Journées de méthodologie statistique de l'Insee 2022

ESTIMATION ET DÉCOMPOSITION DE L'EFFET DE MODE DANS LES ENQUÊTES MULTIMODES (INTERNET/TELÉPHONE)

Gaëlle DABET, Zora MAZARI, Ines OUJIA

Céreq, Équipe ingénierie et gestion d'enquêtes

gaelle.dabet@cereq.fr zora.mazari@cereq.fr ines.oujia@cereq.fr

Mots-clés : Multimode, agrégation, matching, estimation, économétrie

Domaines concernés : multimode, effet de mode

Résumé

Depuis la fin des années 1990, le Céreq a mis en place un dispositif d'enquêtes original qui permet d'étudier l'accès à l'emploi des jeunes sortis du système éducatif une même année. L'enquête Génération, historiquement administrée par téléphone, a été rénovée et intègre désormais un nouveau mode de collecte : internet. La première édition de l'enquête produite en multimode est la Génération 2017, interrogée en 2020.

Des expérimentations multimodes ont été conduites pour tester plusieurs protocoles de collecte. Elles ont mis en avant les enjeux liés à l'introduction d'un mode de collecte en auto-administré : la formulation des questions, les consignes, la disposition des éléments sur une page doivent permettent de collecter des données comparables (téléphone/internet). Ces expérimentations ont également permis des premières estimations des effets de mode sur les variables d'intérêt de l'enquête. Le multimode induit à la fois des effets de sélection – le choix du mode de collecte diffère selon le profil des répondants - et des effets de mesure – un même enquêté peut répondre différemment à la même question selon le mode. Par la suite, le questionnaire Génération a été retravaillé, tant sur le fond que sur la forme, de manière à limiter ces effets.

Cet article s'inscrit ainsi dans la poursuite des travaux présentés lors des deux dernières JMS [1][2]. Il propose une analyse des effets de mode au travers de deux enquêtes multimodes distinctes dont les protocoles de collecte sont sensiblement différents : la première interrogation la « Génération 2017 », réalisée en 2020, et l'enquête Génération « Covid et après ? », ré-interrogation de la « Génération 2010 ». Cette dernière dispose d'un échantillon embarqué avec affectation aléatoire du mode de collecte (deux sous-échantillons monomodes) permettant de contrôler de l'effet de sélection. Pour chaque enquête, différentes méthodes seront implémentées et testées.

Concernant l'enquête Génération 2017, une méthode de matching sur score de propension a été mise en œuvre. Les différentes étapes de son application seront présentées : estimation du score de propension et mesure de sa qualité, choix de la stratégie d'utilisation du score dans l'estimation des effets. Plusieurs méthodes d'appariements ont aussi été évaluées (en fonction des contraintes d'équilibrage) afin d'obtenir des résultats satisfaisants.

Du côté de l'enquête Génération, « Covid et après ? », dans la mesure où les échantillons sont indépendants, l'effet de sélection est *a priori* maitrisé. Il va s'agir de repérer les éventuels biais de mesure en utilisant par exemple une méthode d'ajustement par les pondérations.

Nous proposerons des solutions pour aider à l'utilisation des données d'enquête. Quelles analyses et quelles interprétations seront possibles lors des exploitations ? Nous préconiserons également des ajustements dans l'outil de collecte pour les prochaines interrogations.

Abstract

The 'Génération' surveys study the access to employment of a cohort of young people who completed their training or education in a given year. These surveys, historically carried out only by telephone, are now conducted via both the internet and telephone. This article provides an analysis of the impact of the survey modes through two distinct multimodal surveys with significantly different data collection protocols: 'Génération 2017', and 'Covid et après ?' carried out in 2020 and 2021. For the 'Génération 2017' survey, several propensity score matching methods were implemented in order to obtain satisfactory results. The 'Covid et après ?' survey included two control groups with random assignment of the collection mode to control for the selection effect. An adjustment method based on weights was used on these groups to identify possible measurement biases. The quality of matching and weighting methods and variables having a measurement effect are discussed, along with guidelines for using survey data.

Introduction

Depuis la fin des années 1990, le Céreq a mis en place un dispositif d'enquêtes original - Génération - qui permet d'étudier l'accès à l'emploi des jeunes sortis du système éducatif une même année. Ces enquêtes portent sur les sortants des établissements de formations situés en France et concernent l'ensemble des niveaux et spécialités de formation. Les jeunes d'une Génération sont interrogés 3 ans et 6 ans après leur sortie du système éducatif. Après 20 ans d'existence, le Céreq a décidé de rénover en profondeur ce dispositif de la statistique publique tant d'un point de vue scientifique, sur le champ et le contenu du questionnaire, que méthodologique, sur le mode de collecte notamment. L'enquête Génération, historiquement administrée par téléphone, intègre désormais un mode de collecte supplémentaire : internet. La Génération 2017, interrogée en 2020-2021, constitue la première édition de l'enquête produite en multimode.

L'introduction de ce mode complexifie le processus de collecte. Pour réussir au mieux le passage au multimode, le Céreq a mené diverses enquêtes expérimentales. Elles ont permis de détecter des effets de mode sur certaines variables d'intérêt de l'enquête. En effet, la collecte multimode induit à la fois des effets de sélection (le choix du mode de collecte diffère selon le profil des répondants) et des effets de mesure (un même enquêté répondrait différemment à la même question sur le mode internet ou téléphone). Les effets de sélection ne posent pas de problème dès lors que les réponses via les deux modes sont exploitées ensemble. En revanche, les effets de mesure introduisent des biais dans l'analyse et l'interprétation des données.

Les expérimentations menées auprès des Générations 2010 et 2013 ont permis de formuler des préconisations afin de limiter ces biais et pouvoir les mesurer*, a posteriori*. En amont de la collecte, pour réduire les effets, un ajustement du questionnaire est nécessaire. Une attention particulière doit être portée aux variables présentant un effet de mesure détectées lors des expérimentations (évolution de la formulation par exemple ou suppression dans le questionnaire). Dans la phase de production, pour mesurer ces effets, il est avantageux d'introduire un échantillon de contrôle avec affectation aléatoire du mode de réponse. En aval, l'expertise de différentes méthodes de repérage des effets conduisent à choisir la méthode de matching sur score de propension.

Cet article s'attache à décrire les méthodologies appliquées pour estimer les effets de mode dans deux enquêtes récentes menées auprès des Générations 2010 et 2017. Il s'agit de deux enquêtes aux protocoles sensiblement différents (annexe 1) : l'enquête multimode Génération 2017 (première interrogation de la cohorte, en 2020), qui suit un protocole classique séquentiel puis concurrentiel, et l'enquête multimode Génération « Covid et après ? », réalisée en 2021 auprès de la Génération 2010[1] dont le protocole est similaire et intègre en plus deux sous-échantillons monomodes téléphone et internet. L'analyse des effets de mode de l'enquête Génération 2017 se traduit par la mise en application d'une méthode de matching sur score de propension. Pour l'enquête Génération « Covid et après ? », avec la présence d'un échantillon de contrôle, une méthode alternative est utilisée : l'ajustement par la méthode de pondération inversée. Toutefois, pour compléter l'analyse, un matching a également été réalisé sur cette dernière enquête pour confronter les résultats obtenus avec les deux méthodes (pondération inversée d'un côté, matching de l'autre).

Tout l'enjeu de cet article est de montrer dans quelle mesure les choix établis se révèlent judicieux pour l'estimation des biais de mesure : *Quelle est la robustesse de la modélisation du score de propension ? Les paramètres d'équilibrage définis garantissent-ils un appariement de qualité ?* Une interprétation des résultats obtenus sera également proposée : *Quel est l'origine du biais observé ? Quel est le seuil d'acceptabilité en termes d'écart entre les deux modes de collecte pour une variable présentant un effet de mesure ?* Enfin, concernant les méthodes d'agrégation, différentes solutions seront envisagées selon l'interprétation de l'effet de mesure et le type de question concerné.

[1] Quatrième interrogation, hors dispositif Génération

1. Utilisation d'une méthode d'appariement sur score de propension pour Génération 2017

1.1. Description de la méthode

Comment mesurer les biais induits par l'introduction d'un nouveau mode de collecte – internet ? Différentes méthodes économétriques d'évaluation sont disponibles. Sachant que les résultats [3] sont fortement conditionnés par la méthode choisie, ses hypothèses et le paramétrage de ses spécificités, il est important de déterminer celle qui s'adapte le mieux aux caractéristiques de l'enquête. À savoir, le protocole de collecte mis en œuvre, les particularités du corpus répondant, mais aussi les objectifs des analyses.

La plupart des méthodes consistent à comparer deux populations. La cohorte d'individus de la Génération 2017 est composée de répondants téléphone et internet. Étant donné le nombre de répondants équilibré dans chacun des modes et un protocole proposant le choix du mode, l'utilisation d'une méthode d'appariement sur score de propension semble opportune pour tenter d'éliminer dans un premier temps les biais de sélectivité (ou d'autosélection) et estimer par la suite les biais de mesure.

D'autres méthodes ont été testées lors des expérimentations portant sur les précédentes enquêtes Génération. La méthode d'appariement sur score de propension avait été retenue pour la poursuite des travaux. Les autres méthodes n'ont pas été implémentées sur les nouvelles données de la Génération 2017. Cette méthode consiste à apparier chaque individu traité avec son « jumeau » non traité et possédant les mêmes caractéristiques puis de mesurer l'effet du traitement. Dans le cadre de cette enquête, le traitement[2] s'apparente au fait de répondre au questionnaire par Internet[3]. Lors des analyses des expérimentations de l'enquête Génération, une estimation de l'ATT (l'effet moyen du traitement chez les traités) a été réalisée et sera poursuivie.

Les analyses suivantes porteront uniquement sur les données non pondérées.

Si l'on considère T comme la variable de traitement ($T=1$, répondant choisissant internet et $T=0$, répondant choisissant le téléphone) et les variables de résultats Y_1 et Y_0 correspondant aux réponses de l'individu, alors l'effet causal du traitement est défini pour chaque individu par l'écart $\Delta = Y_1 - Y_0$ qui représente la différence entre ses réponses par internet et celles qu'il aurait déclarées s'il avait répondu par téléphone. Si les variables de résultats sont indépendantes de la variable d'accès au traitement, l'ATT peut être estimé simplement comme la moyenne des différences de résultat observées entre les deux populations.

$$\Delta^{TT} = E(Y_1 - Y_0 | T = 1)$$

Dans les enquêtes multimodes (téléphone/internet), telles que l'enquête Génération 2017, cette propriété d'indépendance n'est souvent pas satisfaite. L'estimateur formé par la moyenne des différences des variables de résultat est affecté d'un biais de sélection car la composition des deux populations n'est pas identique (en termes de caractéristiques individuelles observables).

Toutefois, l'ATT suppose que l'on impose le traitement à un groupe d'individus. Dans le protocole de collecte de l'enquête Génération 2017, le choix du mode est offert. Ainsi, pour pouvoir utiliser et interpréter l'indicateur ATT, on s'appuie sur une hypothèse d'indépendance conditionnelle à des caractéristiques observables des individus. Grâce au matching, une forme de randomisation du mode serait maintenant assurée par les covariables X.

---

[2] Le terme de traitement se réfère aux premiers travaux ayant permis de développer un cadre conceptuel (modèle d'évaluation adapté à la situation dans laquelle un traitement peut être administré ou non à un individu – Rubin, 1974), travaux qui concernaient l'évaluation de l'efficacité des traitements dans le domaine médical. Bien qu'il ne soit pas le plus approprié, il est utilisé en économétrie pour qualifier une intervention publique, une réforme fiscale, une politique de subvention, un programme de formation, ou bien un programme d'aide sociale que l'on cherche à évaluer. [3]

[3] Génération 2017 ne dispose d'aucun échantillon de contrôle. Ainsi, le fait d'être traité dans l'analyse de l'ATT traduit en réalité à la fois le fait de répondre ou non, et le fait de choisir le mode internet ou téléphone. Plus précisément, pour chaque personne, quatre situations sont possibles : répondre par internet, répondre par téléphone, ne pas répondre par internet et ne pas répondre par téléphone. Dans l'effet de sélection sont donc mesurés deux effets, un lié à non-réponse et l'autre lié au choix du mode. Néanmoins pour la suite, les calculs seront poursuivis dans un cadre binaire avec deux événements : répondant par internet ou répondant par téléphone.

La méthode d'appariement sur score de propension repose sur un processus en plusieurs étapes. Pour sa mise en œuvre, on modélise la propension à répondre par internet, et on réalise des appariements en testant différents paramètres. Une fois l'appariement effectué, il s'agit de vérifier l'ajustement des deux populations traitée et non traitée. On s'appuie alors sur deux indicateurs préconisés pour valider le processus : l'amplitude du support commun et la vérification de la propriété équilibrante.

Le support commun, défini par la plage commune de distribution des scores, doit être suffisamment large pour que l'appariement soit possible sans écarter un nombre important d'individus de l'analyse.

La propriété équilibrante résulte de la comparaison des distributions des covariables utilisées pour calculer le score de propension (au sein des deux groupes : les répondants par internet et leurs « jumeaux » appariés, répondants par téléphone). Plus les distributions sont proches, plus l'ajustement du modèle est bon, garantissant la comparabilité des deux groupes et par conséquent un contrôle des biais de sélection.

La pratique habituelle des producteurs d'enquêtes consiste à sélectionner les principales variables d'intérêt pour évaluer la présence d'un effet de mesure. Le choix effectué pour cette première enquête Génération produite en multimode a été de réaliser les analyses sur la quasi-totalité (annexe 2) des variables du questionnaire. Pour analyser les effets, la méthode nécessitera un Matching sur score de propension par modalité pour chaque variable.

L'objectif est d'avoir un diagnostic complet des variables soumises à effet de mesure, dans le but d'améliorer du questionnaire pour les prochaines interrogations. Il ne s'agit pas en revanche de corriger les biais de mesure de l'ensemble des variables affectées. Le nombre de variables à corriger ne pourra être que limité.

1.2. Estimation du score de propension

Pour mettre en œuvre l'appariement, le préalable est d'estimer un score qui mesure, pour chaque individu composant le corpus de répondants, la probabilité de répondre à l'enquête par internet (indépendamment du mode effectif choisi). Le calcul du score porte dans cette analyse sur des données non pondérées. Plus le nombre de variables introduites pour estimer le score de propension sera élevé, meilleure sera la description de la probabilité de traitement, mais il faut veiller à ne pas dégrader le support commun. En appariant les individus ayant répondu sur chaque mode en fonction de leur score de propension[4], on équilibre la distribution des variables X dans les populations de traitement et de contrôle, c'est-à-dire que l'on rend ces deux groupes semblables du point de vue de la distribution des variables agissant sur la probabilité de répondre par internet. Pour autant, il n'est pas possible d'affirmer que la méthode de l'appariement avec score de propension élimine systématiquement le biais de sélection. Ainsi, Heckman, Ichimura et Todd ont montré que, dans certains cas au moins, les biais de sélection résiduels obtenus après application de la méthode d'appariement peuvent être de l'ordre de grandeur de l'effet du protocole lui-même. Par exemple, si le protocole prévoit de relancer une sous-population plus particulièrement par téléphone, le protocole sera davantage explicatif du mode de collecte que le profil du répondant [5].

1.2.1 Champ d'analyse pour le calcul du score

Environ 25 000 individus ont répondu à l'enquête Génération 2017 (champ Céreq). La part des répondants par internet et téléphone sont proches : 50 % par téléphone, 43 % par internet et 7 % sur plusieurs modes. Ces derniers ont été retirés de l'analyse des effets de mesure pour éviter la création d'une indicatrice de mode de collecte pour chacune des variables d'intérêt de l'enquête.

Les répondants proviennent soit de l'échantillon principal (initial), soit de la réserve. Ces deux sous-échantillons n'ont pas été soumis au même protocole de collecte. En effet, les enquêtés de l'échantillon de réserve ont été mis en production cinq mois après ceux de l'échantillon principal. Bien que joints par mail lorsqu'ils en avaient un, ils ont surtout été contactés par téléphone. Ainsi, parmi les 2 000

[4] Rosenbaum et Rubin (1983) insistent fortement sur cet aspect crucial de la méthode : un score de propension adapté est avant tout un outil d'équilibrage (un balancing score).[4]

répondants de la réserve, 92 % ont répondu sur ce mode. De plus, la réserve est constituée d'une population spécifique (diplômés de CAP ou de BAC professionnel). L'analyse des effets de mode portera donc sur l'échantillon principal compte-tenu de la composition spécifique de la réserve induisant de fait un biais de sélection lié au protocole.

Au total, l'échantillon utilisé pour l'analyse est constitué de 21 600 questionnaires dont 10 900 collectés par téléphone et 10 700 par internet (soit 50,4 % contre 49,6 %).

1.2.2 Sélection des variables pour le score de propension

Selon le mode choisi, le profil des répondants est assez différent : les plus diplômés, les personnes disposant d'un mail dans la base, les femmes, les personnes en emploi ont davantage choisi le mode de réponse internet. Il y a donc un effet de sélection important dans l'enquête Génération.

La modélisation du score de propension doit s'appuyer sur des caractéristiques observables, qui doivent être explicatives à la fois du fait de répondre à l'enquête sur le mode internet (processus de sélection) et des réponses apportées aux variables d'intérêt. Les variables choisies ne doivent pas être affectées par un biais de mesure. Les variables auxiliaires de la base de sondage sont de cet ordre, et il est également possible d'utiliser les variables de l'enquête réputées insensibles au mode [6].

La méthode d'appariement sur le score de propension repose sur l'hypothèse qu'il n'existe pas de facteur non observé qui influe à la fois sur la participation par internet et sur la valeur des variables d'intérêt.

Les variables initialement retenues de la base de sondage sont les suivantes :

- Genre
- Age
- Indicatrice d'une classe de sortie année terminale de formation
- Indicatrice d'une classe de sortie suivie en contrat de professionnalisation
- Type de baccalauréat réalisé
- Région du dernier établissement scolaire fréquenté
- Type d'établissement scolaire fréquenté[5]
- Région de résidence à la date de l'enquête
- Indicatrice de résidence en quartier prioritaire de la politique de la ville à la fin des études (QPV)
- Disponibilité d'un mail dans la base

Trois variables de l'enquête ont été initialement sélectionnées, réputées insensibles au mode :

- Mode de cohabitation à la date de l'enquête (habite chez les parents, seul, couple ou en colocation)
- Situation professionnelle à date d'enquête (emploi combiné avec la première position de la PCS, chômage, reprise d'études et autre situation)
- Plus haut diplôme obtenu (17 positions)

En pratique, la spécification du modèle est un processus itératif qui doit répondre à l'exigence d'un support commun suffisamment large pour un appariement de qualité.

[5] Le type d'établissement correspond à un découpage des établissements de formation (Université, écoles du supérieurs, lycées, ….)

Encadré 1 : Présence d'un mail dans la base de sondage et réponse sur internet
L'échantillon de l'enquête Génération 2017 se compose d'une liste d'élèves supposés avoir terminé leurs études en 2016-2017. Il est constitué de différents fichiers administratifs collectés auprès des établissements scolaires. Le taux de mails et de numéros de téléphone fournis est assez différent selon les sources. Par exemple, le fichier des contrats de professionnalisation ne fournit que des mails et la source « Agriculture » n'en fournit aucun.
Une des principales variables expliquant le choix du mode de réponse est la présence d'un mail dans la base de sondage. Pour pouvoir intégrer cette information dans la modélisation, il est nécessaire qu'elle soit corrélée à la fois avec le mode de réponse choisi et les variables d'intérêt de l'enquête. Ce constat a été vérifié dans d'autres enquêtes de la statistique publique. Cette hypothèse ne s'est pas révélée concluante dans l'enquête Génération. En conséquence, cette variable n'a pas été retenue dans la modélisation. En effet, la présence d'un mail est davantage liée à la constitution des fichiers des établissements qu'à un profil spécifique des enquêtés par rapport à l'usage d'internet.

Dans ce processus itératif, l'analyse des effets de mesure sur l'ensemble des variables a permis d'ajuster le modèle de régression en élargissant le champ des seules informations socio-démographiques du modèle initial. Ainsi, les variables de contrôle ont été enrichies de variables sensibles à la thématique de l'enquête. Par exemple, il s'est avéré en première lecture que le fait d'avoir redoublé en classes de primaire ou de collège présentait un effet de mesure. Celui-ci correspondait en réalité à un effet de sélection résiduel lors de la modélisation. En intégrant ces deux variables à la modélisation, d'autres effets de sélection résiduels ont disparu : réalisation de séjour à l'étranger, stage de fin d'études, etc. Le fait d'avoir redoublé signale un parcours scolaire plus difficile ce qui conduit à moins s'auto-saisir d'une enquête qui porte sur les parcours scolaires et professionnels.

Tableau 1 : Covariables constitutives du score de propension

Variables	DDL	Khi-2 de Wald	Pr > khi-2
PCS du père	7	175,9995	<0,0001
Redoublement au primaire	3	171,8265	<0,0001
Mode de cohabitation	5	153,6012	<0,0001
Sortant de contrat de professionnalisation	1	123,5313	<0,0001
Région à la date de l'enquête	19	108,4697	<0,0001
Type d'établissement de formation	12	106,1444	<0,0001
Genre	1	72,4746	<0,0001
Lieu de naissance de la mère	3	68,6647	<0,0001
Situation d'emploi combinée à la PCS	7	60,5909	<0,0001
Contrat à l'embauche de l'emploi à la date d'enquête	4	39,2155	<0,0001
Mention au bac	5	36,492	<0,0001
Région de l'établissement de formation	18	35,3801	0,0085
Genre * Plus haut diplôme obtenu (effet croisé)	16	34,1497	0,0052
Plus haut diplôme obtenu (17 positions)	16	28,4213	0,0281
Âge * Plus haut diplôme obtenu (effet croisé)	16	27,0714	0,0407
Redoublement au collège	2	24,9337	<0,0001
Type de bac réalisé	4	22,4368	0,0002
Classe de sortie année terminale	1	19,8611	<0,0001
Âge	1	13,3723	0,0003
Résidence en QPV	2	11,8258	0,0027
Classe suivie en apprentissage	1	8,4763	0,0036

La catégorie sociale du père est la variable la plus significative du modèle, les enfants de père « profession intermédiaire » répondant plus par internet et ceux de père « artisan-commerçant » moins souvent. Le fait de résider chez les parents, *toutes choses égales par ailleurs*, conduit à répondre davantage sur internet. Cela peut s'expliquer par la présence d'un ordinateur dans le foyer mais aussi un par meilleur taux de mail délivré. En effet, les adresses mail des parents détenues pour les sortants du secondaire sont plus stables que celle de jeunes avec des adresses mail d'écoles ou d'université. Le type d'établissement fréquenté en classe de sortie a aussi un effet significatif, il est corrélé avec la présence d'un mail dans la base. De même, suivre ses études en contrat de professionnalisation joue positivement dans la réponse à internet (aucun numéro de téléphone disponible pour ce profil). La variable région de résidence à la date de l'enquête ressort également, notamment en raison de la présence de résidents à l'étranger parmi les répondants.

1.3. Choix des paramètres du matching

Dans le contexte de cette enquête et dans la perspective d'un appariement par modalité, il a été nécessaire de déterminer une méthode d'appariement unique garantissant la robustesse des résultats. Les contraintes d'équilibrage choisies sont déterminantes dans l'évaluation des effets de mesure. Il s'agit de déterminer les paramètres concernant l'appariement : avec ou sans remise, exact et/ou usage d'un caliper, nombre de voisins. L'appariement avec ou sans remise permet de tester la possibilité d'utiliser un contrefactuel une ou plusieurs fois. Le caliper fait varier la distance entre les individus pour éviter d'apparier des individus trop éloignés. Augmenter le nombre de voisins[6] peut permettre éventuellement de réduire la variance (mais peut occasionner un biais dans l'estimation) [7]. Une contrainte de matching exact a été ajoutée, sur la variable de plus haut diplôme obtenu (principale variable d'intérêt de l'enquête). Elle a été définie selon quatre groupes : non diplômés, diplômés du secondaire, diplômés du supérieur court et du supérieur long[7].

Cinq matchings avec des paramètres différents ont été réalisés sur l'ensemble des modalités afin de choisir le mieux adapté aux spécificités de l'échantillon. Le tableau suivant résume les résultats des appariements réalisés sur la base Individus.

Tableau 2 : Description des tests d'équilibrage de l'appariement et indicateurs de validation

Paramètres	Matching 1	Matching 2	Matching 3	Matching 4	Matching 5
Avec remise	Non	Oui	Oui	Oui	Oui
Caliper[8]	0,15	0,15	0,15	0,10	0,10
Nombre de voisins	1	1	2	1	2

Indicateurs	Matching 1	Matching 2	Matching 3	Matching 4	Matching 5
Nombre de modalités avec propriété équilibrante non vérifiée	47	32	23	30	24
Nombre maximum de contrefactuels	1	40	61	36	51
Part de répondants internet appariés (en %)	65	99	98	99	97

Avant de lancer la procédure d'appariement, les individus ont été triés aléatoirement (tri implémenté dans la fonction « Match » de R). Il est notamment préconisé pour réaliser un matching sans remise (Matching 1). Cette option a été écartée rapidement car le support commun en était très réduit, aux

[6] Lorsqu'il existe un nombre important d'observations, il est parfois possible de sélectionner plusieurs individus dont la « proximité » avec l'individu traité est similaire. Le nombre de « voisins » à retenir repose alors sur un compromis classique biais-variance [4]

[7] Bac+5 et plus

[8] Il est coutume d'appliquer un caliper à 0,2. Compte tenu du bon ajustement observé, il a été réduit à 0,15 puis à 0,10. Un caliper définit à 0,01 a également été testé [2], sans impact significatif sur les résultats, mais abandonné du fait de la réduction du support commun - limite pour quelques modalités avec peu d'individus.

alentours des 65 %. Les matchings suivants s'améliorent nettement avec la méthode avec remise. Le support commun dépasse 95 % des individus traités pour l'ensemble des méthodes avec remise. Il ressort également qu'un caliper à 0,10 donne de meilleurs résultats qu'un caliper à 0,15. En effet, les individus appariés se ressemblent plus sans pour autant dégrader le support commun et tout en respectant la propriété équilibrante pour la majorité des variables introduites dans l'analyse. Le Matching 4 est celui retenu pour l'ensemble des traitements. Il a été préféré au Matching 5 afin de limiter l'introduction de biais dans les estimations. Il est important de noter qu'un socle commun de modalités avec un effet de mesure est systématiquement détecté quel que soit la méthode d'appariement utilisée.

1.4. Évaluation de l'équilibre des covariables entre les groupes traités et non traités

Les estimateurs basés sur le score de propension ont pour but de réduire le déséquilibre de distributions des covariables entre les individus traités et non traités. Deux hypothèses sous-jacentes sont à vérifier : l'indépendance conditionnelle à des caractéristiques observables X (biais de sélection *a priori* contrôlé) et la condition de support commun (les individus se ressemblent suffisamment pour que la comparaison ait du sens). Pour ce faire, on vérifie la qualité du score de propension : d'une part la propriété équilibrante et d'autre part la largeur du support commun.

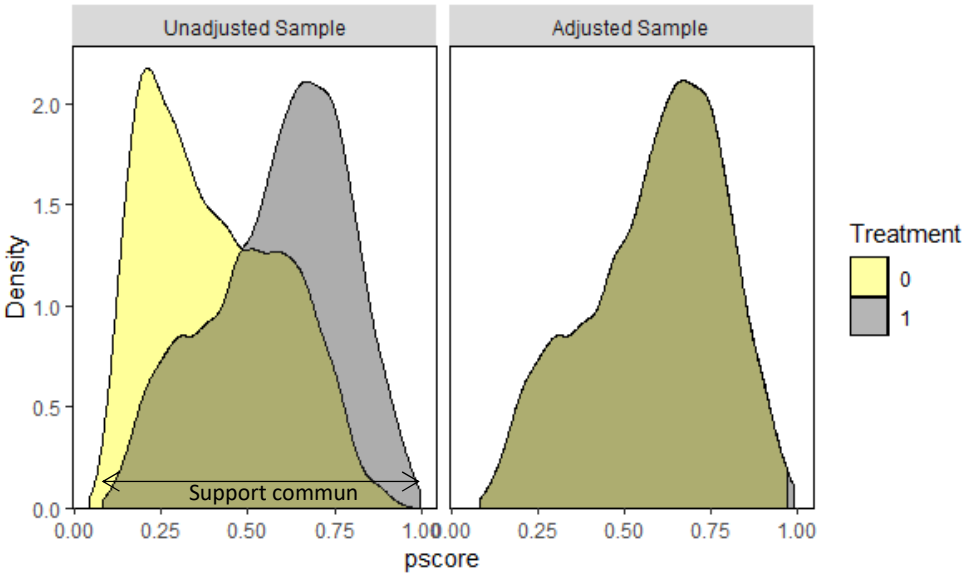
La question du support commun des distributions du score de propension conditionnel au traitement est essentielle dans l'analyse du matching. Son importance a été soulignée par Heckman, Ichimura et Todd qui ont montré qu'un support commun réduit constitue une source prépondérante de biais dans l'estimation de l'effet causal du traitement [5].

Tableau 3 : Répartition des scores de propension selon le mode de collecte

Mode	Moyenne	Minimum	Quartile inférieur	Médiane	Quartile supérieur	Maximum
Téléphone	0,404	0,041	0,236	0,375	0,562	0,972
Internet	0,590	0,080	0,458	0,619	0,739	0,999

En d'autres termes, on définit le support commun comme l'ensemble des individus ayant répondu par internet qui ont leur score de propension compris entre la valeur minimale du score des répondants par internet et la valeur maximale du score des répondants par téléphone. Ce support commun désigne donc les répondants internet pouvant trouver un contrefactuel ayant un score de propension similaire au leur dans le mode auquel ils n'ont pas répondu, le mode téléphone en l'occurrence [2]. Dans notre cas (tableau 3), l'intervalle est compris entre [0,080 et 0,972] traduisant une plage de scores communs satisfaisante.

Figure 1 : Support commun avant et après appariement

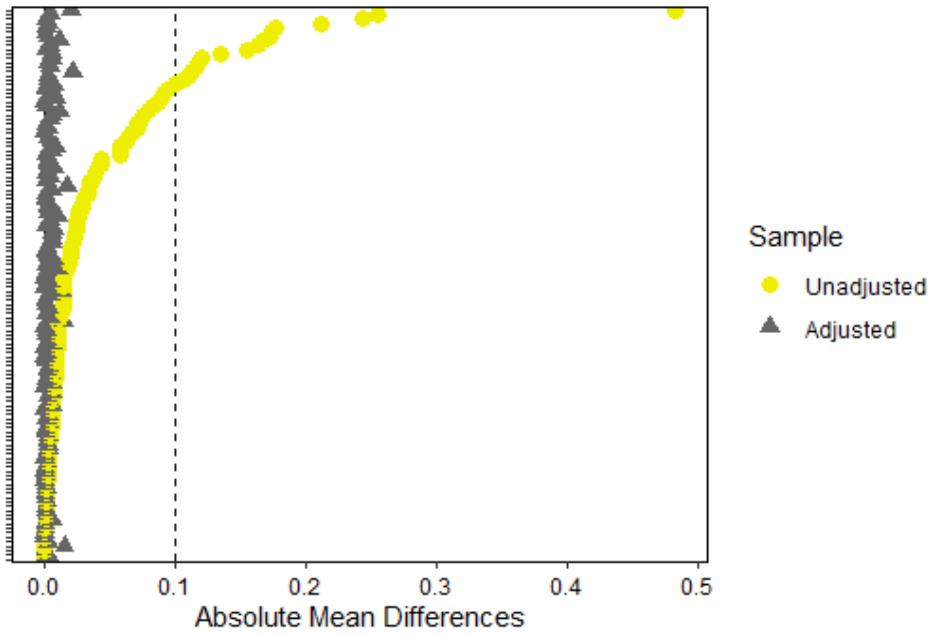


Les distributions du score de propension avant appariement (à gauche) sont assez étendues et symétriques. Ce qui laisse à penser qu'un nombre important d'individus du groupe traité peuvent être matchés avec des unités du groupe non traité dont les caractéristiques initiales sont similaires. Après appariement (à droite), on observe que la majorité des individus traités disposent d'un contrefactuel dont le score de propension est proche (14 individus exclus sur l'ensemble des traités).

1.4.1 Vérification de la propriété équilibrante

Une fois l'appariement réalisé, il convient de s'assurer que la propriété équilibrante du score de propension est vérifiée. On compare les distributions des covariables dans l'échantillon apparié des individus du groupe traité (répondants internet) et non-traité (répondants téléphone). Un seuil d'acceptabilité à 0,10 a été défini. L'objectif est ici de vérifier s'il reste des biais résiduels, auquel cas il faut réajuster le modèle pour améliorer l'estimation du score en introduisant ou en supprimant des covariables.

Figure 2 : Différences standardisées de moyennes avant et après appariement



On observe (Figure 2) un bon ajustement des covariables en deçà de 10 points d'écart entre les modalités du groupe traité et non traité (seuil fixé initialement). En comparant les distributions de chacune des variables constitutives du score, très peu d'écarts subsistent. Finalement, l'écart maximum observé est de moins de 2,5 points.

1.5. Mise en œuvre sous R

Les analyses et estimations ont été produites sous R avec le package *Matching* suivant les recommandations d'Abadie et Imbens [8]. Les procédures et fonctions utilisées pour estimer le score de propension sont inspirées du document de travail de S. Quantin [7].

1.5.1 Un matching par modalité (un exemple)

Afin d'illustrer l'implémentation de la méthode choisie, on s'appuie sur une variable qui était suspectée d'être concernée par un effet de mesure. Il s'agit d'une variable subjective de l'enquête qui décrit le degré d'optimisme quant à l'avenir professionnel. Information récurrente dans les enquêtes Génération, elle est posée à l'ensemble des individus et contribue à éclairer les perspectives professionnelles des jeunes.

PP020 : Comment voyez-vous votre avenir professionnel ? Vous êtes :
1 = Très inquiet
2 = Plutôt inquiet
3 = Plutôt optimiste
4 = Très optimiste

On retient la modalité 2 : « Plutôt inquiet » dans le traitement suivant. Dans le tableau 4, on observe un écart brut d'environ 5 points (0,052) entre les deux groupes sur cette modalité. Le contrôle de l'effet de sélection permettra de vérifier si cet écart est lié (tout ou partie) au mode de réponse utilisé.

Tableau 4 : Proportion d'individus selon le groupe pour la modalité plutôt inquiet

	Effectif	Proportion
Groupe de contrôle (téléphone)	2 255	0,207
Groupe traité (internet)	2 781	0,259

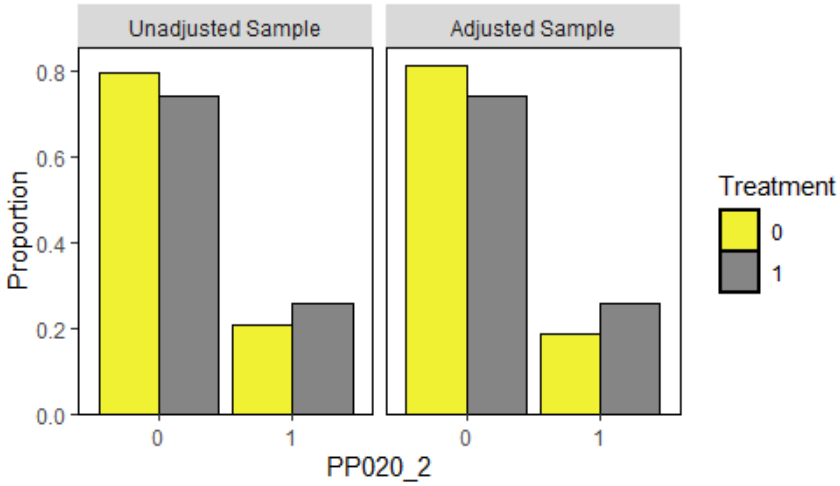
Après avoir délimité le champ des répondants à la question, on utilise le score de propension défini en partie 1.1.3. pour réaliser l'appariement. Le support commun est de 99,87 % (14 individus non appariés) et la propriété équilibrante est vérifiée (figure 2). Pour cette modalité, la p-value est significative (inférieure à 0,05) et indique la présence d'un effet de mesure malgré la correction du biais de sélection.

Tableau 5 : Résumé du matching

	Effectif
Estimate	0,072015
SE	0,0056807
T-stat	12,677
p.val	< 2.22e-16
Original number of observations	21 627
Original number of treated obs	10 734
Matched number of observations	10 720
Number of obs dropped by 'exact' or 'caliper'	14

Après ajustement via l'appariement des deux populations, on estime un effet de mesure de 0,072, creusant davantage l'écart brut de 0,052 initialement constaté. On en déduit donc un effet de sélection de -0,02 via la formule **l'effet de mode = effet de sélection + effet de mesure**. Cet effet de mesure, assez marqué, signifie qu'à profil égal les répondants par internet déclarent 7 points de plus être « plutôt inquiet » pour leur avenir professionnel par rapport à leurs homologues répondant par téléphone.

Figure 3 : Structure des répondants à la modalité 2 de la variable PP020 avant et après appariement



La figure 3 illustre le résultat du matching. L'ajustement de l'échantillon ne change pas la hiérarchie des réponses selon le mode mais fait varier les proportions des répondants.

1.5.2 Industrialisation

Les analyses des effets de mode doivent être ciblées sur les variables d'intérêt de l'enquête. Toutefois, pour cette première enquête Génération généralisée en multimode, une analyse sur l'ensemble des variables a été réalisée afin de n'omettre aucune information éventuellement soumise à effet. Ce travail a permis d'ajuster la modélisation du score de propension en captant des biais de sélection résiduels. Il permettra, pour les prochaines éditions de l'enquête Génération, de pointer une liste de variables à retravailler.

Compte tenu du nombre important de variables dans l'enquête, une procédure automatisée a été construite sur R. Elle met en œuvre un matching par modalité (hormis les variables continues).

Plusieurs étapes (programme en annexe 3) sont nécessaires pour obtenir les résultats permettant l'analyse et la validation du matching :

a) Sélection des variables de l'enquête, des variables de la modélisation et calcul du score de propension pour chaque individu

b) Pour chaque variable i : restriction au champ de la question (suppression des individus non-répondants à la question)

c) Dichotomisation de la variable i à analyser : on obtient j variables (j étant le nombre de modalités de la variable i)

d) Pour chaque modalité j : réalisation du matching et récupération du nombre maximum de contrefactuel

e) Calcul des différences standardisées pour vérifier la propriété équilibrante (test au seuil < 0,10)

f) Création d'un tableau contenant des indicateurs synthétiques du résultat des matching (export Excel possible)

Le tableau 6 est un extrait du tableau (f) en sortie. Est listé l'ensemble des matchings réalisés sur chacune des modalités présentes dans le fichier d'origine (une ligne par modalité).

Tableau 6 : Indicateurs synthétiques d'analyse des matching (extrait)

Modalités	Estimate (Effet de mesure)	Nombre d'observation	Nombre de répondant internet	Nombre de répondant appariés	Nombre d'individus exclus de l'appariement	Test de propriété équilibrante	Nombre d'individus matchés au maximum	P.value
Langue parlée P- Français	-0,09	3 761	1 595	1 593	2	VRAI	33	2,09E-08
Langue parlée P – Autre	0,02	3 761	1 595	1 593	2	VRAI	33	1,41E-01
Langue parlée P – Fr + Autre	0,05	3 761	1 595	1 593	2	VRAI	33	4,16E-05
Langue parlée P - NC	0,03	3 761	1 595	1 593	2	VRAI	33	2,78E-03
Langue parlée M – Fr	-0,12	3 563	1 534	1 524	10	VRAI	26	7,88E-13
Langue parlée M – Autre	0,02	3 563	1 534	1 524	10	VRAI	26	2,48E-01
Langue parlée M – Fr + Autre	0,08	3 563	1 534	1 524	10	VRAI	26	1,27E-07
Langue parlée M – NC	0,01	3 563	1 534	1 524	10	VRAI	26	4,57E-03
Discriminé - Oui	0,01	21 592	10 720	10 714	6	VRAI	31	4,63E-02
Discriminé – Non	-0,15	21 592	10 720	10 714	6	VRAI	31	2,20E-17
Discriminé - NSP	0,13	21 592	10 720	10 714	6	VRAI	31	2,20E-17
Discriminé - NVPD	0,01	21 592	10 720	10 714	6	VRAI	31	2,20E-17

Note de lecture : La modalité **Discriminé - Oui** (estimez-vous avoir été discriminé à l'embauche ; réponse oui) comporte un effet de mesure estimé de 0,1 points (Estimate=0,01). 10 720 répondants internet ont participé au matching et 6 n'ont pas été matchés. La propriété équilibrante est vérifiée pour les covariables utilisées pour le calcul du score de propension. Au maximum, un répondant téléphone a été matché avec 31 répondants internet. (La Pvalue (pval) est < à 0,05 ce qui signifie que l'effet de mesure est significativement différent de 0.

Il s'agit pour partie des données du résumé des matchings (exemple avec la figure 4). Il contient également une indicatrice de vérification de la propriété équilibrante *PropEquok* (TRUE si < 0,10) ainsi que le nombre maximum de contrefactuels par matching : *Maxmatch.*

À l'aide de ce tableau, les modalités détectées comme soumises à effet de mesure ont été extraites en sélectionnant celles avec une propriété équilibrante vérifiée (TRUE), une valeur de l'effet de mesure supérieure à 2 points (Estimate compris dans l'intervalle [- 0,02 ; + 0,02]) et significative (Pval < 0,05).

Cette première phase de décomposition de l'effet de mode a permis de distinguer les biais de sélection des biais de mesure induits par le multimode. Toutefois, de nombreuses questions subsistent malgré les choix qui ont été faits. En effet, il n'y a pas de consensus sur la méthode à adopter pour éliminer complètement les effets de sélection. Par la suite, une analyse détaillée des biais de mesure montrera les difficultés d'interprétation (plusieurs effets pouvant être combinés). Par ailleurs, des interrogations demeurent quant à l'écart acceptable entre les deux modes de réponses, détecté par l'effet de mesure. Ce seuil doit-il être le même pour toutes les variables ?

2. Effets de mesure dans l'enquête Génération 2017: diagnostic et solutions proposées

Une fois le travail de détection des variables soumises à effet de mesure effectué, un travail d'analyse a permis de faire un nouveau tri (notamment en décelant des effets de sélection résiduels) et de comprendre les mécanismes à l'œuvre dans l'apparition de ces effets.

2.1. Que revêtent les effets de mesure ?

L'effet de mesure correspond à l'effet de la modification du comportement de réponse selon le mode de collecte. Les modes de réponse disponibles (internet ou téléphone) induisent l'absence ou la présence d'un enquêteur. En découlent plusieurs sources possibles d'effet de mesure [9] :

— *L'aide apportée par les enquêteurs et l'ergonomie de l'outil* : les reclassements peuvent être proposés par l'enquêteur dans les cas où il détecterait des incohérences dans les réponses fournies par l'enquêté. À l'inverse, l'apport d'exemples par l'enquêteur pourrait aiguiller un enquêté vers certaines

modalités de réponse plus souvent que vers d'autres. Plus généralement, les difficultés de compréhension des concepts ou des modalités peuvent aussi être induites pas l'ergonomie du questionnaire et sont sources de biais de mesure.

— le manque d'intérêt, initial, ou accru par la longueur du questionnaire, conduirait l'enquêté à sélectionner prioritairement les premières modalités proposées sans lire les suivantes, de façon à atteindre plus rapidement la complétion du questionnaire (« *primacy effect »).* Plus globalement, la moindre implication de l'enquêté dans le remplissage de son questionnaire (« *satisficing »)* peut se traduire par des réponses arrondies, le choix des modalités médianes, ou encore l'abandon du questionnaire.

— *la désirabilité sociale* est définie comme la tendance qu'ont les individus à vouloir se présenter sous un jour favorable ou à chercher à se conformer à des attentes normatives [10][11]. Elle pourrait par exemple conduire un enquêté à ne pas avouer avoir renoncé à la recherche d'emploi ou encore à déclarer davantage de démarches de recherche d'emploi face à un enquêteur.

Encadré 2 : Test d'une méthode de matching sur les répondants internet (ordinateur versus smartphone/tablette)

L'enquête Génération propose la possibilité de répondre sur smartphone/tablette. L'ergonomie diffère dans l'affichage de certains types de questions (QCM par exemple) et pour le calendrier d'activité. L'ergonomie du calendrier répond aux exigences du smartphone (remplissage vertical) alors que pour la version ordinateur le remplissage est à l'horizontale. Un appariement a été mis en œuvre (critères choisis inchangés). Le groupe de contrôle est composé des répondants par ordinateur et le groupe des traités par les répondants smartphone/tablette. Aucun effet de mesure n'a été observé. Ainsi, on peut conclure qu'il n'y a pas de différence de remplissage entre les deux types de supports.

2.2. Quelles sont les variables soumises à effet de mesure ?

Dans un premier temps, 63 questions de l'enquête Génération 2017 sont identifiées comme concernées par un effet de mesure sur 250 questions testées. Plus précisément, elles comportent au moins une modalité présentant un effet significativement différent de zéro et supérieur à 2 points. L'écart maximum observé est de 22 points.

Tableau 7 : récapitulatif des modalités identifiées avec effets de mesure

Nombre de modalités	**Table individus**	**Table calendrier**
Propriété équilibrante non respectée	30	20
Ecart non significativement différent de 0	492	128
Ecart < 0,02 significativement différent de 0	260	61
Ecart > 0,02 significativement différent de 0	122	15
Total	904	224

Effet de sélection résiduel : Certains effets de sélection n'ont pu être totalement corrigés au moyen des variables socio-démographiques disponibles. Les jeunes ayant un profil de « réussite » scolaire et professionnelle semblent avoir davantage participé à l'enquête par internet. Ils sont plus intéressés par le sujet de l'enquête. Cette dernière portant sur le parcours scolaire et l'insertion professionnelle, une auto-sélection se ferait par internet pour les enquêtés qui ont eu un parcours scolaire ou d'insertion professionnel plus valorisant.

Ainsi, 30 variables sont identifiées comme provenant d'un effet de sélection résiduel : moins d'emploi durant les études, plus d'évolution de carrière, moins de contact avec Pôle emploi pour les répondants internet par exemple.

Après élimination de ces 30 variables portant un effet de sélection résiduel, **demeurent 33 variables considérées comme réellement touchées par l'effet de mesure** (exemples en Annexe 4). Les biais observés proviennent des effets classiques attendus de désirabilité sociale et de *satisficing* mais aussi de la construction du questionnaire lui-même. Les variables peuvent être classées en quatre catégories :

- Ergonomie et guidance : selon le mode de collecte, certaines questions sont reçues différemment bien que l'affichage à l'écran soit identique. La guidance dans le remplissage du calendrier d'activité demeure différenciée selon le mode malgré un affichage identique (sauf sous smartphone et tablette) du fait de l'aide des enquêteurs.
- *Satisficing* : Cet effet concerne seulement des variables avec une échelle de Lickert
- Désirabilité sociale : cet effet est détecté sur quelques variables subjectives du questionnaire (peu de questions de ce type dans l'enquête).

Tableau 8 : Répartition des modalités impactées selon l'effet de mesure principal

TYPE D'EFFET PRINCIPAL	Nombre de variables
ERGONOMIE - GUIDANCE	**17**
Affichage des questions à choix multiples	5
Consigne « Attendre la réponse spontanée »	6
Proposition d'une modalité « Autre »	2
Difficulté de positionnement dans le calendrier	1
Complexité des cas d'intérim	2
Recherche dans le répertoire SIRENE	1
SATISFICING	**7**
Questions avec échelle de Lickert	4
Questions avec échelle de Lickert avec position médiane	3
DÉSIRABILITE	***9***
Total	**33**

2.2.1 Effet de l'ergonomie et guidance

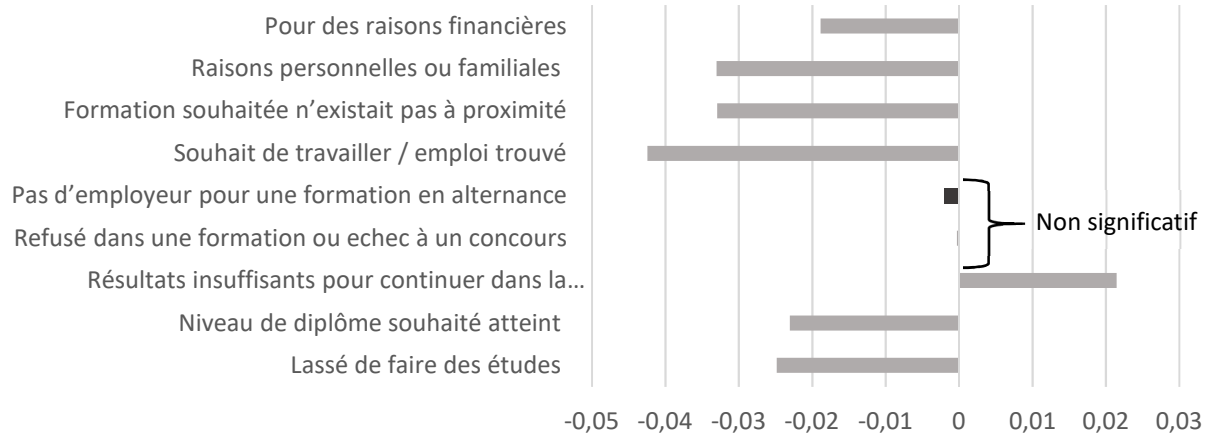
Les cas d'effets de mesure potentiellement en lien avec l'ergonomie du questionnaire sont les suivants : des questions à choix multiples, des questions pour lesquelles les modalités n'ont pas été listées par l'enquêteur, et des questions présentant une modalité « autre ».

- *Questions à choix multiples (QCM)* :

Les questions QCM ne sont pas toutes concernées par un effet de mesure, le cas se présente surtout en cas de présence de modalités subjectives ou de modalités qui peuvent recouvrir des intersections.

Ce type de questions est présenté sous forme de tableau, présentant les modalités en ligne et deux colonnes « Oui », « Non » dans lesquelles il est obligatoire de se positionner. Par téléphone, l'enquêteur propose une modalité et attend une réponse « oui » ou « non » avant de passer à la modalité suivante. Par internet, le répondant voit directement le tableau avec l'ensemble des modalités proposées. Ainsi, il peut réaliser une sélection des modalités pour lesquelles il va cocher « oui ». Par exemple, à la question « Raisons d'arrêt des études », les répondants téléphone déclarent en moyenne 3 modalités contre 2 par internet. Toutefois, la modalité « Vos résultats étaient insuffisants » présente un effet particulier : elle est davantage sélectionnée par internet, ce qui témoigne plutôt d'un effet de désirabilité sociale.

Figure 5 : Pour quelles raisons avez-vous arrêté vos études ? Effet de mesure sur les modalités « oui » (en points)

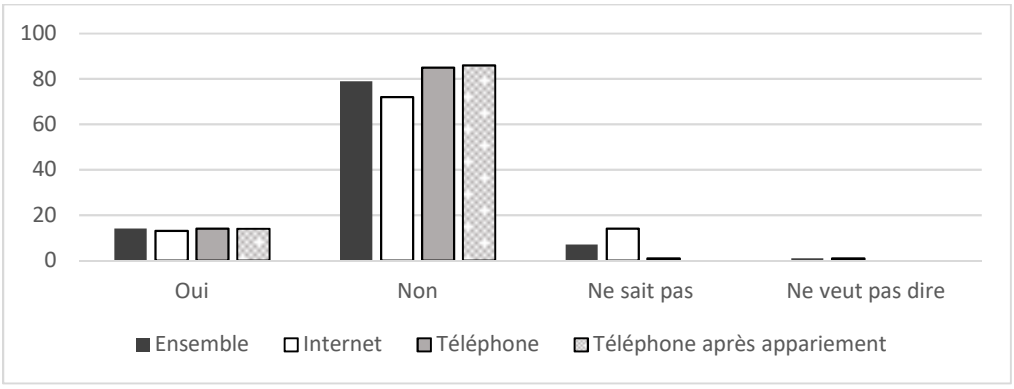


Note de lecture : Les écarts de réponses après appariement sont présentés. Certaines modalités « oui » présentent un effet de mesure. A profil égal, sur internet, les répondants déclarent moins souvent oui à la plupart des modalités proposées

- Questions avec présence de la consigne « Attendre la réponse spontanée » sur le mode téléphone :

Avec un questionnaire dépassant les 30 minutes en moyenne, la consigne *« Attendre la réponse spontanée »* a été intégrée en cours de production, sur certaines questions, pour limiter la durée de passation par téléphone. Elle est proposée également pour éviter à l'enquêteur de citer des modalités telles que « Ne sait pas » ou « Ne veut pas dire ». Par internet, ces modalités apparaissent à l'écran. Il en résulte des réponses systématiquement différentes selon le mode. Par exemple, à la question « Avez-vous été victime de discrimination à l'embauche ? », on observe une sur-sélection de la modalité « Ne sait pas » par internet. Par téléphone, ce type de répondants (indécis, mais qui ignoraient la possibilité de ne pas trancher) se positionne plutôt sur la modalité « non ». En revanche, la modalité « Oui » ne présente aucun effet.

Figure 6 : Déclarations sur le thème de la discrimination selon le mode (en %)



Note de lecture : Les répondants par internet sont moins nombreux à répondre « non » et plus nombreux à répondre « ne sait pas » que les répondants par téléphone. En gommant l'effet de sélection, c'est-à-dire en comparant les répondants téléphone après appariement avec les répondants internet, la différence est toujours présente, il y a donc bien un effet de mesure sur cette variable.

- Questions avec présence de la modalité « Autre » :

Un effet de mesure peut être observé sur cette modalité. Selon le degré de complexité de la question, l'effet se traduit par pour plus ou moins de déclarations « Autre ». Par exemple, au niveau de la question sur la série du bac technologique, avec des sigles complexes, l'enquêteur a plus souvent indiqué en clair le type de bac.

- Un positionnement parfois délicat dans le calendrier d'activité :

Quatre types de situations sont proposés pour décrire le parcours depuis la fin des études[9], dont un état « Autre situation ». Si l'enquêté déclare une séquence de ce type, il doit ensuite préciser sa situation en choisissant dans une liste de propositions. Une saisie en clair est également possible si l'individu ne parvient pas à se positionner. Sur le mode internet, on constate une sur-sélection de « autre situation » et une sur-déclaration via l'ouvert. Ceci peut s'expliquer par le fait que par téléphone, l'enquêteur peut orienter le répondant vers le type de situation correspondant le mieux à sa déclaration (exemple : le chômage partiel est à rattacher à une situation d'emploi). Ainsi, par internet, il y a davantage d'enquêtés qui se sont inclus dans la catégorie « Autre » de manière incorrecte du point de vue des consignes de remplissage du calendrier.

- Situations d'intérim et subtilité du remplissage attendu :

Par internet, les répondants intérimaires ont déclaré des séquences d'emploi avec moins de missions et moins d'entreprises différentes. La consigne du calendrier indique de regrouper les emplois d'intérim en une seule période d'emploi, en indiquant le nombre d'entreprises et de missions. Cette consigne n'a pas forcément été respectée par internet, alors que les enquêteurs ont été sensibilisés à ces cas lors de la formation qui leur a été dispensée. De plus, la présence de plusieurs missions et plusieurs entreprises complexifie le remplissage ce qui a pu générer des abandons en autoadministré. Par conséquent, les parcours complexes sont moins souvent décrits par internet.

- Utilisation d'un webservice pour réaliser des recherches dans la base SIRENE (à la volée) :

Pour chaque emploi décrit, les entreprises employeuses sont recherchées dans la base SIRENE. Par internet, les répondants ont davantage réussi à trouver leur entreprise. Ainsi, un effet est détecté seulement sur la modalité « non codé » qui résulte d'un effet enquêteur/outil : l'enquêteur a pu commettre des erreurs de saisie du nom de l'entreprise ou ne pas avoir reconnu l'entreprise dans la liste du menu (par exemple lorsque l'enquêté a déclaré un sigle).

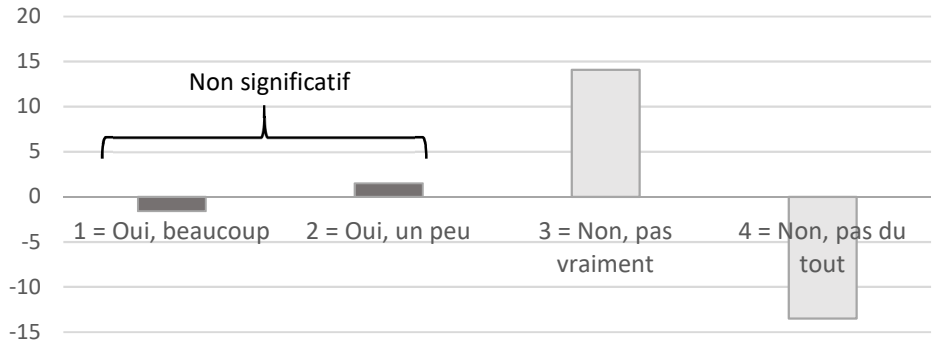
2.2.2 Satisficing

La troisième catégorie de variables concernées par un effet de mesure est constituée de variables avec échelle de Lickert. Ce modèle de questions permet de mesurer la perception d'un individu sur un sujet donné. L'échelle se décline en nombre pair ou avec une modalité médiane. Les réponses sur internet se portent plus souvent sur les modalités intermédiaires (exemple : « Plutôt oui » ou « Plutôt non »). Par téléphone, les réponses sont plus tranchées. On peut apparenter ce type de comportement à du *satisficing*, au sens où le répondant internet fournit ici moins d'effort pour répondre à la question. Le fait de visualiser l'échelle à l'écran peut le tenter de donner une réponse « moyenne », ce qui lui demande moins de temps de réflexion. Ce type d'effet de mesure a été détecté sur l'ensemble des questions à échelle (sauf une, dont l'interprétation est différente, partie 2.2.3.), néanmoins peu nombreuses dans le questionnaire.

[9] Etats proposés dans le calendrier d'activité : Emploi, recherche d'emploi, formation ou reprise d'études, autre situation

Figure 7: Diriez-vous que cet emploi a perturbé le cours normal de vos études ? Effet de mesure

(en points)

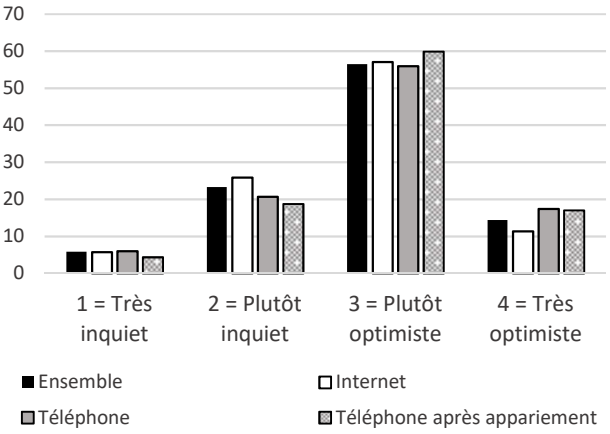


Note de lecture : Les écarts de réponse après appariement sont présentés. Les modalités « non, pas vraiment » et « non, pas du tout » à la question présentent un effet de mesure. A profil égal, sur internet, les répondants déclarent 14 points de plus « non, pas vraiment » et 14 points de moins « non, pas du tout ».

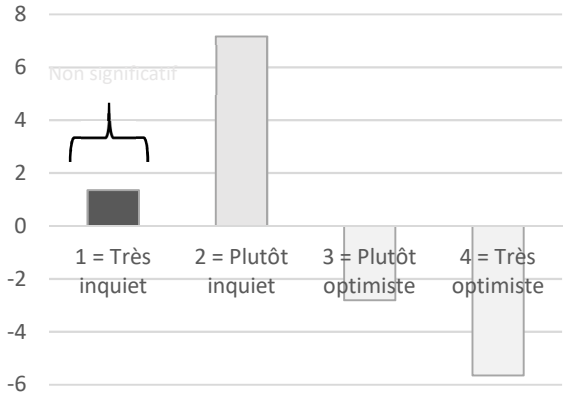
2.2.3 Désirabilité sociale

Pour illustrer le phénomène de désirabilité sociale, un cas de question subjective est celle concernant l'optimisme quant à l'avenir professionnel (partie 1.4.1.). Construite à partir d'une échelle paire, elle présente un effet de mesure dont l'interprétation diffère des autres variables à échelles. On déduit plutôt un effet de désirabilité sociale qu'un effet *satisficing*. En effet, les répondants sont moins positifs sur le mode internet que par téléphone, car ils chercheraient plus souvent à donner une bonne image lorsqu'ils s'adressent à un enquêteur.

Figure 8 : Déclarations à la question sur l'optimisme professionnel selon le mode (en %)



effets de mesure (en points)



Un autre type de questions sur lesquelles se manifeste l'effet de mesure de type désirabilité sociale peut être illustré avec la question posée sur l'utilité des intermédiaires du marché du travail (Pôle emploi ou mission locale). Il s'agit d'une question sur les bénéfices retirés de l'inscription à Pôle emploi ou à une mission locale. Si par internet, les répondants déclarent plus souvent que Pôle emploi ou la mission locale les ont aidés, par téléphone, c'est l'inverse qui se produit. Est-il plus attendu d'exprimer des avis négatifs sur Pôle emploi ?

Encadré 3 : Mesure des effets sur l'ensemble des répondants (principal + réserve)

L'échantillon de réserve a été intégré aux analyses afin de se prémunir de l'oubli d'un éventuel biais. Un score de propension a été imputé via la modélisation de l'étape 1.1.2., il correspond à ce qui aurait été observé si ces individus avaient été soumis au même protocole que les répondants de l'échantillon principal. Un appariement a été réalisé en fusionnant les répondants des deux échantillons. Après comparaison des effets détectés selon les deux groupes analysés (échantillon principal uniquement et échantillon principal avec réserve), on conclut sur les mêmes résultats. Les variables sujettes à l'effet de mesure sont identiques. Seuls les écarts au sein des modalités diffèrent sensiblement.

2.3. Les solutions d'ajustement et les préconisations d'utilisation proposées

Le travail faisant suite à cette phase d'estimation et d'analyse des effets de mesure consiste à réfléchir à la nécessité et la possibilité d'apporter des solutions pour pouvoir analyser les données. Face à une variable impactée par l'effet de mesure, quelles sont les précautions à prendre quant à son utilisation ? Selon la nature du biais et l'objectif visé, il faut déterminer des préconisations et les éventuelles corrections à opérer sur les données.

À ce stade des analyses, plusieurs pistes sont envisagées sans que les choix ne soient encore complètement arrêtés.

- *Des mises à jour dans les variables*

Une des premières solutions envisagées est le regroupement de modalités à diffuser. Certaines variables de QCM ou des questions avec échelle permettent un regroupement de modalités. Cette action permettra de minimiser les écarts dûs à l'effet de mesure.

D'autres variables pourront être recodées notamment pour les questions liées à des difficultés de remplissage du calendrier d'activité. Les questionnaires des prochaines enquêtes Génération devront être modifiés afin de limiter les effets dans les futures données collectées.

- *Aucune action*

Il s'agit dans ce cas d'agréger les données collectées sans opérer de correction sur l'un ou l'autre des échantillons (téléphone ou internet). Ce cas concerne les situations où le biais de mesure est jugé d'ampleur limitée. Il s'agit des questions avec échelle et une modalité centrale, de certaines questions « attente de la réponse spontanée » avec la modalité « Ne sait pas » dont les effets de mesure sont jugés peu importants (<5 points). Cela concerne aussi les variables liées à la sélection des entreprises dans la base SIRENE.

- *Correction de l'effet de mesure*
    - *Corriger par imputation les données provenant de l'un ou l'autre des modes*

Cette opération consiste à imputer des réponses des individus du mode alternatif. Elle peut se faire soit sur l'ensemble des répondants du mode alternatif, soit sur uniquement les individus « porteurs » de l'effet de mesure [6]. La dernière méthode sera retenue[10]. Il s'agira alors de choisir un mode de « référence ». Lors de la création de la base de données comparables avec les enquêtes antérieures, le mode de référence sera nécessairement le mode de collecte historique (téléphone). Néanmoins, le passage au multimode ne doit pas conduire à chercher à tout prix à corriger les effets de mode par rapport au mode historique lorsque ces effets sont censés être bénéfiques pour la qualité de l'enquête. En effet, les principales variables considérées comme devant être imputées ont subi un effet désirabilité sociale et, dans ce cadre, les données collectées sur internet sont considérées comme plus fiables. Dans ce cas, l'enquête multimode est biaisée, mais potentiellement moins que dans les enquêtes monomodes antérieures.

[10] Il s'agit d'une méthode conçue par Stéphane Legleye [12]. Elle vise à repérer des individus ayant des caractéristiques sociodémographiques communes mais un comportement de réponse différent dans chacun des deux modes et d'imputation des réponses sur les individus du mode alternatif porteur de l'effet de mesure.

- *Repondération par calage*

Une première solution de correction par calage consiste à modifier le poids des répondants afin que l'estimation globale de la variable soit conforme à ce qu'on aurait obtenu après correction du biais de mesure. Les réponses des individus ne sont pas modifiées. Cette solution suppose un jeu de poids par variable pour corriger l'effet de mesure. À ce stade, cette solution n'est pas envisagée pour les données de Génération.

Une deuxième méthode consisterait en un jeu de poids unique : un calage est réalisé sur une distribution figée des deux modes, pour toutes les enquêtes. Le biais de mesure n'est pas corrigé mais il est maintenu au même niveau d'une enquête à l'autre. Avec l'hypothèse que le biais de mesure ne varie pas dans le temps, il rend comparables les données pour des enquêtes récurrentes. Cette technique est surtout adaptée lorsque les parts des modes de réponse évoluent peu dans le temps. Il faudrait déterminer si cette méthode est applicable sur les prochaines enquêtes, en fonction de l'évolution de la répartition des répondants sur les deux modes qui sera observée.

- *Utilisation de la modélisation avec le mode de collecte*

Dans le cas d'une agrégation simple des données, contrôler la variable mode (et plutôt en interaction) dans les modélisations effectuées s'avère être une solution. La modélisation sans précaution d'une variable présentant un effet de mesure peut faire apparaître des liens fictifs qui résultent uniquement d'un lien avec le mode de collecte. Pour pallier cet écueil, l'ajout de la variable de mode de collecte permet de capter l'effet de mesure. Ajouter son interaction avec les autres variables du modèle permet de tenir compte d'un éventuel effet de mesure différencié selon les modalités de ces variables, mais ne semble pas changer l'interprétation globale de ces variables [2].

Quelles que soient les solutions adoptées pour agréger les données, des informations seront fournies aux utilisateurs des données de l'enquête afin de les mettre en garde et de les accompagner. Dans le dictionnaire des variables, il est prévu d'afficher une indicatrice d'effet de mesure sur les variables concernées par l'effet de mesure (> seuil de 5 points) et le redressement éventuel opéré. Les chargés d'études seront sensibilisés à l'intérêt de contrôler le mode de collecte dans des modèles lorsqu'ils utilisent ces variables (hors variables imputées).

Les nouvelles variables calculées (recodification ou imputation) seront diffusées en plus des variables brutes en justifiant leur intérêt pour produire des statistiques descriptives ou des modélisations.

L'analyse des effets de mode de l'enquête Génération 2017 a permis de montrer la complexité dans le choix et la mise en application de méthodes d'appariement. Selon le protocole de collecte de l'enquête et la spécificité de l'échantillon d'individus à traiter, le choix de la méthode d'évaluation peut varier et peser sur la robustesse des résultats. Pour aller plus loin, un comparatif entre deux méthodes d'analyse des biais a été réalisé à partir de l'enquête Génération « Covid et après ? ». L'objectif est de mesurer l'efficience d'un échantillon de contrôle dans un protocole de collecte multimode.

3. Utilisation de deux méthodes d'évaluation pour l'enquête Génération, « Covid et après ?»

L'enquête Génération, « Covid et après ? » réinterroge les 8 800 répondants de l'enquête 2017 auprès de la Génération 2010 sur les reconversions liées à la crise sanitaire. Il s'agit d'une cohorte fidélisée ayant déjà répondu à trois interrogations à 3, 5 et 7 ans après la sortie du système éducatif. Cette nouvelle enquête, ne relevant pas du dispositif Génération standard et d'une durée de 12 minutes en moyenne, met en œuvre un protocole à la fois séquentiel et concurrentiel (annexe 1).

Deux échantillons ont été exploités en parallèle. Un échantillon classique multimode d'environ 7 000 individus, ayant abouti à un taux de collecte de 59 % (soit 4 150 individus) avec 1 752 répondants téléphone (42 %) et 2 398 répondants par internet (58 %). Parallèlement, un échantillon embarqué, avec affectation aléatoire du mode de collecte a été mis en place afin de mesurer les éventuels biais liés au mode de collecte. Il est construit par tirage aléatoire parmi les 8 800 individus disponibles, hormis les individus non diplômés, trop peu nombreux du fait d'une attrition plus importante sur cette

catégorie de jeunes[11]. Constitué de 1 800 individus, il représente 20 % de la population à interroger. Les deux sous-échantillons monomodes qui composent l'échantillon embarqué sont composés chacun d'environ 900 individus. 340 individus ont répondu par internet et 397 par téléphone[12]. Les effets de mesure ont été analysés sur l'échantillon de contrôle en utilisant une méthode d'ajustement par les pondérations [13].

Pour évaluer les résultats obtenus, une analyse sur les effets de mode a été effectuée sur l'échantillon multimode (hors répondants de l'échantillon embarqué) dont le protocole de collecte offre la possibilité de choisir le mode de réponse. Une méthode de matching (mêmes contraintes d'équilibrage) telle qu'utilisée dans l'enquête Génération 2017 sera mise en œuvre pour étudier les effets de mode de cet échantillon.

Encadré 4 : Méthode de pondération inversée sur un échantillon embarqué

La méthode de pondération inversée repose sur l'hypothèse qu'il n'existe aucun biais de composition inobservable entre les répondants aux différents modes. Pour être au plus proche de ces hypothèses, une première approche consiste à disposer d'un échantillon de contrôle disjoint [6]. Le tirage aléatoire a permis de construire des échantillons comparables sur des caractéristiques (y compris sur des caractéristiques inobservées dans l'enquête) qui auraient influé sur le choix du mode de réponse à l'enquête s'il avait été proposé. L'objectif est de contrôler le biais de sélection afin de faciliter l'estimation de l'effet de mesure.

Pour étudier les effets de mesure sur l'échantillon de contrôle, une méthode de repondération par l'inverse du score de propension a été utilisée. Considérant la faible taille de l'échantillon, cette méthode a l'avantage par rapport à la méthode par appariement de ne pas écarter d'individus de l'analyse. De plus, du fait de l'imposition du choix du mode, le profil des répondants est relativement proche (figure 11), ce qui suggère d'utiliser une méthode par régression ou par repondération. En définitive, cette dernière méthode est choisie car elle fournit un estimateur sans biais de l'effet de mesure.

Pour calculer l'effet moyen du traitement sur les traités (ATT), la structure des individus répondant par téléphone est rendue comparable à celle des individus du mode internet en repondérant les individus du mode téléphone. Les poids appliqués aux individus sont :

$$w_i^{ATT} = T_i + \frac{(1-T_i)\hat{p}(X_i)}{1-\hat{p}(X_i)} = \begin{cases} 1\ si\ T_i = 1, par\ internet \\ \dfrac{\hat{p}(X_i)}{1-\hat{p}(X_i)}\ si\ T_i = 0, par\ téléphone \end{cases}$$

Avec T_i le traitement et $\hat{p}(X_i)$ le score de propension.

Les biais de sélection ainsi réduits, les données pondérées vont pouvoir servir à estimer l'effet moyen du traitement sur les traités. L'ATT est obtenu par régression pondérée (par w_i^{ATT}) sur le mode de collecte uniquement[13] (estimateur sans biais de l'effet de mesure).

3.1. Modélisations du score de propension à partir de deux échantillons spécifiques

Dans la méthode de pondération inversée, comme pour la méthode d'appariement, une première étape consiste à calculer un score de propension (partie 1.1.) dans un processus itératif à la suite de l'analyse des bais de sélection résiduels. Dans l'enquête multimode, la probabilité de répondre par internet est modélisée. Le principe est d'intégrer dans le modèle de régression des variables qui ne sont pas affectées par un biais de mesure et qui sont explicatives à la fois du fait de répondre par

[11] Ces profils appartiennent uniquement à l'échantillon multimode afin de maximiser le taux de réponse.

[12] Soit un taux de réponse de 38% pour l'échantillon internet et 44% pour l'échantillon téléphone

[13] D'autres facteurs de confusion peuvent être inclus dans la régression. Ce n'est pas adapté aux données de l'échantillon embarqué car il faudrait spécifier un modèle pour chacune des 104 variables étudiées.

internet et des réponses apportées aux variables d'intérêt de l'enquête.
Concernant l'échantillon de contrôle, les répondants des deux sous-échantillons monomodes internet et téléphone ont été fusionnés pour l'estimation du score (probabilité de répondre par internet). Bien que l'appartenance à un échantillon soit aléatoire, le fait de répondre ne l'est pas, car selon le groupe auquel l'individu a été assigné, la propension de répondre à l'enquête n'est pas la même. Un individu non-répondant de l'échantillon téléphone aurait pu choisir de répondre s'il avait appartenu à l'échantillon internet. Ces échantillons monomodes permettent de contenir la sélection en n'offrant pas le choix du mode de réponse, mais un effet de sélection subsiste et se traduit par le fait de répondre ou non. L'effet de sélection dans les échantillons monomodes s'apparente donc à de la non-réponse. Cette non-réponse est différente selon le mode proposé. Cela influe sur la structure des répondants par mode : le modèle estimé sur l'échantillon de contrôle captera ainsi l'effet des caractéristiques individuelles sur le fait d'être répondant téléphone ou internet.

Tableau 9 : Covariables utilisées pour le calcul du score de propension

Modèle échantillon embarqué	Modèle échantillon multimode
Issues de la base de sondage : - Genre - **Plus haut diplôme obtenu (21 modalités,** hors non diplômé**)** - **Mode de cohabitation lors de la première enquête en 2013 (parents, en couple ou seul)** - PCS du père Collectées en cours d'enquête : - Situation professionnelle en 2021 croisée avec la PCS - Typologie de trajectoire en 2020 - Contrat de travail en 2021 - Avoir engagé une démarche de réorientation depuis la dernière enquête en 2017	Issues de la base de sondage : - **Genre** - **Plus haut diplôme obtenu (22 modalités)** - **Contrat de travail en 2017** - **Statut dans l'emploi en 2017** - **PCS du père** - Pays de naissance du père Collectées en cours d'enquête : - **Avoir engagé une démarche de réorientation depuis la dernière enquête en 2017** - **Région d'habitation en 2021** - PCS en 2020 - Avoir des enfants au 1er mars 2020 - Télétravail avant mars 2020 - **Typologie de trajectoire en 2020**

Note de lecture : Les variables en gras sont significatives dans la modélisation (régression logistique Stepwise)

Les informations intégrées dans le modèle du score de propension pour l'échantillon multimode sont plus nombreuses et différentes de celles du modèle pour l'échantillon embarqué. Moins de variables sont significatives pour les échantillons monomodes. Le biais de sélection y est mieux maitrisé.

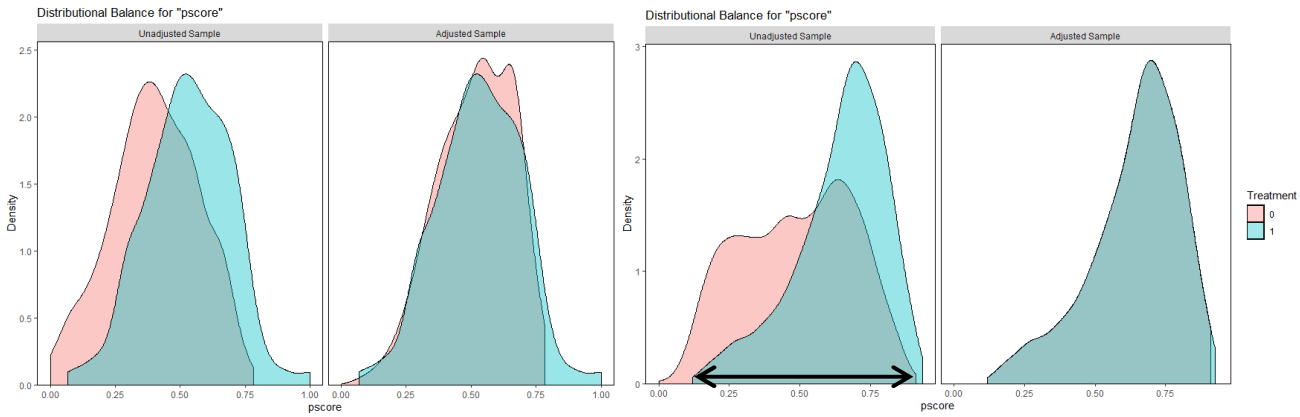
Les deux modélisations révèlent que les diplômés d'un bac professionnel ou technologique, d'un CAP/BEP industriels, *toutes choses égales par ailleurs,* répondent moins souvent par internet. Pour l'échantillon multimode, du fait du choix du mode d'autres variables ressortent, par exemple les femmes et les personnes en emploi salarié en 2017 choisissent davantage le mode de collecte internet.

Les différences observées sur la significativité des covariables, peuvent s'expliquer par une sélection maitrisée mais aussi par la composition de l'échantillon embarqué (absence des non-diplômés) et par la petite taille de l'échantillon.

3.1.1 Distributions du score de propension

La représentation graphique des distributions du score de propension dans les groupes traité (internet) et non traité (téléphone) permet de comparer la densité du score avant et après ajustement et de visualiser le support commun pour la méthode par appariement (partie 1.3.1).

Figure 9 : Distributions des scores de propension avant et après ajustement



Par pondération (échantillon embarqué) *Par appariement (échantillon multimode)*

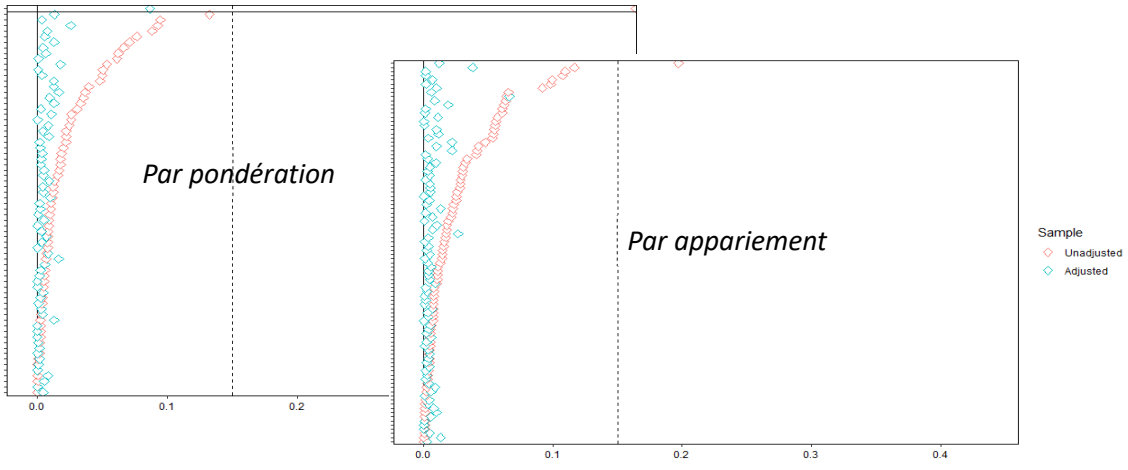
Selon l'échantillon utilisé, l'ajustement est sensiblement différent. En effet, avec l'échantillon embarqué, on observe une distribution du score de propension plus proche entre les deux modes qu'avec l'échantillon multimode. Avec l'échantillon embarqué, les distributions du score selon le mode de réponse sont symétriques et proches l'une de l'autre. Afin de pouvoir réaliser la pondération, les scores de propension des répondants par téléphone doivent être strictement inférieurs à 1, ce qui est bien le cas ici. Avec l'échantillon multimode, les différences de densité du score illustrent les spécificités des répondants selon le mode de collecte. Le graphique des répondants téléphone de l'échantillon multimode semble tronqué et plus étalé. Le protocole de l'enquête téléphonique n'a pas pu être mené à son terme. Le nombre de répondants par téléphone est plus faible, il manque ainsi les profils spécifiques répondant plus difficilement.

Dans les deux cas de figure, le support commun est acceptable et permet pour la majorité des répondants internet de trouver un contrefactuel (seuls 32 individus se trouvent hors de l'intervalle du support pour l'échantillon multimode).

3.1.2 Propriété équilibrante

La propriété équilibrante du score de propension est ensuite vérifiée en comparant les différences de moyennes standardisées des variables du modèle du score de propension avant et après appariement. Les variables introduites dans les deux modélisations ne sont pas similaires. Toutefois, quelle que soit la méthode utilisée, on observe un bon ajustement des covariables (en dessous du seuil de 0,15) produisant des groupes comparables pour les analyses.

Figure 10 : Différences standardisées de moyennes avant et après ajustement



3.2. Estimation et détection des effets de mesure

Pour les deux échantillons, l'effet de traitement est estimé par la différence moyenne des résultats entre les individus traités et non-traités (ATT).

Avec la méthode de pondération inversée, il faut déterminer à l'aide d'un intervalle de confiance si l'effet de mesure s'avère significativement différent de zéro. La méthode de matching, implémentée avec le logiciel R, permet, après contrôle des biais de sélection, d'estimer les effets de mesure et fournit une valeur de test pour déterminer la significativité des écarts observés (encadré 5). Dès lors, on s'appuie sur un tableau de synthèse contenant les résultats du matching (figure 7) pour déterminer les modalités soumises à effet de mesure.

Encadré 5 : Calcul des intervalles de confiance sur les variables de l'échantillon embarqué (bootstrap)

Après pondération, on calcule l'effet du traitement sur les traités par estimation de la différence moyenne entre les groupes. Pour étudier la significativité de l'effet estimé [7], on construit des intervalles de confiance par la méthode du bootstrap [14][15] pour chacune des modalités. On utilise le package *boot* de R : on va réestimer sur données pondérées, l'effet moyen du traitement pour chaque modalité de chaque variable pour un grand nombre d'échantillons tirés avec remise parmi les données initiales. On choisit de réitérer les estimations 2 000 fois. La fonction *boot.ci* du package permet d'obtenir cinq intervalles de confiance [14] (intervalle de confiance normal, studentisé, percentile, etc.).

Pour décrire le résultat du bootstrap, la modalité « Améliorer vos conditions d'emploi (rémunération par exemple) » de la variable « Qu'est-ce qui a motivé ce projet ? » a été utilisée.

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = bootT.outt)

Intervals :
Level      Normal              Basic             Studentized
95%   (-0.3513, -0.0202 )   (-0.3529, -0.0245 )   (-0.3509, -0.0199 )

Level     Percentile            BCa
95%   (-0.3499, -0.0215 )   (-0.3392, -0.0115 )
Calculations and Intervals on Original Scale
```

Cette modalité présente un effet de mesure de 18 points, significativement différent de zéro après calcul des intervalles de confiance. L'effet de mesure est donc avéré quelle que soit la méthode de calcul de l'intervalle. Toutefois, on constate que les intervalles de confiance sont assez larges donc peu précis (probablement lié à la taille de l'échantillon embarqué). Les cinq intervalles de confiance sont reproduits pour toutes les modalités analysées. Il en résulte que pour la majorité des variables traitées, les écarts mesurés ne sont pas significativement différents de zéro. Au total, 6 variables (soit 6 % des variables étudiées) sont impactées par l'effet de mesure selon les intervalles de confiance normal, percentile, basic et student.

[14] L'intervalle normal fait l'hypothèse que la distribution de l'estimateur de l'effet de mesure suit une distribution normale ; l'intervalle de confiance studentisé dépend des percentiles de la distribution de la statistique du test de Student des échantillons ; les méthodes percentile, basic (Reverse Bootstrap Percentile) et BCa (Bias Corrected and accelerated) permettent d'obtenir un intervalle de confiance grâce aux percentiles de la distribution de l'effet de mesure estimé.

3.3. Analyse comparative des variables soumises à effet de mesure

Après la phase de repérage des variables avec effet de mesure réalisée à l'aide des deux méthodes testées, un travail de comparaison a été effectué sur les variables repérées comme porteuses d'un effet de mesure. Il en ressort des divergences. Certains effets de sélection résiduels restent présents dans l'échantillon multimode contrairement à l'échantillon embarqué. En dehors de ces effets de sélection résiduels, 10 variables présentent un effet de mesure avec l'échantillon multimode et 6 avec l'échantillon embarqué (exemples en Annexe 4). Les variables détectées comportent au moins une de leurs modalités présentant un effet de mesure. Les six variables de l'échantillon embarqué sont communes avec celles détectées avec l'échantillon multimode (avec quelques différences sur les modalités). L'hypothèse retenue est que l'échantillon embarqué ne permet pas de détecter l'ensemble des variables soumises à effet de mesure du fait de ses caractéristiques intrinsèques (taille, représentativité) qui ne répondent pas entièrement aux besoins des analyses. On détecte ainsi 4 variables supplémentaires avec l'échantillon multimode. Les variables principales qui ressortent avec un effet de mesure sont les questions sur le rapport au travail, sur l'équilibre entre la vie professionnelle et familiale, sur les difficultés rencontrées, etc. Elles sont majoritairement de nature subjective.

On donne ici quelques exemples pour illustrer les différences dans les résultats obtenus à partir de l'échantillon embarqué et de l'échantillon multimode. La question Q410 (tableau 7) apparaît comme ayant un effet de mesure significatif selon les analyses des deux échantillons. Cependant, quatre des sept modalités ressortent avec un effet de mesure uniquement sur l'échantillon multimode (non significatives dans l'échantillon embarqué, en gris dans le tableau 10). Par ailleurs, l'ampleur des effets n'est pas la même. Le sens des effets, lui, est le même quelle que soit la méthode : sur internet, les répondants sont plus prédisposés à répondre « Oui » à chacune des modalités du QCM.

Tableau 10 : Décomposition et comparaison de l'effet de mode pour la variable Q410 (modalité « oui ») de l'échantillon embarqué et multimode

Votre point de vue sur le travail a-t-il été modifié par la crise sanitaire. Accordez-vous plus d'importance à :	**Echantillon embarqué**			**Echantillon multimode**		
	Effet de mesure	**Effet de sélection**	**Effet de mode**	**Effet de mesure**	**Effet de sélection**	**Effet de mode**
La sécurité de l'emploi	0,09	-0,03	0,05	0,08	-0,05	0,03
La rémunération	0,09	-0,02	0,07	0,15	-0,08	0,06
L'ambiance	0,04	0,02	0,06	0,09	0,02	0,11
L'utilité du travail	0,02	0,00	0,02	0,09	-0,05	0,04
La reconnaissance au travail	0,04	0,00	0,04	0,11	-0,03	0,08
L'éthique	0,07	0,01	0,09	0,11	-0,05	0,06
L'autonomie dans le travail	0,11	-0,01	0,10	0,07	-0,04	0,03

Note de lecture : l'effet de mesure est de 9 points pour la modalité « la rémunération » et de 4 points pour la modalité « l'ambiance ». Les effets grisés ne sont pas significatifs.

D'autres questions ont des dissemblances du point de vue des modalités concernées par l'effet de mesure. Dans l'échantillon embarqué, à la question sur les motifs de reconversion professionnelle (QCM randomisé avec 8 modalités), l'effet de mesure est de 18 points sur la modalité « Améliorer vos conditions d'emploi ». Dans l'échantillon multimode, cette modalité ne présente aucun effet. En revanche, un effet de mesure est détecté sur la modalité « Donner plus de sens à votre travail ». La comparaison des deux jeux de variables présentant un effet de mesure permet tout de même de rendre compte d'un périmètre commun aux deux échantillons analysés. Ces variables semblent être porteuses du même type d'effet quel que soit l'échantillon analysé.

Une synthèse des modalités à traiter, selon l'effet de mesure principal, permet de dresser un bilan de ce qui est observé dans l'enquête Génération, « Covid et après ? ». Tout comme dans l'enquête

Génération 2017, plusieurs types d'effets de mesure peuvent être combinés au sein de la même question (cas des QCM).

Tableau 11 : Répartition des modalités impactées selon l'effet de mesure principal

	Nombre de variables avec effets de mesure	
TYPE D'EFFET PRINCIPAL	**Echantillon embarqué**	**Echantillon multimode**
ERGONOMIE - GUIDANCE		
Affichage des questions à choix multiple (compréhension ou satisficing)	3	3
SATISFICING		
Questions avec échelle de Lickert		1
DÉSIRABILITE	3	6
Total	**6**	**10**

Les solutions préconisées sont similaires à celles proposées pour Génération 2017.

Peu de variables avec effet de mesure sont détectées dans l'enquête Génération, « Covid et après ? ». Une partie des questions ayant un effet de mesure significatif sont les questions subjectives à choix multiples, du fait d'une passation différenciée sur les deux modes même si l'affichage à l'écran est identique. De plus, certaines de ces questions sont longues et complexes, le nombre de modalités énuméré par l'enquêteur et leur ordre fait perdre de vue le sujet de la question et peut prêter à confusion. Un travail sur ces questions et leur formulation est à préconiser dans les enquêtes suivantes afin d'en diminuer les effets (présenter les modalités une à une sur internet pour être plus proche du mode téléphone ?).

L'analyse de l'échantillon embarqué (champ incomplet de la population cible) a permis de détecter un nombre plus faible de variables soumises à effet que l'échantillon multimode [16]. Cependant, il a permis de mieux maîtriser l'effet de sélection car n'est apparu aucun biais de sélection résiduel après repondération. Néanmoins, pour réaliser des analyses plus précises, l'échantillon embarqué, devrait être de taille plus importante afin de diminuer les intervalles de confiance et couvrir l'ensemble de la population.

Conclusion

En 2020-2021, le Céreq a réalisé l'enquête Génération pour la première fois dans son nouveau dispositif multimode. La collecte a été réalisée dans un contexte impacté par la crise sanitaire. La production de cette enquête Génération 2017 a nécessité une lourde phase de préparation : expérimentations, conception du questionnaire et du protocole de collecte multimode internet-téléphone et rédaction de spécifications de développement d'un outil de collecte ergonomique [17]. Ce travail méthodologique et technique en amont de la collecte a été indispensable pour limiter les abandons en cours de collecte et l'apparition de biais de mesure.

Pour estimer l'effet de mesure, une méthode d'appariement sur score de propension a été implémentée sur la quasi-totalité des variables de l'enquête. Ce choix a permis à la fois de n'omettre aucune variable potentiellement biaisée et d'affiner le calcul du score de propension. Dans le cadre de cet article, la pondération n'a pas été utilisée (elle sera intégrée dans la poursuite de ce travail).

Concernant les résultats obtenus, une question subsiste quant au seuil d'acceptabilité en termes d'écart de mesure entre les deux modes de collecte.

Le résultat de ces analyses est globalement satisfaisant, les variables impactées par cet effet de mesure sont en nombre limité dans l'enquête Génération 2017. Certaines variables étaient attendues, notamment des variables avec échelle ou d'autres variables déjà expertisées comme sujettes à biais de désirabilité sociale. Au-delà de ces cas, ce travail de détection exhaustif a permis de vérifier la cohérence du remplissage du calendrier d'activité, élément central de l'enquête. Il s'avère rempli de la même façon sur les deux modes de collecte, sauf pour certains cas particuliers. Les concernant, il faudra faire évoluer l'outil pour les prochaines collectes.

Les recommandations en matière d'identification et de traitement des effets de mode dans les enquêtes sont en constante évolution. Une question sur l'utilisation d'internet dans le questionnaire ou la collecte sur échantillon monomode de contrôle sont des solutions qui n'ont pas été utilisées dans l'enquête Génération 2017 mais pourront l'être dans les suivantes. L'enquête « Covid et après ? » a offert la possibilité de tester le repérage des effets de mesure via un échantillon de contrôle. Les variables avec effet ont été comparées avec celles détectées par un appariement sur score de propension sur échantillon multimode. L'échantillon monomode a détecté moins de variables soumises à effet de mesure que la méthode d'appariement. Cela est en partie imputable à sa faible taille et donc à une variance importante des estimateurs. Si l'utilisation d'un échantillon monomode est envisagée pour les prochaines enquêtes Génération, des questions subsistent : quelle doit être sa taille ? Comment préserver le taux de réponse en réinterrogation ?

La mise en œuvre de ces analyses nous amène à nous interroger sur le contenu du questionnaire Génération et sur des aspects méthodologiques. Les variables présentant un effet de mesure sont majoritairement de nature subjective. Les conclusions des expérimentations avaient déjà pointé la sensibilité de ces variables empreintes de désirabilité sociale. Elles ont donc été limitées dans l'enquête. Faut-il aller jusqu'à renoncer aux questions subjectives dans les enquêtes multimodes ? Il serait regrettable d'éliminer ces questions historiques, d'autant que le mode internet nouvellement introduit dans l'enquête semble plus fiable sur ce type de questions.

D'autre part, on peut penser que la détection des biais de mesure a bien fonctionné pour cette première interrogation de la Génération 2017 car la composition des deux groupes de répondants était équilibrée (environ 50 % par mode). Mais pourrons-nous reproduire ces analyses si un mode devient majoritaire lors des prochaines interrogations ? Par ailleurs, comment effectuer des comparaisons entre Générations ou, pour une Génération donnée, entre les deux vagues d'enquête, si la répartition des modes change fortement ?

On touche ici du doigt la complexification des traitements réalisés en aval de l'enquête, en lien avec le passage au multimode. La mise en œuvre des enquêtes multimodes longitudinales reste un défi pour le Céreq comme pour l'ensemble des producteurs d'enquête.

Annexe 1 – Protocoles de collecte mis en œuvre

Un protocole multimode séquentiel et concurrentiel pour l'enquête Génération 2017

Le protocole de l'enquête Génération 2017 est multimode, séquentiel et concurrentiel. Il a été perturbé par la survenue de la crise sanitaire. Prévue initialement au printemps 2020, l'enquête a été reportée à l'automne, puis prolongée dans le temps du fait de difficulté de collecte. Les enquêtés ont été informés par mail, SMS ou par courrier postal du lancement de l'enquête, et ont reçu des codes pour compléter leur questionnaire sur internet. Les non-répondants sont ensuite relancés par mail et par téléphone.

Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
Priorité INTERNET	Choix INTERNET-TELEPHONE	Priorité TELEPHONE	Priorité TELEPHONE
1 mois	1 mois	2 mois	3 mois
Internet + premiers contacts téléphoniques des individus sans adresse mail	L'ensemble des non-répondants est contacté, y compris ceux qui ont démarré et n'ont pas terminé. Identification de la personne dans le champ de l'enquête par téléphone puis proposition de continuer sur internet + relance sur internet	Relances téléphoniques de l'ensemble des non-répondants, en privilégiant la réponse par téléphone	Relances de la population à faible taux de réponse + ajout de la réserve d'échantillon car l'objectif non atteint

Un protocole multimode avec introduction d'un échantillon embarqué » pour l'enquête Génération « Covid et après ? »

Toutes les interrogations précédentes de Génération 2010 (à 3, 5 et 7 ans) ont été effectuées par téléphone. L'enquête « Covid et après ? » est une nouvelle interrogation de Génération 2010 (hors dispositif) à être réalisée en multimode.

Phase 1	**Phase 2**	**Phase 3**	**Phase 4**
Priorité INTERNET	Choix INTERNET-TELEPHONE	Priorité TELEPHONE	Priorité INTERNET
3 semaines	3 semaines	1 semaine	1 semaine
Internet (97% des individus disposent d'une adresse mail)	L'ensemble des non-répondants est contacté par téléphone et relancé sur internet	Relances téléphoniques de l'ensemble des non-répondants, en privilégiant la réponse par téléphone	Relances sur l'ensemble des non-répondants par internet

Annexe 2 – Sélection des variables analysées dans l'enquête Génération 2017

Certaines variables ont été simplifiées ou écartées. Il s'agit notamment des variables de type nomenclatures, avec de très nombreuses modalités. Par exemple, pour traiter la variable de la spécialité de formation (NSF - nomenclature de spécialités française), seule la première position a été conservée. Par ailleurs, le questionnaire étant très filtré, le nombre de répondants par mode peut être faible sur certaines questions. Les variables sélectionnées respectent la contrainte d'au moins 100 répondants par mode (soit 200 déclarations par question au minimum).

Le jeu de données de l'enquête Génération 2017 est composé de deux tables : la base Individus (534 variables) comprenant les caractéristiques propres aux répondants et des modules de questions posées à des moments précis du parcours, et la base Calendrier (162 variables) qui regroupe l'information longitudinale sur le parcours professionnel. Un traitement différencié a été réalisée pour restreindre le champ d'analyse. Pour la table Individus, 316 variables ont été sélectionnées répondants aux critères énoncés (soit 1 070 modalités). Pour la table Calendrier, la situation à la date de l'enquête a été privilégiée (quel que soit la situation professionnelle décrite – emploi, chômage, reprise d'études ou autre situation) et concerne 119 variables (soit 340 modalités). Concernant les questions posées seulement au cours de situation passées, elles ont été analysées via une autre table Calendrier (Situation au 22ème mois). Enfin, une analyse spécifique a été menée sur une variable numérique : le salaire.

Annexe 3 - Industrialisation du matching

Ce programme est perfectible (1h30 pour 900 modalités).

```
library(Matching)
library(cobalt)
library(here)
library(xlsx)
library(sas7bdat)
library(foreign)
library(stringr)

### a) Constitution de la base d'analyse

#Récupération du pscore

tbl1=merge(BASE,mybase[,c("ident","pscore")], by="ident")
tbl2=tbl1[,-1] # suppression de l'identifiant individu

#Base d'analyse
tblok=tbl2
names(tblok)

res=data.frame()
noms_var=c()

#Variables à tester (en début de fichier)
for (i in 1:301) #Indiquer le numéro de la dernière variable à analyser
{
 ### b) Gestion des non-répondants
 temp=tblok[,c(i,(dim(tblok)[2]-21):dim(tblok)[2])] #19 variables dans le modèle + pscore et modalité à traiter
 temp[,1]=as.numeric(temp[,1])
 temp2=na.omit(temp)
 temp2[,1]=as.character(temp2[,1])

### c) Construction d'indicatrices pour chacune des variables
 temp3= splitfactor(data = temp2, var.name = names(temp2)[1],replace = FALSE, drop.level=NULL,
 drop.first=FALSE)

 #Nouvelle boucle pour chaque modalité de la variable pour le matching
  for (j in (dim(temp2)[2]+1):(dim(temp3)[2]))
 {
  Y <- temp3[,j] #modalité j de la variable i
  Tr <- temp3$MODEOK #variable de traitement
  temp3$PHD <- as.numeric(temp3$PHD4)

  ### d) Matching avec remise et appariement exact sur PHD (plus haut diplôme obtenu)
  MatchATT.out <- Match(Y = Y, Tr = Tr,
                        estimand="ATT",
                        M=1,
                        replace = TRUE, #Avec remise
                        ties = FALSE,
                        caliper = 0.10,
                        X = cbind(temp3$pscore, temp3$PHD),
                        exact = c("FALSE","TRUE")) #FALSE pour pscore et TRUE pour PHD

  gussmatch <- data.frame(MatchATT.out$MatchLoopC)
  freq <- data.frame(table(gussmatch$X2))
  freq2 <- freq[rev(order(freq$Freq)),]
```

```
  maxmatch <- freq2[1,2] #Nombre d'individus maximum matchés

  ### e) Différences standardisées et ratios de variance avant/après appariement

  # Sélection automatique des variables du modèle avec au moins 2 modalités
l=c("PHD17","ETABREG2","ACTUREG","TYPE2","SITDEPCS","CONTRAT","ANTER","QPV_FINETU","HABDE",
"SD080", "SD140","CPRO","APP","BAC_TYPE","BACMENT","PS020","PS050")

  mod=" GENRE"

  for (k in 1:length(l))
  {
   if (dim(table(temp3[,l[k]]))>1)
   {
    mod=paste0(mod," + ",l[k])
   }
  }
  balMatch <- bal.tab(x=MatchATT.out,
                      formula = formula(paste0("MODEOK ~ ",mod)),
                      data = temp3,
                      un = TRUE,
                      disp.means=TRUE,
                      disp.v.ratio = TRUE)

  balmatch1ATT <- data.frame(balMatch$Balance)
  b=max(balmatch1ATT$Diff.Adj)<0.10 #Test de la propriété équilibrante

  ### f) Création du tableau avec résultats pour chaque matching
  res=rbind(res,c(est=MatchATT.out$est,
                  orignobs=MatchATT.out$orig.nobs,
                  origtreatednobs=MatchATT.out$orig.treated.nob,
                  wnobs=MatchATT.out$wnobs,
                  ndropsmatches=MatchATT.out$ndrops.matches,
                  capture.output(summary(MatchATT.out, full=TRUE))[5],
                  b,maxmatch)) #Récupération du résultat du matching
 }
 noms_var=c(noms_var,names(temp3)[(dim(temp2)[2]+1):(dim(temp3)[2]-1)])
}
row.names(res)=noms_var
names(res)=c("Estimate","Nbobs","Nbcawi","NbcawiMatch","Nbdrop","pval","PropEquok","Maxmatch")
res2=res
res2$pval2=str_replace(res2$pval,"p.val.....  ","")
```

Annexe 4 – Synthèse des variables soumises à effet de mesure

Enquête Génération 2017

Le questionnaire de l'enquête Génération comporte des questions communes à la plupart des enquêtes Génération et des questions provenant de modules d'extensions thématiques commandités par des partenaires (Injep, Ministère logement…). Dans l'analyse suivante, il sera question des variables du tronc commun présentant un effet de mesure > 5 points.

Variables	Question	Modalité avec effet	Point écart	Type d'effet	Solution
EP030	Nombre entreprises Intérim	Une seule entreprise intérim	22	REMPLISSAGE CALENDRIER	RECODE
EP040	Nombre de missions d'intérim	Une seule mission	15	REMPLISSAGE CALENDRIER	RECODE
DI010	Victime de discrimination	Non	-14	ERGONOMIE	REGROUPER NSP-NON
		Ne sait pas	14		
EXP050	Diriez-vous que cet emploi a perturbé le cours normal de vos études ?	Non, pas vraiment	14	ÉCHELLE	REGROUPEMENT
		Non, pas du tout	-14		
PS110M3	Raison du choix d'une formation scolaire plutôt qu'apprentissage	Autre	13	ERGONOMIE	SUPPRIMER LA MODALITE M3
AUP010	Autre situation calendrier	Autre	13	REMPLISSAGE CALENDRIER	RECODE
SD090	En quelle langue vous parlait votre père lorsque vous étiez enfant ?	Français	-12	ERGONOMIE	REGROUPER 1 et 3 ? INEXPLOITABLE ?
		Français et une autre langue	10		
SD100	En quelle langue vous parlait votre mère lorsque vous étiez enfant ?	Français	-11	ERGONOMIE	REGROUPER 1 et 3 ? INEXPLOITABLE ?
		Français et une autre langue	10		
COVEA110	Dégradation condition de travail, est-ce toujours le cas ?	Oui	10	DÉSIRABILITE	MODÈLE

Variables	Question	Modalité avec effet	Point écart	Type d'effet	Solution
PHD170M1	Diplôme obtenu	Maitrise	-7	ERGONOMIE	AUCUN
PHD170M3		Diplôme de niveau Bac+4	-4		
PHD170M5		Master	-14		
PHD170M8		Autre	-5		
PP020	Optimisme quant à l'avenir professionnel	Plutôt inquiet	7	DÉSIRABILITE	MODÈLE / IMPUTATION
		Plutôt optimiste	-3		
		Très optimiste	-6		
AC050	Choix d'un entretien avec un chercheur	Oui	-7	DÉSIRABILITE	AUCUN
EAO040	Diriez-vous que vous êtes payé ?	Plutôt bien payé	-6	ÉCHELLE	REGROUPEMENT
		Plutôt mal payé	7		
EAO060M1	Où chercher vous un emploi ?	Dans votre commune	6	ERGONOMIE	RECODE
EAO060M2		Dans votre département	6		
EAO060M3		Dans d'autres régions	3		
QF210	Durée interruption	Interruption de 12 mois	-6	SEUIL	REGROUPEMENT MODALITE 1 et 2
IMT060M1	La mission locale ou PAIO vous ont-ils permis de	Trouver un emploi ou un stage	4	DÉSIRABILITE	MODÈLE / IMPUTATION
IMT060M2		Mieux cibler votre recherche d'emploi	4		
IMT060M4		Percevoir des allocations	-6		

Enquête Génération « Covid et après ? »

Questions	Modalité si QCM	Hypothèse 1	Hypothèse 2	Solution
Rapport au travail	Rémunération	ERGONOMIE	DÉSIRABILITE	MODÈLE
	Reconnaissance travail	ERGONOMIE	DÉSIRABILITE	MODÈLE
	Éthique	ERGONOMIE	DÉSIRABILITE	MODÈLE
	Utilité travail	ERGONOMIE	ERGONOMIE	MODÈLE
	Ambiance	ERGONOMIE	ERGONOMIE	MODÈLE
	Sécurité emploi	ERGONOMIE	ERGONOMIE	MODÈLE
	Autonomie	ERGONOMIE	ERGONOMIE	MODÈLE
Entretien qualitatif		DÉSIRABILITE	DÉSIRABILITE	AUCUN
Equilibre avec la vie professionnelle	Équilibre vie professionnelle/vie familiale	DÉSIRABILITE	ERGONOMIE	MODÈLE
	Équilibre entre travail et activités extra-professionnelles	DÉSIRABILITE	ERGONOMIE	AUCUN
Motif de reconversion	Donner plus de sens à votre travail	DÉSIRABILITE	DÉSIRABILITE	MODÈLE
Ressource mobilisée	un conseil en évolution professionnelle	ERGONOMIE	SÉLECTION	AUCUN
	un accompagnement de votre employeur	ERGONOMIE	SÉLECTION	AUCUN
	un bilan de compétences	ERGONOMIE	SÉLECTION	AUCUN
Difficulté rencontrée	Manque de ressources économiques	DÉSIRABILITE	DÉSIRABILITE	AUCUN
Optimisme avenir professionnel		DÉSIRABILITE	DÉSIRABILITE	MODÈLE
Opinion sur le télétravail		ÉCHELLE	ÉCHELLE	REGROUPEMENT

Bibliographie

[1] Barret C., Dzikowski C., « La collecte par Internet est-elle l'avenir des enquêtes Génération du Céreq ? », *12èmes Journées de Méthodologie Statistique*, 2015.

[2] Cissé M., Barret C., « Agrégation de données multimode : Impact sur la modélisation des variables présentant un effet de mesure », *13èmes Journées de Méthodologie Statistique*, 2018.

[3] Fougère D., « Les méthodes économétriques d'évaluation », *Revue française des affaires sociales*, pp 105-128, janvier 2010.

[4] Rosembaum P., Rubin D., « The central role of propensity score in observational studies for causal effects », *Biometrika*, pp 70:41-55, 1983.

[5] Heckman J., Ichimura H., Todd P., « Matching as an Econometric Evaluation Estimator », *Review of Economic studies*, vol 65, pp 261-294, 1998.

[6] F. Beck, L. Castell, S. Legleye, A. Schreiber, « Le multimode dans les enquêtes auprès des ménages : une collecte modernisée, un processus complexifié », *Courrier des statistiques N°16*, Janvier 2022, Insee, p. 7-28

[7] Quantin S., « Estimation avec le score de propension sous R », Méthodologie statistique, document de travail, INSEE, 2018.

[8] Abadie,Imbens G. W., « Large Sample Properties of Matching Estimators for Average Treatment Effects », *Econometrica*, pp 74:235-267, January 2006.

[9] Vinceneux K., « Mode de collecte et questionnaire, quels impacts sur les indicateurs européens de l'enquête Emploi ?», Méthodologie statistique, document de travail, INSEE, octobre 2018.

[10] Krauter, F., Presser, S., & Tourangeau, R. « Social desirability bias in CATI, IVR, and Web surveys: The effect of mode and question sensitivity. » *Public Opinion Quarterly*, vol 5, n°72, pp. 847-865, 2008.

[11] Shapiro, R., Siegel, A. W., Scovill, L. C., & Hays, J. (1998). « Risk-taking patterns of female adolescents: What they do and why. » *Journal of adolescence*, 21(2), 143-159

[12] Legleye, S., « Effets de sélection, imputations et effets de mode : les dernières tendances en matière de multimode », Séminaire de Méthodologie Statistique, 2017.

[13] Hirano, K., Imbens, G. W., & Ridder, G. « Efficient estimation of average treatment effects using the estimated propensity score » *Econometrica*, 2003, 71(4), 1161-1189.

[14] Davison, A.C. and Hinkley, D.V., « *Bootstrap Methods and Their Application »*, Chapter 5. *Cambridge University Press.*, 1997

[15] Fromont M., Vimond M., « Bootstrap et rééchantillonnage », Atelier SFdS – Partie 1, 2012

[16] Noack E., Sigot J.C. « Collecte multimode de l'enquête DEFIS Salariés : À la recherche d'un effet de mesure », *13èmes Journées de Méthodologie Statistique*, 2018.

[17] Bouvet N., Dabet G., Gaubert E., Olaria M., Oujia I., Mazari Z., Vignale M., Wierup E-L., « Conception et évaluation d'un outil de collecte multimode (internet - téléphone) : quelles spécificités pour quels résultats ? », *14èmes Journées de Méthodologie Statistique*, 2022.