fois moins important que pour l'enquête principale téléphonique.

Au-delà des limites citées de la base de sondage, une autre piste essentielle peut être évoquée pour expliquer un tel écart : les problèmes d'ergonomie du questionnaire Web et la durée du questionnaire (34 minutes en moyenne par téléphone) provoquant de nombreux abandons en cours de collecte.

2.3. Les abandons en cours de collecte

Legraphe suivant représentela courbedesurviedesrépondantsayant été identifiés dans le champen fonctionde l'avancementdans lequestionnaire. Cesabandonsencours depassationneseretrouvent pas avecla mêmeintensitédans les collectespartéléphone et internet. En particulier, letauxd'abandonestquatrefoismoindre par téléphone que par internet. Onretrouve icilerôleessentieldutélé-enquêteur,forméspécifiquement à l'enquête, chargéd'assumerla complexitéduquestionnaire et dereformulerlesquestionsencas d'incompréhension. Onretrouveleconstatsouventétablidansles étudesméthodologiquesd'enquêtés moins impliqués dans les enquêtes auto administrées.

Graphiquedes abandonsen cours de questionnaire



Définition des zones :

1 : filtre pour identifier les sortants dans le champ

2 : module parcours scolaire

3 : saisie des séquences (emploi, chômage, formation…) durant les trois premières années de vie active

4 : description des séquences

Note de lecture:90%desrépondants dans le champ ontrempliplus de 20% duquestionnaire.

Enplusdecettemoindreimplication,deuxtypesd'abandonsdistinctssemblentseproduireaucours dela passationduquestionnaireCawi avec d'une part, unabandondiffusquisetraduitparunedécroissancedelacourbedesurvieavecunefaible pente et d'autre part desabandonsbeaucoupplusmassifspourcertainesquestions.

Lesabandonsdiffuspourraientêtreattribuésen partie aumanque d'ergonomieduquestionnaireCawi. En effet, plusieurschoix de développements rendaient laborieuse la passation par internet :

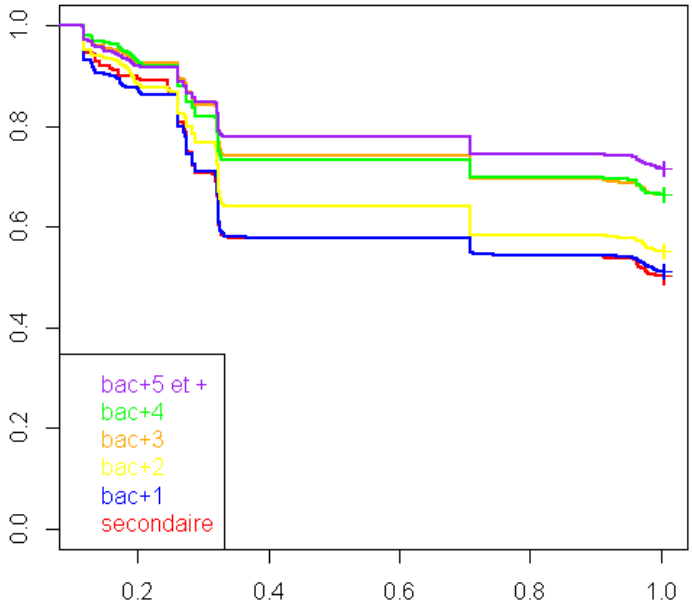
- Unequestionpar page
- Un bouton « valider » peu accessible : utilisation fréquentede la barre de défilement pour l'atteindreen bas de page.
- Le click sur les boutons radios difficile
- Une lenteur dusystème
- Un Cawi peu esthétique
- L'absencede titre de partie compliquant la compréhensiondudéroulement duquestionnaire.

Lesabandonsmassifsont été repérés dans différentes situations :

- Les questions avec consignes longues
- Les menus déroulants (communes, professions, entreprise, …) : absence de recherche intuitive, défilement dans le menu fastidieux
- Après une demande de correction de la saisie (au cours du remplissage du calendrier professionnel)
- A la fin de la description de la première période du calendrier d'activité. Ces abandons concernent exclusivement les sortants dont la première période est inférieure à 4 mois. Il pourrait s'agir des sortants dont le parcours à l'issue du système éducatif comporte un grand nombre de séquences différentes, de courte durée, et qui jugent trop lourde la description de l'intégralité de leur parcours.

Parailleurs, on observeunedifférenciationdesabandonsselon le niveau d'étude des enquêtés.Les répondantsquiabandonnentlemoins par internet ont les niveaux de sortie les plus élevés.Le graphe suivant présente les abandons en cours selon les niveaux de sortie agrégé allant du secondaire à bac+5 et plus. Cela peut s'expliquer en considérant le fait que les plus diplômés ont moins souvent besoin de l'accompagnement de l'enquêteur pour naviguer dans le questionnaire et ils ont également en moyenne des questionnaires plus courts du fait d'un parcours professionnel plus simple à décrire. A l'inverse, pour les bas niveaux de qualification, un questionnaire reposant entièrement sur l'écrit peut poser problème aux jeunes avec une appétence en lecture ou en bureautique potentiellement moins élevée. D'autre part, ce sont les jeunes les plus exposés à un long questionnaire du fait d'un parcours professionnel potentiellement plus complexe. Cette différenciation des abandons selon la classe de sortie est également présente pour la collecte Cati, mais les écarts sont nettement moins marqués.

Courbe d'abandon pour les répondants internet selon le niveau de sortie



3. Le redressement des données Cawi

Du fait de l'indisponibilité fréquente du mail et des abandons en cours liés avec le niveau de formation, les populations Cati et Cawi ne sont pas directement comparables. En effet, en l'absence d'emails pour les sortants de rectorat, plus de 90% des répondants Cawi sont des sortants de l'enseignement supérieur, alors que dans la population générale ils sont un peu plus de 50%. Ainsi, seuls 185 sortants du secondaire ont répondu au Cawi. Par conséquent, les étapes de correction de la non-réponse et de redressement se feront sur les sortants du supérieur uniquement pour travailler sur des populations comparables.

Toujours en raison d'une information incomplète sur les mails, et de types de formation sélectionnés exhaustivement pour la collecte Cati, certains niveaux de formation du supérieur contiennent très peu voire pas de répondant. Par la suite, dans la partie consacrée à la comparaison des réponses des deux modes de collecte, on travaillera sur les niveaux de formation qui sont communs aux collectes Cati et Cawi. Ainsi, les niveaux de formations qui ne contiennent aucun sortant dans la collecte Cawi ne seront pas pris en compte dans les statistiques calculées à partir des données Cati. Cet ensemble sera appelé $S_{restreint}$et constitue une restriction des données Cati.

Par ailleurs, sur cette restriction du Cati, et sur les données du Cawi, sera défini un domaine formé par l'agrégation des niveaux du supérieur qui contiennent suffisamment d'effectifs Cawi. L'idée ici est de comparer des populations susceptibles d'être atteintes. Le domaine formé par l'agrégation de ces strates sera noté S_d, que ce soit pour le Cati ou pour le Cawi.

Dans cette partie nous allons voir les corrections qui ont été apportées aux probabilités d'inclusion du Cawi pour prendre en compte le fait que le tirage ait été fait dans le complémentaire de S_1. Ensuite nous aborderons la modélisation de la non-réponse qui repose dans notre cas sur trois modèles logistiques successifs, enfin nous présenterons les étapes relatives au calage des observations du Cawi sur celles du Cati. Toutes les étapes de redressement sont faites sur l'ensemble des sortants du supérieur appartenant à $S_{restreint}$ et non uniquement sur le domaine S_d.

3.1. Correction des probabilités d'inclusion du Cawi

L'échantillon de l'enquête internet S_2ayant été tiré dans le complémentaire de l'échantillon de l'enquête par téléphoneS_1, les probabilités d'inclusion finales des unités k dans l'échantillon S_2 sont égales à $(1-\pi_k^1)\times\pi_k^{\frac{2}{1}}$et non pas à $\pi_k^{\frac{2}{1}}$. Où π_k^1est la probabilité de tirage dans le premier échantillon, et $\pi_k^{\frac{2}{1}}$ la probabilité de tirage dans le deuxième échantillon conditionnellement au premier tirage. En effet :

$1_{k\in\overline{S_1}})$

Les probabilités de tirage sont conservées par le tirage équilibré, les probabilités d'inclusion pour les observations du Cawi sont donc $(k\in S_2)=\pi_{k_{tirage_{Cawi}}}\left(1-\pi_{k_{tirage_{principal}}}\right)$, ce sont les probabilités d'inclusion qui seront utilisées dans les traitements post-collecte, où $\pi_{k_{tirage_{principal}}}$est la probabilité d'inclusion lors du premier tirage et $\pi_{k_{tirage_{Cawi}}}$les probabilités d'inclusion utilisées pour le tirage de l'échantillon Cawi.

3.2. Une modélisation de la non-réponse reposant sur plusieurs modèles successifs

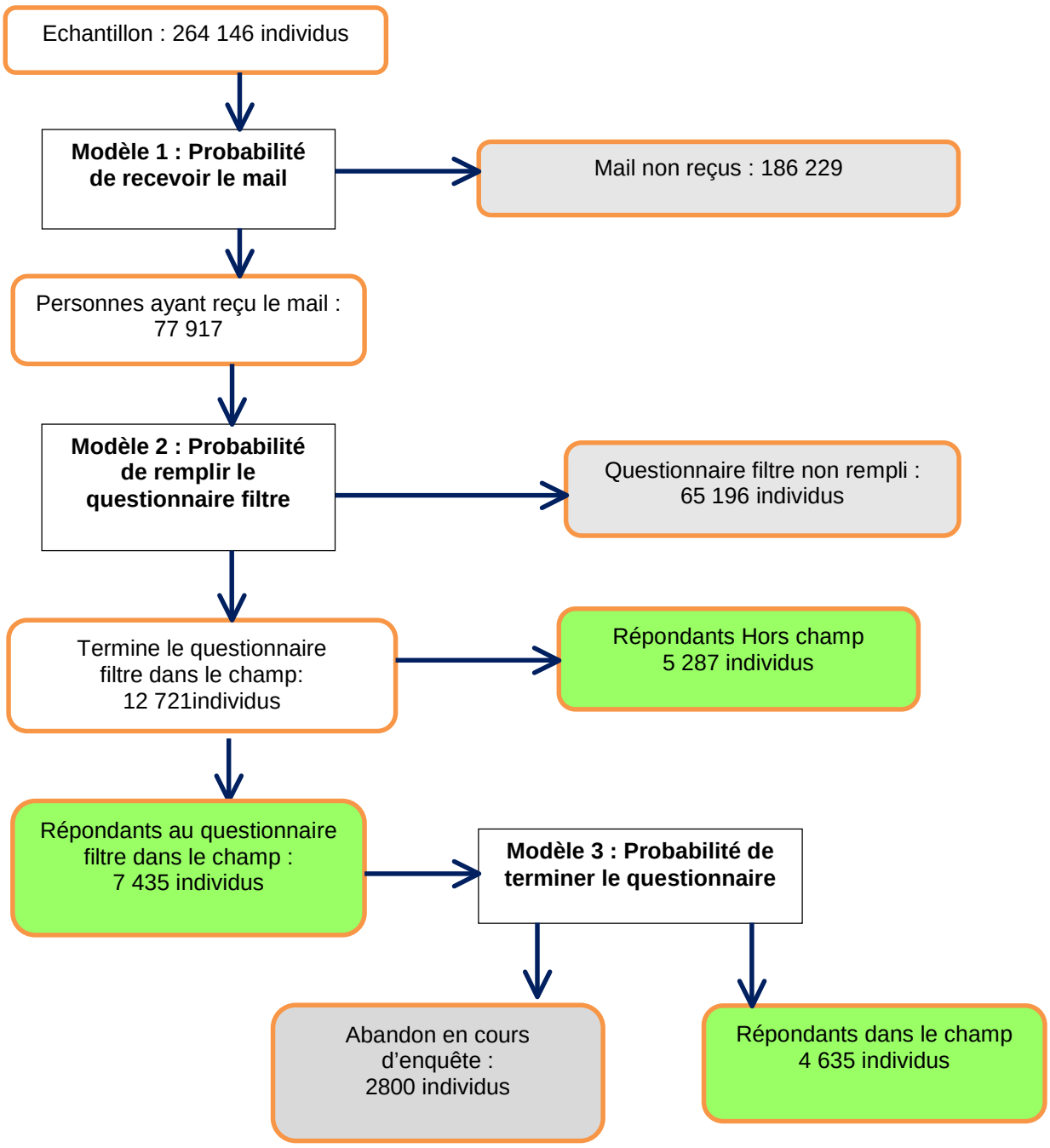
Nous avons vu précédemment que les hors champ Céreq sont nombreux à répondre. Ils sont identifiés dans la partie filtre du questionnaire et exclus au même moment. La modélisation de la non-réponse sera donc constituée de modèles successifs séparant les étapes de la modélisation de la non-réponse portant sur la population entière des étapes concernant uniquement les répondants dans

le champ. Le schéma suivant montre la succession des étapes de la correction de la non-réponse.

Peu de variables de la base de sondage étaient mobilisables pour modéliser la non-réponse. Les variables explicatives sont les suivantes :

- la classe d'âge (F_ANAIS) *(retenu dans le modèle 2(M2))*
- la variable sexe (sexe) *(retenu dans M2)*
- l'indicatrice de résidence en ZUS au moment de la sortie de la formation initiale (ZUS),*(M2)*
- la variable niveau de formation (STRAF, 32 positions), *(M1)*
- la variable niveau de formation agrégé (STRASUPT, 5 positions) *(M2, M3)*
- la région de l'établissement de formation (REG). *(non retenue)*
- le nombre de connexions de l'enquêté (NB_CONNEXION), cette variable est une para-donnée et ne fait pas partie des variables de la base de sondage.*(M3)*

Schéma de modélisation de la non-réponse



A l'issue de cette phase de modélisation de la non-réponse on obtient des probabilités de réponses qui s'échelonnent de $8 \cdot 10^{-3}$ à 0,21 avec une médiane à 0,10. Pour améliorer la robustesse des estimations des probabilités de réponse qui pourraient pâtir d'un modèle mal spécifié, un regroupement des observations a été effectué par la méthode du score. Les observations ont donc été triées selon la probabilité de répondre $\boldsymbol{p_k}$. Les n premières observations servent à former la première classe, les n suivantes la deuxième etc. Pour chaque groupe homogène de réponse la probabilité de réponse des observations est remplacée par la probabilité moyenne dans le groupe homogène de non-réponse. Cette étape est en relation avec deux objectifs qui relèvent du compromis biais variance :

- diminuer le nombre de classe pour assurer une meilleure stabilité des estimations en réduisant les poids extrêmes dus aux probabilités de répondre très faibles.
- augmenter le nombre de classe pour avoir des groupes suffisamment petits pour conserver des classes homogènes, et ainsi avoir un estimateur corrigé de la non-réponse qui s'écarte peu de l'estimateur par expansion $\boldsymbol{y_\pi}$ sans biais.

3.3. Calage des observations du Cawi sur celles du Cati

L'enquête Génération est habituellement calée sur les données de l'enquête Emploi pour satisfaire des besoins de cohérence avec les données publiées par l'Insee (Institut national de la statistique et des études économiques) et la Depp (Direction de l'Evaluation, de la Prospective et de la Performance). Dans le cadre de la comparaison des modes de collectes entre téléphone et internet, dont l'objectif est une comparaison en interne, les marges de calages ne seront pas nécessairement celles utilisées pour le calage des données Cati.

L'intérêt du calage est ici de rapprocher les structures Cati et Cawiet de s'affranchir en partie des effets de sélection qui ont eu lieu au moment de la collecte et donc d'étudier les différences de réponses sur des populations comparables. Cette étape de calage se fait sur les niveaux de formation du supérieursuffisamment représentés dans les collectes Cati et Cawi.

On pourra noter, ici dans le cas des calages sur la variable STRAF des niveaux de formation détaillés, que les poids calés finaux sont indépendants de la distance choisie pour effectuer le calage. En effet, la variable formation est utilisée seule, par conséquent le système que l'algorithme ou la résolution directe doit résoudre est un empilement de N équations simples indépendantes où N est égale au nombre de modalité de la variable STRAF. Les poids calés sont donc strictement égaux à l'issue du calage quelle que soit la distance choisie.

Globalement les rapports de poids qui représentent l'effort que doit faire le calage pour amener la structure initiale sur la structure finale, semblent corrects. La plupart des rapports de poids sont proches de 1 et ils sont compris entre 0,38 et 1,73 ce qui est acceptable selon l'usage.

4. Comparaison des structures de réponses Cawi et Cati

Ce rapprochement sur une même dimension des données Cawi et Cati autorise les comparaisons entre les structures des autres variables du questionnaire. Dans un premier temps nous abordons la comparaison des modes de collecte de manière descriptive par l'intermédiaire d'indicateurs résumant, pour chaque variable, les écarts entre les structures des réponses Cati et Cawi. Ensuite, nous tenterons d'expliquer ces écarts par quelques variables relatives aux questions, comme par exemple le nombre de modalités, ou la taille de la population considérée. Enfin, nous présenterons la méthode utilisée pour estimer les variances des estimations. Celles-ci sont faites par un calcul direct qui prend en compte les étapes d'échantillonnage, de non réponse et de calage.

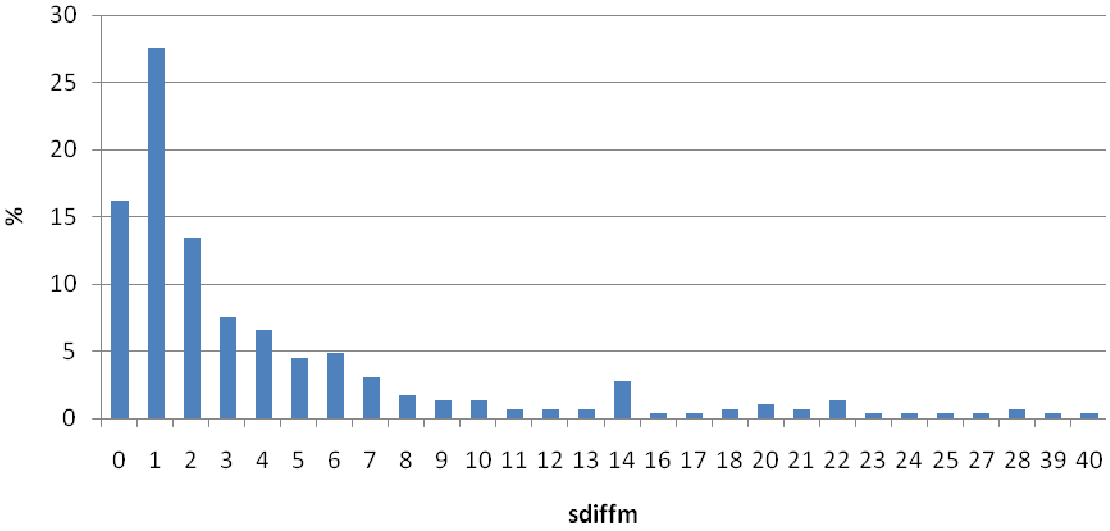
4.1. Description des écarts entre les réponses Cati et Cawi

Un premier tri à plat systématique sur chacune de ces variables montre que pour une part importante des questions ou des variables reconstruites, le faible nombre de répondantsCawi et la restriction au domaine $\boldsymbol{S_d}$ entrainent que certaines variables peuvent n'avoir aucune observation pour une ou plusieurs de leurs modalités. Il s'agit par exemple des questions relatives aux sortants de thèses. Ces derniers qui ont été tirés principalement pour la collecte Cati sont trop peu présents dans l'échantillon

Cawi pour que l'on puisse exploiter les questions les concernant. D'autres variables ne sont pas directement comparables à cause d'un nombre de modalités trop important pour les populations concernées dans la collecte Cawi. Un travail de regroupement des modalités pourra être fait par la suite pour intégrer ces variables dans ce travail de comparaison des modes de collectes.

Après avoir retiré ces variables non directement comparables le travail porte sur un ensemble de 285 variables. Pour chacune des variables a été calculée la somme des écarts en points et en valeur absolue entre les deux structures Cawi et Cati, cette somme est noté *sdiff.* Cette somme permet de représenter la distance entre les deux structures pour une variable donnée. Cette quantité divisée par le nombre de modalités est appelé *sdiffm.* Le graphique suivant montre la distribution des écarts de points moyens par modalité *sdiffm* pour les 285 variables.

Structure des différences entre les collectesCawi et Cati



Not
e de lecture : 27% des variablesont une différence entre les structures Cati et Cawi d'un point en moyenne par modalité

On constate que pour plus de la moitié des variables directement comparables, les écarts moyens par modalité sont relativement faibles, au plus 2 points d'écarts en moyenne par modalité. En revanche, près de 30% des variables présentent des différences de plus de 5 points en moyenne par modalité.

Une tentative d'explication de ces écarts a été faite à partir des informations suivantes : la taille des populations concernées par les questions considérées, le nombre de modalités de ces questions, la statistique du khi-deux entre les modalités des questions et les variables de calage ainsi que la variable caractérisant l'avancement dans le questionnaire.

Les questions posées ici sont les suivantes :

- Est-ce que les écarts augmentent lorsque la taille du champ de la question diminue ? Auquel cas l'aléa de sondage pourrait jouer un rôle important.

- Est-ce qu'une corrélation importante entre les variables de calage et les variables peut diminuer significativement les différences constatées entre les structures ? Auquel cas le calage aurait un réel intérêt sur l'estimation des variables bien corrélées avec les variables de calage.

- Est-ce qu'un nombre de modalités plus élevé peut entrainer une dispersion plus grande des réponses et de manière différente selon le Cawi par l'effet de primauté et le Cati par le recencyeffect ?

- Est-ce que l'avancement dans le questionnaire est lié aux écarts ?

Des modèles de régression ont été mis en œuvre. Pour obtenir des résidus plus proches de la normalité, ce ne sont pas directement les variables *sdiff* ou*diffm* qui ont été expliquées mais leur racine carrée, après avoir testé une transformation par le logarithme. Les modèles ne permettront pas de trancher clairement ces diverses questions.

La corrélation entre la taille de la population concernée par la question considérée et la différence totale *sdiff* est légèrement négative. La corrélation entre le nombre de modalité est la différence *sdiff* est légèrement positive. Ce sont les seules relations entre les variables explicatives et les différences *sdiff* qui sont significatives.

On constate sur les résultats de la régression que même si la relation entre la statistique du khi-deux et les variables sdiff est négative, ce qui va dans le sens espéré d'une diminution de la variabilité lorsque la corrélation avec les variables de calage augmente, cette relation n'est pas significative. L'apport du calage sur la variance des estimations semble assez restreint dans le cas présent. De même, il n'y a pas de lien entre l'avancement dans le questionnaire et les différences constatées *sdiff*.

Les liens sont faibles entre les variables explicatives et les différences sdiff et sdiffm. Même si quelques corrélations significatives sont trouvées, la variance expliquée des écarts entre les structures des deux populations est de l'ordre de quelques pourcent au final. La grande partie des écarts entre les structures Cawi et Cati ne sont pas explicables par des critères simples et suggèrent donc qu'il y a des effets de sélection et des effets de mesure importants qui subsistent. Il est nécessaire de regarder au cas par cas les facteurs potentiels qui peuvent expliquer ces différences entre les deux collectes. Après cette présentation des différences constatées, nous allons passer à l'estimation de la variance des estimations afin de déterminer si les différences observées sont significatives et ainsi mieux interpréter ces écarts.

4.2. Une estimation de la variance par calcul direct

Afin de ne pas envisager des effets de sélection et des effets de mesure sur chaque écart observé, il convient de déterminer si les différences sont significatives. Pour cela nous avons procédé à une estimation de la variance par calcul direct, en composant les formules associées respectivement aux étapes de tirage d'échantillon, à la correction de la non-réponse et à l'étape de calage. Cette partie a été réalisée avec l'appui méthodologique de la division sondage de l'Insee. Comme l'enquête Génération mobilise essentiellement des données qualitatives, ce seront les variances des proportions qui seront calculées à partir des variables linéarisées.
Le choix de l'estimation de la variance par calcul direct a été fait pour des raisons de coût de calcul et des raisons méthodologiques. La base de sondage comporte en effet plus d'un million de lignes et le temps de calcul de la macro cube croit rapidement avec le nombre d'observations. De ce fait, Les tirages d'échantillons équilibrés par strates avaient été effectués en mobilisant plusieurs PC au sein du département. Il aurait été très lourd de procéder à un calcul de précision par réplications sur une population fictive de plus d'un million d'observations, étant donné que le minimum de réplications pour obtenir un écart type est de l'ordre de 30. Nous aurions pu calculer les variances par bootstrap non pas sur une population fictive de la même taille que celle de la base de sondage mais sur une population pondérée. Cependant nous n'avions pas de garantie sur la validité de cette méthode.

4.2.1. Estimation de la variance due à l'échantillonnage équilibré

L'intérêt du sondage équilibré est de diminuer la variance des variables d'intérêts, la variance étant la variance des résidus sur les variables d'équilibrage. De plus, lorsque l'équilibrage est exact, la structure de l'échantillon correspond exactement à la structure du système des variables auxiliaires spécifiées. La variance du sondage équilibré est estimée par l'approximation de variance de Deville Tillé. Etant donné que le sondage a été stratifié par région et selon des catégories de formation, la variance totale sera la somme sur chaque strate des variances. De plus comme la population a été restreinte à l'ensemble $S_{restre\,\mathrm{int}}$ ce seront uniquement des variances calculées sur ces strates restreintes du supérieur qui seront sommées en s'intéressant en particulier au domaine S_d .

Cette formule de variance considère que le tirage équilibré est un tirage poissonnien réalisé sous les contraintes d'équilibrage. Sa variance est celle d'un sondage poissonnien avec des expressions

particulières des paramètres $C_{k,h}$ .L'application de l'approximation de variance de Deville Tillé pour estimer la variance de l'estimateur d'Horvitz-Thomson du total de la variable d'intérêt Y sur la strate h, notée $\hat{V}_{DT}(\hat{Y}_{\pi}^{h})$,est donnée par :

$$\hat{V}_{DT}(\hat{Y}_{\pi}^{h}) = \sum_{k\in S_h} c_{k,h}(y_k - \tilde{y}_{k,h})^2$$

où $\tilde{y}_{k,h} = x'_k(\sum_{l\in S_h} c_l x_l x'_l)^{-1}\sum_{l\in S_h} c_l x_l y_l$ avec $x'_k = (x_{k1},\cdots,x_{kq})$ et $c_{k,h} = \frac{(1-\pi_k)}{\pi_k^2}\frac{n_h}{n_h - q}$.

La quantité $(y_k - \tilde{y}_{k,h})$ correspond en fait au résidu de la régression dans S_h de la variable d'intérêt Y sur les variables de calage X, pondérée par $c_{k,h}$: elle se calcule donc très facilement à l'aide d'une proc reg, et la quantité $\hat{V}_{DT}(\hat{Y}_{\pi}^{h})$ en découle directement.

La variance de l'estimateur d'Horvitz-Thompson du total d'une variable d'intérêt Y liée au seul aléa d'échantillonnage est donc estimée par :

$$\hat{V}_{DT}(\hat{Y}_{\pi}) = \sum_h \hat{V}_{DT}(\hat{Y}_{\pi}^{h}) = \sum_h \sum_{k\in S_h} c_{k,h}(y_k - \tilde{y}_{k,h})^2$$

4.2.2. Estimation de la variance en prenant en compte le processus de non-réponse

L'approximation de la variance avec non réponse, se fait en se basant sur la variance obtenue dans un sondage à deux degrés, où la phase de non réponse est considérée comme le deuxième degré du sondage.

On applique dans ce cadre la formule de Rao pour estimer la variance des unités primaires à partir de l'échantillon des unités secondaires, c'est-à-dire ici les répondants

T_i est le total de la variable Y dans les unités secondaires. La formule de Rao permet d'exprimer la variance du total comme la somme d'un terme quadratique et de la somme des variances des T_i

$$\hat{V}(\hat{T}) = Q(\hat{T}_1,...,\hat{T}_m) + \sum_{i\in S_I}\left(\frac{1}{\pi_i^2} - q_i\right)\hat{V}(\hat{T}_i)$$

La variance du premier terme est due au tirage équilibré, le deuxième terme correspond à la correction à apporter à l'estimation de la variance en prenant en compte l'ajout de variance due à la non-réponse.

Dans notre cas la deuxième phase correspond au processus de non-réponse, où une unité secondaire correspond à un individu. La variable d'intérêt est la variable $T_i = y_i$, le total sur l'unité secondaire i. elle est estimée par $\hat{T}_i = \frac{y_i \mathbb{1}_{i\in R}}{p_i}$, avec p_i la probabilité de répondre de l'individu i et $I_{i\in R}$ l'indicatrice d'appartenance à la population des répondants.
L'estimation de la variance prenant en compte l'aléa de sondage et la non-réponse est donnée par :

$$\hat{V}(\hat{Y}_{\pi,CNR} = \sum_{k\in S}\frac{y_k}{\pi_k p_k}) = \sum_h\sum_{k\in S_h} c_{k,h}(z_k - \tilde{z}_{k,h})^2 + \sum_{k\in S}\left(\frac{1}{\pi_k^2} - q_k\right)(1-p_k)z_k^2$$

$$\text{avec } z_i = \frac{y_i \mathbb{1}_{i\in R}}{p_i}.$$

Le premier terme correspond à la variance d'équilibrage et le deuxième au terme additionnel dû à la prise en compte de la non-réponse.

4.2.3. Estimation de la variance de l'estimateur à l'issue de la phase de calage

Lorsque la taille de l'échantillon augmente l'estimateur calé converge vers l'estimateur par la régression. De ce fait il est de biais négligeable, et sa variance est celle des résidus obtenus par la régression des variables d'intérêts sur les variables de calage. De manière générale si on a un estimateur calé T_W , la variance de ce dernier est donnée par :

$$V(\hat{T}_W) \approx V(\sum_{k \in s} d_k \varepsilon_k)$$

Où $\varepsilon_k = y_k - \sum_{\alpha=1}^{K} \hat{b}_\alpha X_k^\alpha$, est le résidu de y sur les variables explicatives X^α ($\alpha = 1,2,\ldots K$), et les d_k sont les poids des observations.

On va calculer la variance à partir des résidus de la variable d'intérêt sur les variables de calage avec comme pondération les poids avant calage prenant en compte le poids de tirage et le poids lié à la correction de la non-réponse. L'estimation de variance en prenant en compte les trois étapes de tirage d'échantillon, de correction de la non-réponse et de calage est donné par

$$\hat{\mathrm{V}}(\hat{\mathrm{Y}}_{\text{calé}}) = \sum_{\mathrm{h}} \sum_{\mathrm{k} \in \mathrm{S_h}} \mathrm{c_{k,h}} (\zeta_{\mathrm{k}} - \tilde{\zeta}_{\mathrm{k,h}})^2 + \sum_{\mathrm{k} \in \mathrm{S}} \left(\frac{1}{\pi_{\mathrm{k}}^2} - \mathrm{q_k} \right) (1 - \mathrm{p_k}) \zeta_{\mathrm{k}}^2$$

$$\text{Où } \zeta_{\mathrm{i}} = \frac{\varepsilon_{\mathrm{i}} \mathbb{1}_{\mathrm{i} \in \mathrm{R}}}{\mathrm{p_i}}$$

4.2.4. Calcul de la variable linéarisée

Les variables d'intérêts de l'enquête Génération sont essentiellement des variables qualitatives. Les statistiques sont donc des ratios. Cette statistique n'est pas linéaire et l'approximation de la variance a été faite en calculant la linéarisée de la proportion P. Soit la modalité m_v d'une variable v donnée. On va calculer la linéarisée de la proportion des sortants satisfaisant la modalité m_v sur le nombre de sortants concernés par la question v.

$P = \frac{\sum_{k \in U} \mathbb{1}_{\mathrm{i} \in \mathrm{m_v}}}{\sum_{k \in U} \mathbb{1}_{\mathrm{i} \in \mathrm{v}}} = \frac{M_v}{N_v}$, proportion calculée sur la population totale U .

On considère une fonction de totaux dérivable $f(t_1,...,t_q)$, avec t_1 à t_q les totaux des variables y_1 à y_q .

La variable linéarisée u_k est définie par :

$$u_k = \sum_{i=1}^{Q} \frac{\partial f}{\partial t_i}(t_k,...,t_q) y_{ki}$$

On considère une fonction de totaux correspondant au ratio $\theta = \frac{t_1}{t_2}$

La variable linéarisée est alors :

$$u_k = \frac{\partial f}{\partial t_1} y_1 + \frac{\partial f}{\partial t_2} y_2 \text{ Soit } u_k = \frac{1}{t_2} y_1 - \frac{t_1}{t_2^2} y_2$$

Avec y_1 indicatrice $\mathrm{I}_{i\in m_v}$ égale 1 si l'observation satisfait la modalité m_v

y_2 indicatrice $\mathbb{1}_{i\in v}$ d'appartenance au champ de la question

$$\text{Donc } u_k = \frac{1}{N}(\mathbb{1}_{i\in m_v} - P\mathbb{1}_{i\in v})$$

Qui est estimé par l'estimateur par substitution

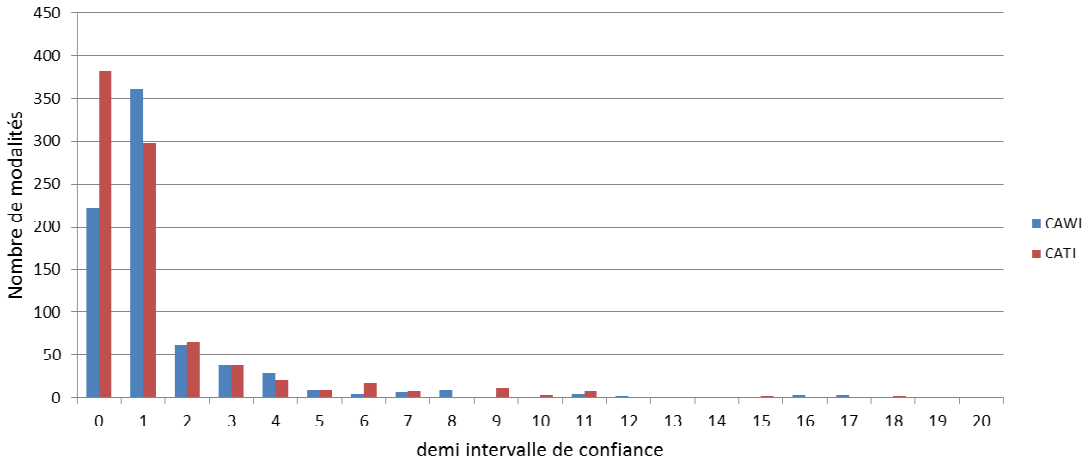
$$\hat{u}_k = \frac{1}{\hat{N}}(\mathbb{1}_{i\in m_v} - \hat{P}\mathbb{1}_{i\in v})$$

Pour chaque estimation on fait une estimation sur domaine. Le domaine n'est pas uniquement défini par rapport à S_d il faut aussi tenir compte du champ de la question. Si l'observation appartient à S_d et appartient au champ de la question alors $\hat{u}_k = \frac{1}{\hat{N}}(\mathbb{1}_{i\in m_v} - \hat{P}\mathbb{1}_{i\in v})$, sinon $\hat{u}_k = 0$

Ces observations qui n'appartiennent pas au domaine participent tout de même au calcul de la variance, en effet dans les étapes de régression sur les variables de calage, on peut avoir des valeurs qui ne soient plus égales à zéro, et dans ce cas participent à l'augmentation de la variance. Ces termes peuvent être très gênants sur les domaines extrêmement petits, où la variance peut être rédhibitoire.

Le graphe suivant présente les deux histogrammes des variances pour les collectes Cawi et Cati. En abscisse les variances estimées, et en ordonnée le nombre de modalités correspondant.

Comparaison des variances Cawi et Cati



Les variances Cati en rouge sont très souvent plus faibles que les variances Cawi, il faudrait voir les effets de sondage pour comparer les deux tirages.

La variance apportée par le terme lié à la non-réponse est relativement faible. Quelles que soient les options choisies le second terme spécifique à la non réponse ne représente que quelques pourcent de la variance totale, et n'agrandit l'intervalle de confiance que de quelques dixièmes de points. Ceci est surprenant vu l'ampleur de la non-réponse dans l'enquête Génération. Il est à noter, que le problème principal de la non réponse, en plus de diminuer la taille de la population considérée et d'augmenter la variance, est de générer un biais du fait que les non-répondants n'ont pas nécessairement le même comportement que les répondants.

4.3. Comparaison statistique du Cawi et du Cati

Les tests d'égalité des proportions sont faits pour chaque modalité des variables comparables. Le test effectué est un test d'égalité des moyennes. Les hypothèses de ce test sont que les échantillons sont indépendants, et que les observations sont indépendantes et identiquement distribuées. Dans notre cas les échantillons ne sont pas indépendants du fait que le tirage ait été fait dans le complémentaire de l'échantillon cati. Cependant après avoir retiré les individus qui n'étaient tirés que pour la collecte cati, du fait que les échantillons cumulés représentent environ un tiers de la base de sondage et que l'on peut supposer que les individus n'ont pas de lien entre eux, on fera comme si les échantillons étaient indépendants.

L'hypothèse nulle est que la différence des moyennes est égale à zéro.
L'hypothèse alternative est que la différence des moyennes est différente de zéro.

La statistique de test est $T = \frac{\mu_{cawi} - \mu_{cati}}{\sqrt{\sigma^2_{cawi} + \sigma^2_{cati}}}$

On se place dans le cas bidimensionnel. L'hypothèse nulle est rejetée si la statistique de test en valeur absolue est supérieure à 1,96, avec un risque de première espèce égale à 5%.
Près de 40% des variables ne présentent aucune différence de structure entre les collectes Cawi et Cati. A l'inverse, pour 30% d'entre elles, aucune des modalités ne présentent de proportions égales entre les deux collectes. Enfin pour trois variables sur dix on a une égalité statistique partielle entre les structures Cawi et Cati.

Pour illustrer les différents cas rencontrés dans la comparaison des structures des réponses entre les deux populations quelques exemples sont présentés.

- Variable situation à la date d'enquête SITDE, soit 3 ans après la sortie du système éducatif.

Comparaison des réponses Cawi et Cati de la variable SITDE

Situation à la date d'enquête	Cawi	Cati	p-valeur
emploi	76,5 (±1,1)	75,9 (±0,7)	0.48
chômage	11,9 (±0,8)	12,5 (±0,5)	0.34
autres inactivités	3,8 (±0,5)	2,7 (±0,3)	<0.01
formation	1,2 (±0,3)	1,7 (±0,2)	0.03
reprise d'étude	6,4 (±0,6)	7,2 (±0,4)	0.11

Cette variable ne présente pas beaucoup de différence de structure entre les deux collectes. Il n'y pas de différences significatives sur la part d'emploi, de chômage et de reprise d'études.En revanche, d'après le test les écarts entre les deux modes sur la part d'autres inactivités et de formation semblent significatifs.

- Variable potentiellement liée à un phénomène de désirabilité sociale.

Comparaison des réponses Cawi et Cati concernant l'opinion sur l'avenir professionnel

Comment voyez-vous votre avenir professionnel ?	Cawi	Cati	p-valeur
plutôt inquiet	32,0 (±1,1)	25,5 (±0,7)	$<10^{-3}$
plutôt optimiste	52,1 (±1,2)	72,0 (±0,8)	$<10^{-3}$
ne sait pas	16,8 (±0,9)	2,5 (±0,3)	$<10^{-3}$

On observe que les différences entre les structures Cati et Cawi sont grandes. De fait, cette question n'a pas été posée de la même manière dans le cadre de l'enquête Cati et dans l'enquête Cawi. En effet,pour la passation par internet, la modalité « ne sait pas » était directement affichée. Dans l'enquête Caticette modalité n'était pas énoncée par le télé-enquêteur. Ce dernier cochait la réponse

« ne sait pas » uniquement si c'était la réponse spontanée de l'enquêté.Cela explique facilement pourquoi cette modalité est moins fréquemment renseignée dans l'enquête Cati.
Pour pallier cette difficulté une possibilité serait de ne pas proposer dans l'enquête par internet la modalité « ne sait pas », et de l'afficher seulement si le répondant tente de passer à la question suivante sans répondre. Ce choix rapprocherait les passations internet et téléphone.

En dépit de ce phénomène, on constate que les répondants Cati sont plus pessimistes que les répondants Cati. En effet, il y a un écart de 6,5 points entre les deux collectes sur la modalité « plutôt inquiet ».Même en se plaçant dans la situation extrême où l'on « reclasserait » tous les « ne sait pas » du Cawidans la modalité « plutôt optimiste » et ceux du Cati dans la modalité « plutôt inquiet », il subsisterait encore un écart de 4 points.

On est donc surement ici en face d'un phénomène de désirabilité sociale, qui se traduit par le fait que le répondant choisisse plus particulièrement une réponse correspondant à une norme sociale ou une modalité plus facile à énoncer en présence d'un enquêteur à savoir la modalité « être plutôt optimiste ».

- Variable potentiellement liée à des effets de sélection et de mesure importants

Pour certaines variables on observe des différences significatives sans qu'il y ait de différence d'énonciation des modalités a priori. Ainsi sur la question concernant les « petits boulots » exercés en cours d'études nous avons les résultats suivants :

Comparaison des réponses Cawi et Cati concernant les petits boulots

Pendant vos études, <en dehors de votre contrat d'apprentissage/ emploi régulier> avez-vous eu des jobs de vacances ou des petits boulots ?	**Cawi**	**Cati**	**p-valeur**
1 = Oui, souvent (plus de 3 par an)	18,6 (±0,9)	32,6 (±0,8)	$<10^{-3}$
2 = Oui, parfois (3 ou moins par an)	57,9 (±1,2)	42,5 (±0,8)	$<10^{-3}$
3 = Non, jamais	23,5 (±1,0)	24,8 (±0,7)	0,07

Dans cette question il y a peu de modalités, on peut supposer qu'il n'y a pas d'effet de mémoire, du type recencyeffect. Cette question n'est a priori pas sensible non plus. La différence pourrait porter sur la compréhension des deux premières modalités : Est ce que l'on comprend les critères de ces deux modalités aussi bien par téléphone que lorsqu'on peut les lire. En supposant que le télé-enquêteur précise correctement les modalités, on peut supposer que les différences constatées sur cette question reflètent des différences inhérentes aux populations Cawi et Cati, on serait alors seulement sur un effet de sélection.

Conclusion

Les comparaisons sont limitées à cause du nombre d'observations. On a été obligé de restreindre l'étude dans un premier temps à un domaine du supérieur, puis nous avons choisi les variables qui correspondaient à un champ suffisamment large pour que l'on puisse les comparer sans difficulté. Les premiers résultats des comparaisons nous apprennent que pour près de la moitié des variables de la table individus on ne peut pas faire de comparaison directement sauf à faire des regroupements de modalités. Pour les deux tiers des variables comparables les différences en moyenne semblent assez faibles.

Il est à noter que la variance obtenue se fait en calant les données Céreq sur une source extérieure (en particulier sur l'enquête emploi). La variance est donc sous-estimée, puisque l'on ne prend pas en compte la variance de l'enquête sur laquelle on cale les données.

Une partie de la variance des estimations sur domaine vient des observations qui n'appartiennent pas au domaine mais qui pourtant participent à la variance totale. Pour améliorer les estimations sur domaine on peut ajouter les indicatrices des domaines d'intérêts dans les variables de calage. Mais ceci suppose de connaitre au préalable les domaines d'intérêts.

L'intérêt de la partie calage dans le contexte de comparaison des modes de collecte est de rapprocher des structures Cati et Cawi dans l'objectif de les rendre les plus comparables possibles. Cette étape participe donc en partie à la correction des effets de sélection. L'objectif est principalement de considérer les deux populations sur une structure qui permet d'analyser la cohérence des variables les plus centrales de l'enquête. Cependant ce calage ne permet plus de comparer les populations sur le niveau de sortie.

Les différences constatées peuvent être principalement dues à des effets de sélection non corrigés qui serait dus à des caractéristiques inobservables ou à des effets de mesure dus à la différence de passation entre les deux collectes. Il convient dès lors d'essayer de mettre en oeuvre d'autres méthodes pour essayer de déterminer de quelles natures sont les différences observées.

5. Lematching : une tentative pour distinguer les effets de sélection et de mesure

5.1. Présentation de la méthode

A l'issue de cette phase de comparaison des structures, le constat est que pour une grande partie des variables et questions il y a des différences significatives. Cependant ces écarts peuvent être expliqués, il peut s'agir d'effet de sélection, les répondants Cawi n'étant pas nécessairement les mêmes que les répondants par téléphone. Il peut s'agir aussi d'effets de mesure provenant du fait que les répondants n'ont pas le même comportement et ne perçoivent pas nécessairement la même information dans les deux contextes de passation.

Pour essayer d'apporter des éléments de réponses sur ces questions de la nature des écarts, une méthode de matching habituellement utilisée dans le cadre d'évaluation des politiques publiques va être mobilisée dans le contexte de la comparaison des modes de collecte. Pour essayer de prendre en compte les effets de sélection et les effets de mesure il est possible d'appliquer à la méthodologie d'enquête des méthodes économétriques,décrites par D. Fougère [6], pour évaluer les effets des dispositifs dans le cadre de politiques publiques. Idéalement pour mesurer l'effet d'une politique publique il faut qu'il y ait tout d'abord une association aléatoire entre les bénéficiaires potentiels et le dispositif pour éviter de confondre les effets du dispositif avec des caractéristiques individuelles sous-jacentes qui agiraient si la sélection était spontanée. Lorsqu'un dispositif est mis en place, sans cette expérimentation aléatoire préalable, les méthodes de matching sont parfois mobilisées pour corriger les effets de sélection potentielle dans l'accès au dispositif et déterminer si les bénéficiaires du dispositif voient leur situation changer par rapport à la variable d'intérêt. Dans le cadre de la comparaison des modes de collectes, le dispositif à évaluer est la passation par Cawi et l'effet du dispositif et l'écart de mesure entre les passations Cawi et Cati.

La variable de traitement T est égale à 1 pour les enquêtés répondant par Cawi et 0 par Cati, les répondants par téléphone constituent ici la population de contrôle. La variable d'intérêt est notée Y. Deux variables latentes sont associées à cette variable d'intérêt, Y1 et Y0 qui correspondent aux valeurs potentiellement renseignées par le répondant respectivement par internet et par téléphone. Ces variables latentes ne sont observées que partiellement et on a la relation :

$$Y = T * Y_1 + (1-T) * Y_0$$

L'hypothèse de cette approche par matching est de considérer que, conditionnellement aux variables qui expliquent la sélection, les variables latentes sont indépendantes du traitement. C'est-à-dire que conditionnellement aux covariables, il n'y a plus d'effet de sélection et l'assignation du traitement peut être considéré comme aléatoire. D'après la propriété de Rubin si on est dans ce cas d'indépendance des variables latentes Y_1 et Y_0 par rapport aux covariables expliquant la sélection, alors il y a aussi indépendance par rapport à la probabilité de recevoir le traitement, c'est à dire le score de propension. Les individus seront donc appariés par leur score de propension. Pour déterminer les contrefactuels de l'individu i, c'est-à-dire les individus du groupe de contrôle ayant des caractéristiques similaires, des estimateurs à noyau ou des appariements par les plus proches voisins peuvent être mobilisés. L'effet de mesure sera l'écart moyen estimé entre les individus traités et leur contrefactuel.

La méthode de matching permet de mobiliser en plus des variables initiales de la base de sondage, les réponses fournies par les enquêtés, ce qui permet d'enrichir le modèle de sélection. Il est difficile de mettre en œuvre cette méthode sur l'intégralité des variables du questionnaire. Cette méthode est appliquée à titre illustratif sur deux des variables qui ont déjà été données en exemple plus haut à savoir la situation à la date d'enquête et l'opinion par rapport à l'avenir professionnel. Deux modèles de sélection seront présentés : un modèle peu explicatif puis un modèle plus complet.

5.2. Applications

5.2.1. Application avec un modèle de sélection peu explicatif

Le premier exemple concerne la question portant sur l'opinion par rapport à l'avenir professionnel. Pour cette variable nous avions vu que les structures divergentes du Cati et du Cawi pouvaient être interprété en partie par un biais de désirabilité sociale, et donc de manière plus générale par un effet de mesure important. C'est l'hypothèse que nous allons essayer de vérifier dans ce qui suit en essayant d'isoler une différence significative

Comparaison des réponses Cawi et Cati concernant l'opinion sur l'avenir professionnel

Comment voyez-vous votre avenir professionnel ?	Cawi	Cati
plutôt inquiet	32,0 (±1,1)	25,5 (±0,7)
plutôt optimiste	52,1 (±1,2)	72,0 (±0,8)
ne sait pas	16,8 (±0,9)	2,5 (±0,3)

Pour modéliser la probabilité de répondre par internet et ainsi construire le score de propension qui permettra d'apparier les observations, le modèle logistique suivant a été exécuté sur le pool constitué par la concaténation des observations Cati et des observations Cawi, ces deux populations étant restreintes au domaine S_d .

- La variable à expliquer est l'indicatrice valant 1 si l'enquêté a répondu par internet et 0 sinon.
- Les variables explicatives sont le niveau des formations agrégées et le sexe.

Nous nous intéressons au fait de déclarer la deuxième modalité « Plutôt optimiste ». La variable à expliquer est donc l'indicatrice égale à 1 si le répondant déclare être plutôt optimiste et 0 sinon. Nous allons essayer de déterminer si la différence constatée vient des caractéristiques individuelles. Dans ce cas, après matching, il n'y aurait plus de différence entre le groupe de contrôle et le groupe des répondants par internet. Dans le cas d'un effet de mesure, même si la sélection est bien expliquée, nous devrions avoir des différences significatives dans le fait de déclarer être optimiste entre les deux modes de passation.

Le paramètre calculé ici à l'issue de cette méthode de matching est l'ATT (averagetreatmenteffect on the treated), c'est-à-dire l'effet moyen du traitement sur la population traitée, ici il s'agit de déterminer si la passation par internet conduit les répondants à plus ou moins se déclarer être optimiste que les répondants Cati. L'effet moyen donnera la différence entre les proportions de répondants Cawi qui répondent cette modalité et leur contrefactuel du Cati.

A l'issue de l'appariement, l'effet moyen de la passation par internet est de -0,23 pour la déclaration de la deuxième modalité. Ainsi la proportion de répondants Cawi déclarant être plutôt optimiste est 23 points inférieure à celle de leurs contrefactuels. Il s'agit de la totalité de la différence constatée. Ce qui est peu étonnant parce que le modèle de sélection utilisé ne rapproche pas beaucoup plus les populations que ne l'a déjà fait le calage. La p-valeur liée au test de nullité de l'ATT est de $2\ 10^{-16.}$.

Pour les autres modalités on procède de la même manière. Pour les première et troisième modalités les effets moyens sur le groupe internet sont respectivement de 0,025 avec une p-valeur de 0,0049 et de 0,13 avec une p-valeur de $2\ 10^{-16}$. Il y aurait effectivement pour cette variable un effet de mesure important.

Toujours avec le même modèle de sélection nous allons appliquer la même méthode pour la variable SITDE.

Comparaison des réponses Cawi et Cati de la variable SITDE

Situation à la date d'enquête	Cawi	Cati	ATT	p-valeur
emploi	76,5 (1,1)	75,9 (0,7)	-0,04	$2,6\ 10^{-7}$
chômage	11,9 (0,8)	12,5 (0,5)	-0,02	$7\ 10^{-4}$
inactivité	3,8 (0,5)	2,7 (0,3)	0,006	$4\ 10^{-2}$
formation	1,2 (0,3)	1,6 (0,2)	-0,005	$1,7\ 10^{-2}$
reprise d'étude	6,4 (0,6)	7,1 (0,4)	-0,009	$3,5\ 10^{-2}$

Sur cette variable, il n'y avait pas a priori de problèmes d'effets de mesure. Les modalités emploi chômage et reprise d'étude ont des proportions statistiquement égales entre les deux collectes et de légères différences sont présentes pour les modalités inactivité et formation. Il convient alors de s'intéresser essentiellement aux différences qui sont ressorties comme significatives lors de l'étape précédente d'estimation de la variance. Il apparait que sur ces deux modalités les écarts ressortent ici comme des effets de mesure significatifs et les différences ne peuvent pas être considérées comme des effets de sélection.

Pour toutes ces différentes modélisations, les résultats d'appariement avaient été contrôlés par le test du balancing score qui permet de vérifier qu'après conditionnement par les covariables, expliquant le fait de répondre par internet, les groupes des contrefactuels et des individus traités aient bien les mêmes moyennes sur les covariables. Dans ce cas, on considère qu'il y a bien une indépendance entre le fait de répondre par internet et l'appartenance à l'un ou l'autre des groupes.

Sur les deux exemples, les différences existantes entre les deux modes de collectes ressortent en tant qu'effets de mesure. En effet le modèle expliquant la sélection ne rapproche pas beaucoup plus les deux populations Cawi et Cati que les étapes de redressement. De ce fait, d'autres modèles pour modéliser le score de propension ont été testés pour affiner la sélection. Le modèle expliquant la sélection contient désormais en plus des variables citées, les variables REGION et Q25_7. Cette dernière est une indicatrice valant 1 si le sortant a arrêté ses études pour entrer sur le marché du travail et 0 sinon.

5.2.2. Application avec un modèle de sélection plus complet

Les résultats pour les variables portant sur l'opinion par rapport à l'avenir professionnel OP6 et la situation à la date d'enquête SITDE sont les suivants :

Comparaison des réponses Cawi et Cati concernant l'opinion sur l'avenir professionnel

Comment voyez-vous votre avenir professionnel ?	Cawi	Cati	ATT	p-valeur
plutôt inquiet	32,0 (1,1)	25,5 (0,7)	0 ,016	6 10-2
plutôt optimiste	52,1 (1,2)	72,0 (0,8)	-0,21	2.10-16
ne sait pas	16,8 (0,9)	2,5 (0,3)	0,12	2.10-16

Comparaison des réponses Cawi et Cati de la variable SITDE

Situation à la date d'enquête	Cawi	Cati	ATT	p-valeur
emploi	76,5 (1,1)	75,9 (0,7)	-0,01	$9\ 10^{-2}$
chômage	11,9 (0,8)	12,5 (0,5)	-0,023	$2\ 10^{-4}$
inactivité	3,8 (0,5)	2,7 (0,3)	8.10-4	$7,8\ 10^{-1}$
formation	1,2 (0,3)	1,6 (0,2)	-0,005	$1,2\ 10^{-2}$

reprise d'étude	6,4 (0,6)	7,1 (0,4)	-0,03	$3,5\ 10^{-15}$

En ayant des modèles un peu plus explicatifs de la sélection, on s'aperçoit que des différences de mesure qui étaient significatives avec le premier modèle de sélection ne le sont plus. Les écarts constatés contiennent donc moins d'effets de sélection. Le modèle 2 correspondant à l'ajout des variables REGION et Q25_7 (l'indicatrice justifiant l'arrêt des études pour entrer sur le marché du travail)satisfait le test du balancing score sur toutes les variables sauf sur un des niveaux de formation agrégé.

D'autres modèles incluant des variables additionnelles telles que l'âge ou une indicatrice égale à 1 si le sortant a atteint le niveau d'étude désiré ont été testés. Ces modèles plus étoffés ne permettaient cependant pas de satisfaire le test du balancing score. De manière plus générale, il est conseillé de ne pas essayer d'expliquer la sélection du dispositif à partir de modèle trop bien ajusté. Le risque étant que les supports des scores de propension ne se recouvrent pas entre le groupe des traités et le groupe de contrôle. Le cas extrême étant un score de propension de 1 pour les individus traités et de 0 pour le groupe de contrôle, auquel cas il n'y aurait pas d'appariement possible.

On s'aperçoit que les effets de modes ainsi mesurés dépendent complètement de la manière dont on a spécifié le modèle logistique pour expliquer le fait de répondre par internet. Dans le cas où le modèle ne prendrait pas assez en compte les caractéristiques individuelles, la plupart des effets de sélection seraient comptabilisés parmi les effets de modes. On voit ici le problème de dissociation des effets de sélection et des effets de mesures. Cette méthode de matching suppose que la sélection soit complétement expliquée par les observables. Ceci est une hypothèse forte, les caractéristiques inobservables influençant la sélection, et/ou les caractéristiques observables non contrôlées, seront donc comprises dans les effets de mesure.

6. Quelle stratégie pour agréger les données multimodes ?

Lors d'une collecte multimode, comme vu précédemment, nous pouvons être confrontés à un effet de sélection et à un effet de mesure.

La population peut être partagée en 4 catégories : celle qui souhaite répondre à l'enquête uniquement par téléphone (T), cellequi souhaite répondre à l'enquête uniquement par internet (W), celle qui répondra à l'enquête indifféremment sur le téléphone ou sur internet (TW) et enfin celle qui ne répondra pas à l'enquête quel que soit le mode de collecte (NR). Dans le cas d'une enquête monomode téléphonique, seules les populations T et TW répondront à l'enquête (et donc W et NR ne répondent pas). Dans le cas d'une enquête monomode internet, seules les populations W et TW répondront à l'enquête (et donc T et NR ne répondent pas). Enfin dans une enquête multimode, ce sont les populations T, W et TW qui répondront à l'enquête (et seule la population NR ne répond pas). Autrement dit, l'effet de sélection dans une enquête multimode correspond à une augmentation de la couverture. Cet effet peut donc être considéré comme un effet bénéfique des enquêtes multimodes.La phase de modélisation de la non-réponsesera relativement proche de celle de l'enquête monomode mais pourra intégrer des variables supplémentaires du type disponibilité et origine du mail en plus de celles déjà présentes sur le téléphone.

L'effet de mesure est lié au fait qu'un même individu répondra potentiellement différemment à une même question selon le mode de passation (en autoadministré sur internet ou en entretien avec un télé-enquêteur formé à l'enquête). Dans ce cas, indépendamment de l'identification de cet effet de mesure, plusieurs questions se posent : y-a-t-il une réponse meilleure que l'autre ? Comment agréger les données issues des deux modes alors que les comportements de réponses diffèrent selon le mode ?

Il n'y a pas de consensus dans la littérature sur la manière d'agréger les données provenant de plusieurs modes de collecte. Certains méthodologues pensent qu'il n'y a pas de mode de collecte idéal et que par conséquent une agrégation simple des données des différents modes de collecte donne une meilleure image de la réalité, les erreurs de mesure d'un mode de collecte étant compensées par celles de l'autre.

Pour d'autres, cette stratégie peut dégrader les estimateurs si les erreurs ne se compensent pas mais s'ajoutent. Dans ce cas, il serait préférable de considérer qu'il y a un mode de collecte prépondérant

et de corriger les effets de mesure pour faire correspondre les réponses du mode de collecte alternatif.

Différentes stratégies sont envisageables et présentées ci- après.

6.1. L'agrégation simple

Il s'agit de la méthode la plus simple à mettre en œuvre et probablement la plus fréquemment utilisée. Elle consiste à ne pas prendre de précautions particulières et à empiler les données issues des deux modes.
Cette méthode repose sur l'hypothèse forte qu'il n'y a aucun effet de mesure sur aucune variable ou bien que les biais de mesure de chaque mode se compensent. L'inconvénient principal est qu'en présence d'un effet de mesure sur une variable, l'estimation sera potentiellement biaisée et dépendra directement de la part respective des deux modes de collecte. En revanche, cette méthode présente plusieurs avantages : simplicité de mise en œuvre et fichier de diffusion classique.

6.2. Le calage sur marges

Cette méthode consiste à caler les résultats d'un mode sur l'autre sur un sous-ensemble de variables sur lesquels aurait été détecté un effet de mesure (après « annulation » de l'effet de sélection). La correction des poids permettrait alors d' « annuler » l'effet de mesure sur les variables retenues pour le calage.

Plus précisément, l'opération se ferait en trois étapes : la première consiste à effectuer un premier calage qui rentre dans la phase « classique » de redressement de la non réponse et d'harmonisation avec des sources externes. Dans une deuxième phase, on identifie les variables qui présentent un effet de mesure entre les deux modes, puis on sélectionne parmi elles (par exemple à l'aide d'une analyse de données) celles qui serviront à corriger les différences de mesure dans le calage final. Pour finir, ce calage en une seule étape intègre les variables qui ont servi au redressement et celles qui servent à la correction de l'effet de mesure.

Dans ce cas, il y a le choix d'un mode de collecte prioritairequi sert de référence pour le mode alternatif. Le choix d'un mode peut se faire soit pour des raisons historiques (en choisissant le mode utilisé dans les précédentes enquêtes pour limiter la rupture de série dans les comparaisons temporelles), soit pour des raisons de qualité (le mode retenu serait alors jugé de meilleure qualité que l'autre), soit pour des raisons de mode prédominant (si un des deux modes est très majoritaire parmi les répondants, celui-ci pourrait être choisi comme référence). Ce choix d'un mode de référence présente deux inconvénients majeurs. D'une part,il s'applique uniformément sur toutes les variables. Or, nous avons vu précédemment que la référence pouvait changer selon le type de variable. D'autre part, si le choix se fait sur le mode prédominant, des problèmes de comparabilité entre enquêtes successives se poseront si le mode majoritaire change dans le temps.

Par ailleurs, cette méthode, reposant sur l'utilisation d'un correctif de pondération unique, fait implicitement l'hypothèse que l'effet de mesure est constant sur toutes les variables qui n'ont pas servi au calage. L'inconvénient est donc le risque de biaiser l'estimation de ces autres variables y compris quand elles ne présentaient pas de problème de mesure.

Cette méthode présente les avantages d'une mise en œuvre par des méthodes standards ainsi que la mise à disposition d'un fichier de diffusion classique.

6.3. Introduction d'un correctif variable par variable

Cette méthode consiste pour une variable donnée présentant un effet de mesure à introduire un correctif permettant de « gommer » l'écart de mesure. Autrement dit, pour cette variable, on cale le résultat d'un mode sur l'autre.

Cette méthode nécessite également de définir un mode de référence mais à la différence de la précédente, cette référence peut changer d'une variable à l'autre. Ainsi, pour une variable d'opinion on pourra choisir la collecte internet comme référence, car jugée moins sensible que la collecte par téléphone au phénomène de désirabilité sociale. A l'inverse, pour une variable complexe, le téléphone

pourrait faire office de référence car le téléenquêteur prend en charge la difficulté de la question et maintient l'attention de l'enquêté.

Outre l'intérêt de changer de référence en fonction de la question, elle permet également de ne pas appliquer de correctif en cas d'absence d'effet de mesure sur la variable à considérer.

La méthode nécessite une automatisation des traitements relativement facile à mettre en œuvre (en partant toujours du postulat que l'effet de mesure a été estimé au préalable pour chaque variable). Pour chaque variable, une variable de correction du poids sera créée. Cela présente un inconvénient majeur, quasiment rédhibitoire, de créer un fichier de diffusion non standard. En effet, pour estimer une variable, il sera nécessaire de multiplier la pondération individuelle et le correctif de la variable. Cela complexifie grandement l'exploitation des données, en particulier dès que l'utilisateur souhaite croiser deux variables ou plus.

6.4. Omission d'un mode de collecte et imputation

Il s'agit ici de ne prendre en compte qu'un seul mode de collecte pour une variable qui présenterait un effet de mesure. Comme précédemment le mode de référence retenu peut, en théorie, changer en fonction de la variable.Cette méthode ne nécessite pas de connaître précisément l'ampleur de l'effet de mesure mais uniquement de déterminer si une variable est sensible au mode de collecte.

Dans ce cas, cette méthode est très facile à mettre en œuvre puisqu'il suffit pour une variable de considérer comme données manquantes les réponses données au mode secondaire.

Cette méthode peut s'appliquer en traitement post-collecte mais également en amont de la collecte. En effet, on peut déterminer a priori les questions qui posent un problème de mesure (par exemple à la suite d'une expérimentation) et décider de ne poser cette question que dans un seul mode (celui jugé de référence pour cette question).

La limite principale de cette méthode est de générer artificiellement des données manquantes sur certaines variables. En l'absence de traitement ad'hoc, cela provoque du biaiset signifie pour les utilisateurs de devoir retirer de leur analyse soit les individus concernés, soit les variables concernées.

Cette méthode (qu'elle soit réalisée en amont ou en aval) provoque donc volontairement de la non-réponse partielle sur certaines variables. Il y a donc la possibilité de réaliser des traitements classiques de non-réponse partielle (par exemple par régression ou hot deck).

<u>Imputation par régression sur le mode de référence</u>
Pour une variable Y, présentant un effet de mesure, on choisit le mode de référence (MR). On modélise Y sur les individus ayant répondu à MR à l'aide des variables qui ne présentent aucun écart de mesure. Ce modèle est ensuite utilisé pour prédire Y pour les individus ayant répondu au mode alternatif (MA).
Cette méthode permet de prendre en compte les spécificités de la population répondante au mode alternatif (sous réserve que le modèle estimé sur MR s'applique sur les sous-populations spécifiques de MA).

<u>Imputation par hot deck</u>
Plusieurs méthodes de Hot Deck peuvent être envisagées. Par exemple, on réalise une typologie des individus sur les variables ne présentant pas d'effet de mesure. On choisit aléatoirement un individu (le donneur) MR dans la classe dans laquelle se trouve l'individu MA. Ensuite, on impute pour toutes les variables présentant un problème de mesure,les valeurs du donneur. Pour cette méthode, il y a un risque de perdre les spécificités de la population répondante au mode alternatif.

Cette méthodeautorise en théorie de créer deux groupes de variables à imputer : celui pour lequel le téléphone est le mode de référence et celui pour lequel internet est le mode de référence. Cela pose tout de même une difficulté. En effet, cette méthode est envisageable si la part des données imputées n'est pas trop importante. Autrement dit, si le mode de référence est assez largement majoritaire parmi les répondants. Il serait en effet difficile d'accepter cette méthode si plus de 30% des individus répondent sur le mode alternatif provoquant donc plus de 30% d'imputation. Par conséquent, le choix

du mode de référence devrait se faire de manière uniforme sur l'ensemble des variables et devrait représenter a minima 70% des réponses.

Ces méthodes d'imputation de non réponse partielle présentent cependant le défaut important d'« oublier » totalement l'information donnée par l'enquêté (du mode alternatif)en la remplaçant par une valeur imputée.

6.5. La transformation de réponse

L'idée de cette méthode est de tenir compte de la réponse donnée par l'individu ayant répondu dans le mode alternatif.

Prenons le cas d'une variable dichotomique {0,1} identifiée comme ayant un problème de mesure.

Imaginons le cas fictif suivant :

	sur l'ensemble des répondants		après correction de l'effet de sélection	
	Mode de référence (MR)	Mode alternatif (MA)	Mode de référence (MR)	Mode alternatif (MA)
0	30%	65%	40%	55%
1	70%	35%	60%	45%

L'idée de cette méthode est de corriger les réponses d'une partie des répondants MA pour « s'approcher » de la répartition MR. Dans l'exemple donné, après avoir corrigé la sélection, il y a un écart de mesure entre les deux modes de +15/-15 points. Une première approche serait donc de modifier la réponse de 15% des individus ayant répondu par MA.
Pour ce faire, on utilise un modèle logistique estimé sur MR en utilisant uniquement des variables ne présentant pas d'effet de mesure. Ce modèle permettrait de prédire sur MA la probabilité estimé de répondre « 1 ».
Intuitivement, on pourrait classer les individus MA qui ont répondu « 0 » selon cette probabilité prédite et modifier la réponse de 15% des individus qui ont la probabilité la plus élevée de répondre « 1 ». La distribution dans le Cawideviendrait alors (50% « 0 » et 50% « 1 »).
Le côté déterministe de cette méthode incite à ajouter de l'aléa. Autrement dit, on tire aléatoirement parmi ceux qui ont répondu « 0 », proportionnellement à la probabilité prédite, les individus (15% de l'ensemble) dont on va modifier la réponse. On retrouverait également une répartition (50%-50%).

Ces étapes pourraient être répétées sur les données corrigées pour évaluer la robustesse de la méthode. On peut également envisager de reproduire cette méthode par sous-populations (par exemple par niveau de diplôme) plutôt que sur la population entière.

Cette méthode de transformation de réponse repose sur l'hypothèse que l'estimation du modèle n'est pas affectée par l'effet de sélection. Pour pallier ce défaut, on peut par exemple utiliser des modélisations du type biprobitqui tiennent compte du biais de sélection.

Une généralisation de la méthode précédente est envisageable pour des variables continues discrètes, qualitatives ordonnées ou non ordonnées, en utilisant les modélisations adaptées de type polytomique ordonnée ou non ordonnée. En revanche, il sera plus délicat de prendre en compte les effets de sélection dans ces modèles.

Nous considérons que ces méthodes, notamment les plus difficiles à mettre en œuvre comme celles qui reposent sur un modèle ou une typologie (imputation ou transformation de réponses), ne peuvent être utilisés que dans le cas où le nombre de variables présentant un effet de mesure est relativement limité. Les efforts devront donc, comme cela a déjà été dit précédemment, se porter avant tout sur la réduction de cet effet de mesure en amont, par une bonne formulation des questions et une ergonomie adaptée.

Toutes les méthodes d'agrégation proposées qui prennent en compte une correction d'un effet de mesure reposent sur l'hypothèse forte de bonne séparation préalable des effets de sélection et de mesure. Le nouveau protocole expérimental mis en place en 2015 tentera de progresser sur ce point en améliorant d'une part le modèle de sélection dans le modèle de matching et d'autre part en proposant un protocole permettant de gommer sur un sous échantillon (du moins d'un point de vue théorique) l'effet de sélection lié au mode de collecte. Cette nouvelle expérimentation permettra ainsi de mettre en œuvre les méthodes d'agrégation proposées ci-dessus (et éventuellement d'autres) pour arbitrer sur la méthode qui sera retenue dans l'avenir du dispositif.

7. Une nouvelle expérimentation

7.1. Objectifs généraux de cette expérimentation

Le premier objectif est d'expérimenter un système d'interrogation multimode (internet et téléphone) où la découverte des aspects pratiques d'organisation (base de données unique pour le questionnaire Cati et le questionnaire Cawi, basculement d'un mode à l'autre, gestion des relances par mail et par téléphone, …) est aussi importante que les développements méthodologiques.
Après le bilan de la première expérimentation de 2013, plusieurs chantiers sont envisagés pour améliorer l'expérience web :

- Prendre en compte les enseignements de la première expérimentation (gestion de l'emailing de masse, adresse de l'expéditeur, qualité du mail-avis,…)
- Améliorer l'ergonomie et donc faciliter la passation du questionnaire en auto-administré
- Posséder plus d'outils pour distinguer les effets de sélection des effets de mesure
- Mettre en œuvre les premières stratégies d'agrégation des données

Deux échantillons, tirés parmi les répondants de la 1ère vague de l'enquête principale, sont prévus dans le cadre de cette expérimentation :

- un échantillon de 4446 individus, pour lequel internet est le mode de collecte prioritaire et le téléphone permettrait de garantir les taux de réponses attendus,
- un échantillon embarqué de 1000 individus, pour lequel les répondants auront un mode de collecte imposé aléatoirement.

7.2. Amélioration de l'ergonomie du Cawi

Les abandons en cours, en partie liés au manque d'ergonomie, ont imposé de préciser plus clairement dans le cahier des charges du marché le niveau d'exigence concernant la conception du questionnaire Cawi. Les attentes sont les suivantes :

- Les pages de consignes seront remplacées par des consignes courtes, apparaissant au moment venu à l'aide de fenêtres pop-up.

- Le calendrier d'activité qui est à la fois le cœur de l'enquête et la partie qui provoque le plus d'abandon devra satisfaire les caractéristiques suivantes :
  - La saisiedes séquences d'activité devra être interactive et facilement utilisable avec la souris.
  - La saisie de l'information se fera de manière optimisée (un clic pour le début de période, un clic pour la fin de période, ou alors une sélection de la période avec un « cliquer-glisser »).
  - La correction d'une information saisie devra se faire de la manière la plus optimale possible (par exemple, clic droit sur la période puis supprimer,rallonger ou raccourcir une période déjà saisie sans être obligé de ressaisir l'intégralité de l'information,intégrer une période entre deux périodes déjà existantes.
  - La consultationd'une période saisie sera facilitée (par exemple, clic droit puis consulter, ou passage de la souris qui génère l'affichage d'une info-bulle).
  - Les contrôles de cohérence se feront en direct (une seule situation décrite pour un mois donné, pas de mois non renseigné) avec alerte sous forme de pop-up.

- Un récapitulatif modifiable à la finde lasaisie du calendrier, alternant des couleurs différentes sur les séquences consécutives.

- Une aide en ligne sera aussi proposée permettant de disposer d'information plus complète, de tutoriel vidéo pour la passation du calendrier d'activité. Le numéro vert pour appeler un téléenquêteur ainsi que l'adresse électronique dédiée à cette opération de collecte seront présents sur chacune des pages pour obtenir de l'aide.

- La gestion des menus (communes, professions, titres et diplômes, …) devra être optimisée pour faciliter la recherche.

- Les questions à choix multiples seront testées sous trois formes (en « cases à cocher », en tableau ou succession de questions oui/non)

7.3. Outils supplémentaires pour distinguer les effets de sélection des effets de mesure

La méthode de matching mise en œuvre lors de la première expérimentation n'a pas permis de conclure définitivement sur les parts des effets de sélection et de mesure quant aux écarts constatés (entre le Cati et le Cawi). Cette nouvelle expérimentation nous permettra d'améliorer cette première méthode et de tester un nouveau protocole permettant potentiellement de distinguer ces effets.

7.3.1. Ajout de variables pour la méthode de matching

Dans une première approche des questions seront ajoutés permettant de prendre en compte la propension des répondants à utiliser internet afin de mieux spécifier le modèle de sélection dans une approche par matching. Les questions envisagées pour essayer de mieux contrôler les effets de sélection sont les suivantes :

Combien de fois avez-vous utilisé internet ces trois derniers mois que ce soit à domicile, sur le lieu de travail ou ailleurs ?
Tous les jours ou presque tous les jours 1
Au moins une fois par semaine mais pas tous les jours 2
Au moins une fois par mois mais pas chaque semaine 3
Moins d'une fois par mois 4

Pour réaliser vos démarches de la vie quotidienne, consultation de votre compte bancaire ou démarches administratives vous utilisez internet :
Rarement 1
Parfois 2
Souvent 3
Systématiquement 4

7.3.2. Utilisation d'un échantillon embarqué

Une seconde approche est d'embarquer un échantillon expérimental dans l'échantillon de l'enquête multimode. Sur cet échantillon, de taille réduite (1000), le mode de collecte des répondants serait déterminé de manière aléatoire, ce qui aurait pour effet de gommer les effets de sélection et donc d'obtenir deux populations aux caractéristiques a priori identiques sur chaque mode de réponse. Les différences observées entre les réponses Cawi et les réponses Cati sur cet échantillon seraient alors essentiellement des différences dues aux effets de mesure.

Dans ce protocole, les enquêtésseront informés par deux lettres avis (papier et numérique). Les deux courriers mentionneront un lien web et un numéro vert. Ensuite un système de relance est mis en place par mail (10 relances), téléphone (20 appels sur le numéro qui a servi à réaliser la première enquête, dont un message laissé sur répondeur), et un SMS. Deux cas se présentent :

- l'enquêté à partir du courrier papier ou de l'un des mails entreprend lui-même de répondre soit en utilisant le lien web, soit en appelant le numéro vert.
- L'enquêté est contacté directement par téléphone et accepte de répondre au questionnaire.

Dans ces deux cas, on considère que l'enquêté a accepté de répondre au questionnaire. C'est à ce moment, que l'enquêté est invité à répondre au questionnaire sur le mode qui lui a été assigné aléatoirement. Dès lors, trois cas se présentent :

- L'enquêté est déjà sur le mode qui lui a été assigné. Dans ce cas, il poursuit naturellement le questionnaire (de manière totalement transparente)
- L'enquêté doit changer de mode.
  - ⇨ S'il est au téléphone, l'enquêteur, à l'aide d'un plan de dialogue adapté l'invitera à répondre au questionnaire sur internet (avec une vérification du mail et l'envoi d'un mail automatique)
  - ⇨ S'il est sur internet, l'enquêté sera invité à appeler le numéro vert pour passer l'enquête par téléphone ou prendre un rendez-vous téléphonique.

L'affectation du mode de collecte se faisant après l'acceptation de répondre au questionnaire, ce protocole devrait permettre de gommer les effets de sélection lié au mode. La difficulté principale est liée au fait que les enquêtés risquent de ne pas répondre au questionnaire (malgré leur acceptation initiale) en particulier lorsqu'on leur impose de basculer sur un autre mode de collecte. L'expérimentation de ce protocole nous donnera des renseignements précieux sur la viabilité d'une telle méthode.

Le protocole du premier contact réel (lettre avis papier et numérique, relance téléphonique) est identique pour l'échantillon multimode « classique » et l'échantillon embarqué afin d'assurer que les caractéristiques des populations au moment du contact soient le plus proche possible. Les populations répondantes seront différentes car les protocoles de réponses ne sont pas les mêmes (multimode « classique » VS affectation aléatoire d'un mode de réponse). Un redressement spécifique tiendra compte notamment du changement de mode.

Au final, avec cette prochaine expérimentation, nous disposerons de deux approches pour appréhender les effets de sélection et de mesure. La comparaison des résultats nous permettra d'avoir des indications sur la robustesse des conclusions.

8. Conclusion

L'objectif de cet article est de fournir des éléments de réponses à la question posée : la collecte par internet est-elle l'avenir des enquêtes Génération du Céreq ?

La première expérimentation de collecte monomode internet révèle plusieurs faiblesses. Sil'indisponibilité des adresses email et de l'absence de lettre avis papier ont indéniablement joué sur la non-réponse en amont, cette étude a permis de déterminer l'ampleur des abandons en cours de collecte et les principales questions qui les génèrent. En particulier, la présence de longues consignes, la saisie d'information très spécifique liée à la description de l'activité semblent rédhibitoires pour certains enquêtés sur le web. Pour alléger cette phase du questionnaire, des solutions ergonomiques semblent envisageables pour remplacer les séries de questions des séquences par un calendrier plus interactif ou ludique à renseigner. Il pourrait s'agir aussi de remplacer ou d'associer aux consignes écrites d'autres formats pour expliquer ce qui est attendu, une animation ou des pop-up interactifs par exemple. Une veille sur les collectes internet permettra de repérer les exemples de développement de Cawi qui pourraient être utilisées dans l'enquête Génération.

Cette étude a permis de déterminer, à l'aide d'un calcul de variance, l'ampleur des écarts entre les données Cati et Cawi : 40% des variables comparables n'ont pas de différence significative. A l'inverse 30% des variables présentent des écarts importants sur toutes leurs modalités entre les deux modes de collectes. Ces écarts s'expliquent à la fois par un effet de sélection et par un effet de mesure.Pour tenter de distinguer ces deux effets, des méthodes de matchingont été mises en œuvre.

Les premiers résultats montrent qu'il est difficile de conclure sur la nature des différences observées entre les deux collectes. Dans les études traitant de ce sujet, il apparait que la distinction des deux effets demeure difficile.

En supposant avoir résolu le problème d'identification des effets de mesure, le principal enjeu dans le contexte d'enquête multimode est de déterminer comment agréger les données issues de chaque mode en tenant compte du problème de mesure. L'étude propose une réflexion sur quelques stratégies d'agrégation des données en présentant les hypothèses, les avantages et inconvénients. La plupart d'entre elles nécessitent de définir au préalable un mode de collecte de référence. Selon les méthodes, cette référence sera unique pour toutes les variables du questionnaire ou bien pourra être déterminée en fonction du type de variable.

Pour la prochaine expérimentation réalisée en 2015 par le Céreq,cette fois-ci en multimode Cati-Cawi, le premier objectif sera d'améliorer la formulation des questions et l'ergonomie pour réduire les abandons en cours de collecte ainsi que les effets de mesure entre les deux modes.
Ensuite, il sera nécessaire de mieux caractériser les spécificités des sortants qui répondent par internet, afin de spécifier un modèle de sélection adéquat. Pour aider ce travail d'identification des effets de modes et des effets de mesures plusieurs actions pourraient être menées. L'ajout de variables dans le questionnaire portant sur la propension des répondants à utiliser l'informatique pourrait permettre de mieux contrôler les effets de sélection. En effet même si l'outil informatique est démocratisé, il ressort d'une étude Pisa que l'utilisation et la maitrise des outils informatiques sont hétérogènes. Au-delà de cet ajout, le nouveau protocole mettra en place un échantillon expérimental embarqué. Cet échantillon expérimental sera construit pour associer aléatoirement un mode de collecte aux sortants qui acceptent de répondre. En effet, les méthodes de matching supposent que les variables expliquant la sélection soient observables. D'autre part, il est déconseillé de spécifier un modèle de sélection qui soit trop ajusté. L'ajout d'un échantillon expérimental permettrait potentiellement de mieux contrôler les effets de sélection et les effets de mesures. Sur cet échantillon, en effet, il n'y aurait plus d'effet de sélection, les différences observées seraient alors uniquement de l'effet de mesure. Les deux méthodes (matching et échantillon embarqué) pourront d'ailleurs être confrontées. Enfin, cette nouvelle expérimentation permettra dans un contexte multimode de tester les différentes méthodes d'agrégation des données.

La mise en place dans le dispositif Génération d'une collecte par internet, en multimode avec le mode téléphonique,implique de résoudre de nombreuses difficultés et provoquera potentiellement une rupture de série dans le dispositif. Pour autant, il y a de réels motifs pour basculer à terme sur un tel dispositif car ces difficultés (disponibilité du mail, ergonomie, effet de mesure, agrégation) semblent surmontables sans oublier la réduction sensible du coût de l'enquête. De plus, la possibilité de répondre par internet parait désormais un mode particulièrement adapté pour la cible de notre enquête : les jeunes.

Bibliographie

[1] Anna-Maria Schielicke. (s.d.). "Don't know" the difference.An experimental comparison between Web and CATI.
[2] Annette, J., Caroline, R., & Peter, L. (2008, February). Assessing the Effect of
Data Collection Mode on Measurement.
[3] Dillman, D. A., Phelps, G., Tortora, R., Swift, K., Kohrell, J., Berck, J., et al. (2009). Response rate and measurement differences in mixed-mode surveys using mail, telephone, interactive voice response (IVR) and the Internet.

[4] Frippiat, D., & Marquis, N. (2010). Les enquêtes par internet en sciences sociales: un état des lieux. population.
[5] Krosnick, &jon, L. C. (2010). Comparing oral interviewing with self-administrated computerized questionnaires an experiment.
[6] Fougère D, les méthodes économétriques d'évaluation, RFAS 1-2-2010
[7] de Peretti G., Razafindranovona T. (2014), Les enquêtes multimode : attention aux effets de mode, Statistique et société, vol2, N°2