

## **APPRENDRE DES PARADONNEES POUR AMELIORER LES PROTOCOLES DE COLLECTE : L'EXEMPLE D'EPICOV**

Géraldine Charrance (\*), Paul Cochet (\*), Aude Leduc (\*\*), Marine Bondon (\*), Thomas Merly-Alpa (\*)

(\*) Ined, Service des Enquêtes et Sondages

(\*\*) Insee, Direction des statistiques démographiques et sociales

Correspondantes : geraldine.charrance@ined.fr ; aude.leduc@insee.fr

**Mots-clés** : Paradoxonnées, collecte, multimode, Fidéli, Covid-19

**Domaine concerné** : 4.6 Paradoxonnées, 15. Mesures et impact de la pandémie de Covid

### **Résumé**

*La collecte de la vague 1 de l'enquête EpiCov s'est déroulée entre le 2 mai et le 1<sup>er</sup> juin 2020. Elle a reposé, en métropole, sur un échantillon de 350 000 personnes issu de Fidéli. Deux modes de collecte ont été employés : internet, complété par le téléphone pour un cinquième des lots aléatoires composant l'échantillon. Environ 130 000 personnes ont répondu en métropole dont 30 000 par téléphone. La collecte a été l'occasion de recueillir des paradoxonnées sur les deux modes de collecte : nombre de connexions au site internet de collecte, matériel utilisé, jour et heure de remplissage du questionnaire, type de numéros de téléphone disponibles, historique des appels, protocole de relance, etc.*

*Cet article rend compte de l'exploitation des paradoxonnées de EpiCov avec un double objectif : i) documenter les potentialités de Fidéli en tant que base de sondage pour une collecte multimode ; ii) aboutir à des préconisations pour les protocoles de collecte d'enquêtes futures.*

*Pour répondre à l'objectif d'évaluation de Fidéli comme base de sondage, une analyse de la disponibilité et de la validité des coordonnées de contact présentes dans le fichier est menée et déclinée en fonction des caractéristiques sociales disponibles dans Fidéli (sexe, âge, décile de niveau de vie, quartier prioritaire de la politique de la ville, etc.).*

*L'apport de l'enrichissement annuel est également étudié, globalement et en fonction des profils socio-démographiques. Par la suite, l'article s'intéresse à la rentabilité des différents lots du protocole de collecte. Il s'attache à mettre en lien les taux de réponse selon les protocoles (monomode versus multimode), les caractéristiques sociales des enquêtés, les modes de collecte et la richesse des coordonnées de contact à disposition (adresse mail ou non, numéro de téléphone ou non, numéro de mobile ou non, etc.). Grâce à la multiplicité des coordonnées disponibles, le protocole de relance a été particulièrement étoffé dans l'enquête EpiCov (mails, SMS, messages vocaux, etc.). L'efficacité des différents moyens de relance est analysée et questionnée.*

*Le nombre de connexions et, pour les lots concernés, d'appels nécessaires au remplissage du questionnaire ainsi que le mode de réponse finalement utilisé sont également analysés et déclinés en*

*fonction des protocoles et des profils sociodémographiques. Une analyse des abandons est également prévue, notamment en lien avec le type de matériel utilisé pour le remplissage du questionnaire en ligne. En dernier lieu, un focus est fait sur les lots multimodes, divisés en deux groupes. Pour certains lots, la collecte s'est déroulée selon un multimode concurrentiel où les collectes internet et téléphone ont été ouvertes dès le début. Pour un autre lot, elle s'est déroulée selon un multimode séquentiel, la collecte téléphone ayant été ouverte dans un second temps, après la collecte internet. Une comparaison est menée pour en tirer des conclusions sur la manière d'exploiter au mieux la complémentarité des modes de collecte. La taille importante de l'échantillon EpiCov permet de nuancer les résultats selon différents types de sous-populations d'étude.*

## **Abstract**

*The Fideli database is of growing interest to survey designers in France. Rich in contact information (postal addresses, telephone numbers, e-mail addresses, etc.), it offers new and innovative possibilities in terms of survey methodology, facilitating the implementation of mixed-mode surveys in particular. However, we still have little experience with this database as a sampling frame: quality of contact information, efficiency and relevance of mixed-mode protocols, etc. Through the analysis of paradata of the first wave of the EpiCov survey, for which 350 000 people living in metropolitan France were sampled during the 2020 spring, this paper aims to fill this gap and, where possible, to provide recommendations for future survey protocols.*

## Introduction

L'enquête EpiCov (Épidémiologie et Conditions de vie liées au Covid 19, voir encadré 1) a été conçue dans l'urgence au début de la crise sanitaire par l'Inserm et la DREES en collaboration avec l'Insee et Santé publique France. La nécessité de recourir à une collecte « à distance » du fait des restrictions sanitaires en France a eu un effet accélérateur sur l'expérimentation au sein des enquêtes de la statistique publique auprès des ménages, de collectes multimodes par Internet et téléphone. Actuellement, ligne directrice de la transformation des enquêtes ménages de la statistique publique, le multimode (Beck et al., 2022) renvoie à la pluralité des modes de contact utilisés pour inviter les personnes à répondre aux enquêtes (courrier, téléphone, mail) et aux différents modes de collecte possibles pour répondre au questionnaire (par téléphone, sur internet, en face-à-face, etc.). De ce point de vue, l'évaluation du protocole de l'enquête EpiCov fournit des enseignements intéressants pour concevoir à l'avenir des protocoles multimodes. L'enquête a été échantillonnée dans Fidéli (Fichiers démographiques sur les logements et les individus). La richesse en coordonnées de contact de cette base de sondage a, ainsi, permis de diversifier les modes de contact. De plus, l'échantillon a été découpé en plusieurs lots aléatoires sur lesquels différents protocoles ont été expérimentés, soit monomode (Internet : CAWI), soit multimode (Internet et téléphone : CAWI et CATI).

### Encadré 1 : La vague 1 de l'enquête EpiCov

La vague 1 de l'enquête EpiCov [Warszawski et al., 2021] s'est déroulée du 2 mai au 1er juin 2020 à partir d'un échantillon de 371 000 personnes de 15 ans ou plus tirées dans Fidéli. Parmi elles, 350 000 personnes résidaient en métropole et 21 000 résidaient dans les trois départements et régions d'outre-mer (DROM) concernés par l'enquête (Guadeloupe, Martinique, La Réunion). La collecte a donc débuté durant le premier confinement en métropole qui a duré jusqu'au 11 mai.

L'échantillon, de grosse taille, a été découpé en 20 lots aléatoires. En métropole, chaque lot comportait donc 17 500 personnes. Les personnes des lots 5 à 20 n'avaient la possibilité de répondre à l'enquête que par l'intermédiaire d'internet. Les personnes des lots 1 à 3 pouvaient répondre par Internet ou par téléphone. Les deux modes de collecte étaient concurrentiels. Les personnes du lot 4 pouvait répondre en début de collecte uniquement par Internet, puis des moyens enquêteurs se sont libérés ce qui a permis d'engager une collecte téléphonique sur les non-répondants Internet. Ce lot est donc passé d'une collecte monomode (Internet) à une collecte multimode concurrentielle (Internet et téléphone) en cours de terrain (à partir du 14<sup>ème</sup> jour de collecte).

Par ailleurs, deux questionnaires ont été proposés. Un questionnaire « long » aux personnes des lots 1 et 20 et un questionnaire « court » à l'ensemble des autres lots. La durée des deux questionnaires différait d'environ 10 minutes en moyenne (20 minutes pour le questionnaire court, 30 minutes pour le questionnaire long). Enfin, une sous-partie de l'échantillon se voyait dans le questionnaire proposer l'envoi d'un « kit » d'auto-prélèvement afin d'effectuer un test sérologique de présence d'anticorps liés au Covid 19.

Le protocole d'annonce et de relance de l'enquête pour animer la collecte a été particulièrement fourni. Selon la variété des coordonnées à disposition pour les personnes échantillonnées et le temps qu'elles mettaient à répondre, les personnes ont pu recevoir un courrier d'annonce postal et un courrier de relance postal, trois mails, trois SMS, deux messages vocaux couplés avec l'envoi d'un mail ou d'un SMS. Le calendrier de ces envois est présenté en Annexe 1. Toutes les coordonnées à disposition pour le ménage ont été utilisées pour joindre les personnes échantillonnées (donc y compris les coordonnées des conjoints ou des parents).

Le taux de réponse à la vague 1 en métropole a été de 37,6 % ; il a été de 35,3% pour les lots internet et de 46,7 % pour les lots multimodes.

L'analyse présentée ici porte uniquement sur l'échantillon métropolitain. Les collectes multimodes se déroulent différemment dans les DROM qu'en métropole, en témoigne le taux de réponse par Internet beaucoup plus faible dans les DROM qu'en métropole (15,4 % contre 35,3 %). L'étude pourrait être prolongée pour en tirer des enseignements spécifiques aux collectes d'enquêtes par Internet et téléphone dans les DROM.

Fondé sur l'exploitation des parodonnées de la première vague d'EpiCov (voir encadré 2), l'objectif de cet article est multiple. Il vise dans un premier temps à documenter les potentialités de Fidéli comme base de sondage et d'en mettre en évidence la richesse tout autant que les limites (1). Si cette base de sondage permet de multiplier les canaux pour joindre les personnes, les coordonnées de contact à disposition sont inégalement distribuées, en termes de disponibilité mais aussi de qualité, selon les caractéristiques sociodémographiques des individus. Ensuite, il identifie pour chaque mode de collecte, internet ou téléphone, quelles ont été les stratégies les plus rentables pour favoriser la participation à l'enquête ou *a minima* pour établir un contact avec la personne sélectionnée (2). Dans un troisième temps, il cherche à évaluer ce qu'apporte un volet téléphonique en complément d'une collecte sur internet sur les comportements de non-réponses totale et partielle et sur la structure des répondants (3). Enfin, il s'intéresse plus spécifiquement aux lots exploités en multimode afin de rendre compte de l'effet d'une ouverture plus tardive de la collecte par téléphone, que ce soit en termes de participation, de rythme de collecte, de mode de réponse effectivement utilisé ou encore de coûts de collecte, qui sont autant d'aspects à prendre en compte lors de la conception d'enquêtes (4).

L'objectif global est de formuler des recommandations pour de futures enquêtes quant à la mise en œuvre de protocoles multimodes, en particulier lorsqu'ils sont couplés aux spécificités d'une base de sondage comme Fidéli, sur laquelle les études menées demeurent encore rares.

Enfin, cet article enrichit la réflexion sur les exploitations possibles des parodonnées d'une enquête. Souvent laissées de côté (par rapport aux données issues des questionnaires), les parodonnées ont des potentialités en termes d'enseignements en matière de collecte. Cet article permet de capitaliser sur la manière d'analyser statistiquement ces données et d'en tirer des enseignements.

#### **Encadré 2 : Données mobilisées**

Dans cette étude, les différents types de données décrites ci-dessous ont été exploitées. La plupart d'entre elles peuvent être qualifiées de « parodonnées » car il ne s'agit pas des données du questionnaire mais de données caractérisant la collecte. Ces données ont été fournies par Ipsos, le prestataire en charge de la collecte d'EpiCov, d'après leurs formats de remontée de ce type de données.

- données de gestion : elles participent à la gestion de la collecte au quotidien par le prestataire. Elles comprennent, pour chaque personne de l'échantillon, le calendrier des différentes relances, les retours éventuels en plis non distribués (PND<sup>1</sup>), mail ou SMS non distribués, les types de numéro de téléphone à disposition pour joindre la personne échantillonnée ou son ménage, les informations de connexion (support, navigateur, système d'exploitation, etc.) en cas de questionnaire rempli par Internet, le mode sur lequel le questionnaire a été complété (Internet ou téléphone) et la date.

- historique des appels : comme son nom l'indique, il s'agit d'une base de données qui comporte une ligne par appel téléphonique spécifiant l'issue de l'appel (permettant notamment de savoir s'il y a eu contact avec le ménage ou non, interview ou non), le moment de l'appel (jour et heure exacts) et le type de numéro de téléphone utilisé pour cet appel.

- échantillon Fidéli : il permet de décrire l'échantillon à l'aide de variables sociodémographiques qui sont donc disponibles pour les répondants et les non-répondants (sexe, âge, niveau de vie, situation familiale, statut d'occupation du logement, tranche d'unité urbaine, zonage quartier prioritaire de la politique de la ville, lieu de naissance, etc.).

- statut de réponse à l'enquête : statut (répondant/non-répondant/hors-champ) défini par l'Insee au regard de la définition du champ de l'enquête et de la complétude des questionnaires.

---

<sup>1</sup> Anciennement dénommés NPAI.

## 1. Richesse des informations de contact dans Fidéli et complémentarité de l'enrichissement annuel

### 1.1. Disponibilité des informations de contact

L'échantillon a été tiré dans le millésime 2019 de Fidéli (logements et leurs occupants au 1<sup>er</sup> janvier 2019, revenus de l'année 2018). La base Fidéli comporte des adresses postales, des numéros de téléphone et des adresses mail. Un enrichissement téléphonique a été mené sur les lots 1 à 4, ouverts à une collecte téléphonique. Pour les lots 1 à 3, il a été mené sur la totalité des personnes, ces lots ayant été ouverts dès le début à une collecte CATI. En ce qui concerne le lot 4, l'enrichissement téléphonique a été mené en cours de collecte, uniquement sur les non-répondants restant dans ce lot à ce moment de la collecte. Lors de l'enrichissement, les numéros issus de l'annuaire n'ont été retenus que s'ils différaient des numéros issus de Fidéli.

Le Tableau 1 présente les informations de contact à disposition pour joindre les personnes des lots 1 à 3 spécifiquement, soit un total de 52 500 personnes, afin de rendre compte plus justement de l'apport de l'enrichissement annuel. Le rattachement des coordonnées à la personne échantillonnée ou à une autre personne du ménage, qui est le conjoint le plus fréquemment, fait référence à ce qui est fait dans les fichiers fiscaux. En cas de mineur échantillonné, les coordonnées sont celles des déclarants fiscaux de son ménage, c'est-à-dire le plus souvent les parents.

Tableau 1: Disponibilité des informations de contact dans l'échantillon EpiCov

	Effectifs	Part dans l'échantillon (en %)
Au moins un mail issu de Fidéli pour joindre le ménage	37 428	71,3
Mail Fidéli de la personne échantillonnée	30 552	58,2
Mail Fidéli d'une autre personne du ménage	26 774	51,0
Au moins un numéro de téléphone issu de Fidéli	37 515	71,5
Mobile Fidéli de la personne échantillonnée	16 432	31,3
Mobile Fidéli d'un autre personne du ménage	13 087	24,9
Fixe géographique Fidéli	18 539	35,3
« 09 » Fidéli	1 735	3,3
Un numéro issu de l'enrichissement annuel	6 249	11,9
Mobile Annuaire	132	0,3
Fixe Annuaire	4 971	9,5
« 09 » Annuaire	1 146	2,2
Au moins un numéro (Fidéli ou Annuaire)	43 764	83,4
Au moins un numéro (Fidéli ou Annuaire) ou un mail Fidéli	47 335	90,2

Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 3, métropole.

Lecture : Parmi les 52 500 personnes composant les lots 1 à 3, 37 428 personnes ont pour leur ménage au moins un adresse mail renseignée dans Fidéli. Cela représente 71,3 % de cet échantillon.

Une adresse mail est disponible pour joindre 71,3 % des ménages ; pour 58,2 % des ménages l'adresse mail est attribuée dans les fichiers fiscaux à la personne échantillonnée spécifiquement. Depuis 2014, il est possible de renseigner deux adresses mail dans la déclaration de revenus : ainsi, pour 37,9 % des personnes échantillonnées, deux adresses mail sont disponibles pour écrire à leur ménage sans que l'on sache a priori s'il s'agit d'adresses distinctes ou non. Aucun enrichissement n'a été mené sur les adresses mail.

En ce qui concerne les numéros de téléphone, un enrichissement annuaire a eu lieu. Il permet d'augmenter de 11,9 points la part de l'échantillon pour laquelle un numéro de téléphone est disponible, la faisant passer de 71,5 % à 83,4 %. Les numéros retrouvés par enrichissement annuaire sont majoritairement des numéros de fixes géographiques<sup>2</sup> (79,5 %), puis des numéros « 09 » (18,3 %) et enfin des numéros de mobile (2,1 %).

La déclaration fiscale, dans sa forme la plus récente, permet de renseigner deux numéros de téléphone (un mobile et un fixe) par déclarant, donc jusqu'à quatre numéros par foyer fiscal. Une fois l'enrichissement annuaire mené, la majorité de l'échantillon dispose d'un unique numéro de téléphone (61,5 %), 19,2 % dispose de deux numéros de téléphone et seulement 2,7 % de plus de deux numéros<sup>3</sup>.

Dans l'échantillon, 5 165 personnes ne possèdent ni adresse mail, ni numéro de téléphone pour les joindre ou joindre leur ménage. Cela représente 9,8 % de l'échantillon.

## 1.2. Caractéristiques des personnes selon les différents modes de contact à disposition

Dans la suite, on s'intéresse aux caractéristiques des personnes selon la disponibilité de coordonnées de contact les concernant. Il s'agit d'évaluer s'il y a des populations plus difficiles à joindre que d'autres et de mettre en perspective l'apport de l'enrichissement annuaire pour joindre les gens. On estime ainsi la probabilité de n'avoir aucun numéro de téléphone, aucun mail, et aucun numéro de téléphone ni mail, en fonction des caractéristiques sociodémographiques disponibles dans Fidéli<sup>4</sup>. On estime également la probabilité d'avoir un numéro de mobile et celle d'avoir uniquement un numéro retrouvé *via* enrichissement annuaire (Tableau 2). Les coefficients présentés sont les effets marginaux issus d'une modélisation logit.

Tableau 2 : Corrélation entre les caractéristiques sociodémographiques et la disponibilité de coordonnées de contact

	Probabilité de n'avoir...		Probabilité d'avoir...		
	aucun numéro de téléphone	aucune adresse mail	aucun numéro de téléphone ni adresse mail	un numéro de mobile pour la pers. échant.	comme numéro, uniquement un numéro annuaire
Constante	14,7***	21,6***	7,2***	21,1***	10,8***
<b>Sexe</b>					
Femmes	Réf.	Réf.	Réf.	Réf.	Réf.
Hommes	-0,8**		-0,6**	14,1***	
Inconnu		26,3***		-19,3***	
<b>Âge</b>					
15 - 24 ans		-10,0***	-2,3***	-15,4***	-2,1***
25 - 34 ans	-2,4***	-10,5***	-3,0***	14,8***	-4,3***
35 - 44 ans	-1,8**	-7,4***	-2,4***	11,0***	-2,8***
45 - 54 ans	Réf.	Réf.	Réf.	Réf.	Réf.
55 - 64 ans	1,9**	9,6***	2,9***	-6,4***	3,6***
65 - 74 ans	2,0**	21,8***	4,2***	-9,9***	5,6***
75 - 84 ans	2,2**	43,7***	6,2***	-15,1***	9,4***

<sup>2</sup> On entend par là des numéros de fixe commençant par un indice géographique : 01, 02, 03, 04 ou 05.

<sup>3</sup> Il s'agit d'un décompte minimal. En effet, au cas où un fixe géographique ou un « 09 » était renseigné pour la personne échantillonnée et également pour son conjoint, on a considéré qu'il s'agissait du même numéro.

<sup>4</sup> Il s'agit du sexe, de l'âge, du fait d'habiter ou non en quartier prioritaire de la politique de la ville, du décile de niveau de vie, de la situation familiale, du lieu de naissance, de la taille d'unité urbaine du lieu d'habitation et du statut d'habitation (locataire, propriétaire, ...).

85 ans et plus	2,1*	46,7***	5,9***	-16,7***	9,3***
Quartier prioritaire de la politique de la ville					
Oui		12,2***	2,2***	-3,1***	4,0***
Non	Réf.	Réf.	Réf.	Réf.	Réf.
Niveau de vie					
1er décile	5,6***	13,7***	5,2***	-4,0***	2,9***
2ème décile	4,1***	12,0***	3,6***	-3,8***	2,3***
3ème décile	2,5**	8,3***	2,3***	-2,9***	1,8**
4ème décile		3,2***	1,1*		
5ème décile	Réf.	Réf.	Réf.	Réf.	Réf.
6ème décile	-2,2**	-3,8***	-1,5**	1,6*	
7ème décile	-2,8***	-8,1***	-2,3***	2,1*	-2,5***
8ème décile	-3,4***	-11,3***	-3,1***	3,8***	-3,3***
9ème décile	-4,8***	-14,6***	-4,2***	4,4***	-4,1***
10ème décile	-5,0***	-17,6***	-5,2***	6,0***	-3,9***
Inconnu	27,0***	40,9***	34,2***	-11,6***	
Situation familiale					
Pers. seule	6,1***	4,7***	3,3***	14,7***	
Couple sans enfant	Réf.	Réf.	Réf.	Réf.	Réf.
Couple avec enfant	-2,9***	-3,9***	-1,9***	-1,6**	
Famille monoparentale				14,0***	
Ménage complexe		3,3***		5,6***	1,5*
Pays de naissance					
France	Réf.	Réf.	Réf.	Réf.	Réf.
Étranger	1,1*	4,9***	2,1***		
Inconnu				-8,0**	6,5**
Tranche d'unité urbaine (en nombre d'habitants)					
Rural	-1,6**		-1,1***		2,7***
2 000-<10 000	-1,1*				1,1*
10 000-<50 000	Réf.	Réf.	Réf.	Réf.	Réf.
50 000-<200 000	-0,4	-0,6	0,0	1,3	0,2
200 000-<2 000 000	-1,3*	-3,7***	-0,9**	3,8***	-0,5
Agglo. de Paris	-2,7***	-5,3***	-2,3***	5,4***	-1,0*
Statut d'occupation du logement					
Propriétaire	Réf.	Réf.	Réf.	Réf.	Réf.
Locataire	5,4***	5,7***	3,5***	7,4***	-1,8***
Gratuit	5,0***	2,8	1,5	1,5	1,2
Autre	-4,2***	-14,1***	-4,5***	20,6***	-3,5*

Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 3, métropole.

Lecture : Parmi les 52 500 personnes composant les lots 1 à 3, la probabilité de n'avoir aucun numéro de téléphone pour joindre la personne échantillonnée est estimée à 14,7 % pour la situation de référence. Avoir 85 ans ou plus, augmente cette probabilité de 2,1 points par rapport à avoir entre 45 et 54 ans. L'effet est significatif au seuil de 5 %.

Note : \*\*\* : p-value < 0,001 ; \*\* : p-value < 0,01 ; \* : p-value < 0,05 ; lorsque le coefficient est manquant, c'est qu'il n'est pas significatif au seuil de 5 %.

On constate qu'il y a assez peu d'écart dans la disponibilité des coordonnées (mail, numéros de téléphone) entre les femmes et les hommes, si ce n'est concernant les numéros de mobile. Les hommes ont beaucoup plus fréquemment un numéro de mobile dans l'échantillon que les femmes. Cet écart peut s'expliquer de deux manières, dans le cadre d'un ménage composé d'un couple hétérosexuel :

d'une part, il est possible que ce soit plus souvent l'homme qui remplisse la déclaration de revenus et ainsi qu'il renseigne ses propres coordonnées dans la déclaration. D'autre part, cela tient peut-être aux déclarations communes. Il semblerait que, pour les déclarations avant 2014<sup>5</sup>, le numéro de mobile, en cas de déclaration commune, puisse avoir été rattaché systématiquement à l'homme qui était, par défaut, classé en déclarant 1 dans les fichiers fiscaux. Dans ce dernier cas, certains numéros de mobile rattachés à un homme dans Fidéli seraient en fait ceux de leur conjointe<sup>6</sup>.

Les jeunes ont plus fréquemment que les autres des coordonnées. Cependant, les plus jeunes, les 15-24 ans ont moins fréquemment un numéro de mobile renseigné, car ils sont moins souvent déclarant fiscal du ménage. Les plus âgées ont renseigné moins fréquemment un numéro de téléphone dans leur déclaration de revenus. L'effet est encore plus marqué pour les numéros de mobile.

Les personnes appartenant aux premiers déciles de niveau de vie ont moins de coordonnées renseignées que celles appartenant aux derniers déciles.

Il y a moins de coordonnées pour joindre les personnes seules que pour les couples sans enfant mais en cas de coordonnées renseignées, ce sont plus fréquemment des numéros de mobile, comme pour les familles monoparentales.

Il y a plus fréquemment des coordonnées renseignées parmi les personnes habitant dans les grandes villes.

Il est moins fréquent de disposer d'un numéro de téléphone ou d'un mail pour les locataires que pour les propriétaires. Les locataires ont, en revanche, plus fréquemment que les propriétaires un numéro de mobile dans les données.

En ce qui concerne l'enrichissement annuaire, la régression donne des résultats qui vont généralement dans le même sens que ceux de l'estimation de la probabilité de n'avoir aucun numéro de téléphone ni mail. L'enrichissement annuaire contribue donc à atténuer les écarts sociodémographiques dans la disponibilité de coordonnées dans Fidéli.

### **1.3. Évaluation de la qualité des informations de contact disponibles dans Fidéli**

La possibilité de joindre les enquêtés avec ces coordonnées est évaluée avec les parodonnées d'EpiCov. Cela concerne la qualité des adresses au travers des retours des lettres-avis en plis non distribués, celle des adresses mail et des numéros de téléphone avec les retours automatiques à l'envoi des mails et des SMS et la part des numéros de téléphone n'ayant jamais pu aboutir à un contact.

#### **1.3.1. Évaluation des adresses postales**

La classification des courriers en plis non distribués (PND) est faite par La Poste. Dans EpiCov, les adresses dataient du 1<sup>er</sup> janvier 2019 (Fidéli 2019) et la collecte a débuté le 2 mai 2020, il y avait donc un écart plutôt important entre le millésime de tirage et la collecte.

Sur l'ensemble des 350 000 lettres avis d'ouverture envoyées en métropole, 14 673 sont revenues en PND, soit 4,2 % des adresses. En ce qui concerne les relances, 20 822 PND ont été reçus sur 237 114 personnes relancées par courrier, soit un taux de PND de 8,8 %. Au total, 31 245 adresses ont eu au moins un PND, soit 8,9 % des adresses ; 4 250 adresses ont eu deux PND, soit 1,2 % des adresses. Les différences de taux de PND observés entre le courrier d'annonce et le courrier de relance peuvent tenir au fonctionnement des services de La Poste pendant le premier confinement où la distribution du courrier (et donc le retour en PND) a été particulièrement perturbée.

---

<sup>5</sup> À cette époque, le formulaire de déclaration des revenus ne permettait de renseigner qu'un numéro de téléphone et un mail sans qu'on puisse savoir à quel déclarant il était rattaché. Il semblerait que la Direction générale des Finances publiques classait l'homme, en cas de déclaration commune, systématiquement comme référent 1 et lui raccrochait le numéro de téléphone mobile déclaré.

<sup>6</sup> Une statistique qui pourrait étayer ce point est le fait qu'il y a beaucoup plus de femmes qui ont répondu avec le mobile d'une autre personne du ménage. Dans les lots 1 à 3, il y a 883 femmes qui ont répondu avec le mobile d'une autre personne du ménage contre 206 hommes.



L'estimation de la probabilité d'avoir eu au moins un retour en PND en fonction des caractéristiques sociodémographiques présentes dans Fidéli fait ressortir que le classement en PND est très lié au statut d'occupation du logement (Tableau 3). Les PND sont plus fréquents pour les locataires et les personnes logées gratuitement que pour les propriétaires. C'est également le cas pour les plus jeunes. Cela met bien en avant la corrélation entre déménagements et classement en PND et donc la plus grande difficulté à joindre par courrier postal les locataires et les jeunes pour leur proposer de participer aux enquêtes.

Tableau 3 : Corrélation entre les caractéristiques sociodémographiques et la probabilité d'avoir eu au moins un courrier retourné en PND

	Probabilité d'avoir eu au moins un courrier retourné en PND
Constante	3,5***
<b>Sexe</b>	
Femmes	Réf.
Hommes	
Inconnu	-1,3***
<b>Âge</b>	
15 - 24 ans	2,2***
25 - 34 ans	3,5***
35 - 44 ans	2,0***
45 - 54 ans	Réf.
55 - 64 ans	-0,9***
65 - 74 ans	-1,4***
75 - 84 ans	-1,4***
85 ans et plus	
<b>Quartier prioritaire de la politique de la ville</b>	
Oui	-0,9***
Non	Réf.
<b>Niveau de vie</b>	
1er décile	0,4***
2ème décile	
3ème décile	
4ème décile	
5ème décile	Réf.
6ème décile	
7ème décile	
8ème décile	0,4**
9ème décile	0,5***
10ème décile	0,5***
Inconnu	5,7***
<b>Situation familiale</b>	
Personne seule	0,5***
Couple sans enfant	Réf.
Couple avec enfant	-1,1***
Famille monoparentale	-0,7***
Ménage complexe	-0,5***

Pays de naissance	
France	Réf.
Etranger	0,2***
Inconnu	
Tranche d'unité urbaine (en nombre d'habitants)	
Rural	-0,6***
2 000-<10 000	-0,4***
10 000-<50 000	Réf.
50 000-<200 000	
200 000-<2 000 000	0,5***
Agglo. de Paris	-0,4***
Statut d'occupation du logement	
Propriétaire	Réf.
Locataire	10,5***
Gratuit	2,6***
Autre	8,6***

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 20, métropole.

*Lecture* : Parmi les 350 000 personnes composant les lots 1 à 20, la probabilité d'avoir un courrier retourné en pli non distribué est estimée à 3,5 % pour la situation de référence. Être locataire augmente cette probabilité de 10,5 points. L'effet est significatif au seuil de 0,1 %.

*Note* : \*\*\* : p-value < 0,001 ; \*\* : p-value < 0,01 ; \* : p-value < 0,05 ; lorsque le coefficient est manquant, c'est qu'il n'est pas significatif au seuil de 5 %.

### 1.3.2. Évaluation des adresses e-mails

Dans les cas où une adresse mail était disponible, un mail-avis a été envoyé, en plus de la lettre-avis. Au total, 246 019 mails ont été envoyés pour annoncer la collecte. L'envoi de ces mails a été lissé sur plusieurs jours pour éviter une réception en SPAM. Ipsos a des outils de classification des retours des mails. Ceux-ci indiquent qu'un total de 6 197 mails (2,5 %) sont arrivés en SPAM et 11 628 mails (4,7 %) ont, d'après ces retours, une forte probabilité de ne pas avoir été distribués.

### 1.3.3. Évaluation des numéros de mobiles via les retours de SMS

Un SMS avis a accompagné la lettre-avis et le mail-avis dans les cas où un numéro de mobile était disponible pour le ménage. Au total, 155 319 SMS ont ainsi été envoyés pour annoncer la collecte. L'envoi était également lissé sur plusieurs jours. D'après les outils de classification des retours d'Ipsos, 7 251 (4,7 %) ont une forte probabilité de ne pas avoir été distribués. Cela pourrait correspondre à des numéros erronés.

### 1.3.4. Évaluation des numéros de téléphone de tous types via le taux de contact

L'historique des appels a été exploité pour les trois lots ayant été ouverts dès le début de l'enquête à la collecte téléphonique. Le Tableau 4 décrit la part des numéros (par type) n'ayant pas permis d'aboutir à un contact avec le ménage. Si aucun « décroché » ne se produit, cela peut être en lien avec un numéro erroné ou avec un filtrage systématique des appels. Dans tous les cas, l'interview par téléphone s'avère impossible. Comme la collecte sur ces lots est multimode, certaines personnes ont pu répondre au questionnaire par internet ; il est probable que leur comportement vis-à-vis des appels téléphoniques soit différent des autres enquêtés. Le calcul est donc mené de deux façons, d'une part en incluant les répondants internet et d'autre part en les excluant.

Tableau 4 : Nombre de numéros de téléphone sans aucun décroché par type de numéro

	Nombre de numéros sans aucun décroché	% parmi ces types de numéro appelés	Nombre médian d'appels sur ces numéros sans aucun décroché	Nombre de numéro sans aucun décroché	% parmi ces types de numéro appelés	Nombre médian d'appels sur ces numéros sans aucun décroché
	Répondants internet exclus			Répondants internet inclus		
Mobile Fidéli de la pers. échant.	2 943	27,4%	10	4 890	33,3%	5
Mobile Fidéli d'une autre personne du ménage	1 957	33,7%	7	3 467	40,0%	4
Fixe géographique Fidéli	3 506	34,5%	6	5 416	40,7%	5
« 09 » Fidéli	220	56,6%	10	309	60,7%	6
Fixe géographique Annuaire	1 309	32,2%	16	1 698	35,9%	13
« 09 » Annuaire	525	57,1%	17	652	59,6%	16
Mobile Annuaire	28	31,8%	16,5	49	39,2%	10
Numéro « terrain »	159	11,9%	10	276	16,8%	5

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 3, métropole.

*Lecture* : Au sein des lots 1 à 3, 2 943 numéros de téléphone de type mobile Fidéli de la personne échantillonnée ont été appelés sans que cela ne mène à un « décroché ». Cela représente 27,4 % des numéros de mobile Fidéli de personnes échantillonnées. Le nombre médian d'appels pour tenter de les joindre a été de 10 appels. Ces résultats s'entendent répondants internet exclus de l'échantillon.

La part des numéros de téléphone qui ne conduisent pas à un « décroché » est importante : si on exclut les répondants internet de l'échantillon, elle oscille entre 27,4 % et 57,1 %, mis à part les numéros de téléphone « terrain », c'est-à-dire ceux qui ont été donnés par les enquêtés au cours de la collecte (*via* la hotline ou un enquêteur), pour lesquels le taux est, comme attendu, plus faible (11,9 %). Elle est la plus élevée parmi les « 09 » (qu'ils soient issus de Fidéli ou de l'annuaire), puis les fixes géographiques, les mobiles Fidéli d'une autre personne du ménage, les mobiles annuaire, et enfin les mobiles Fidéli de la personne échantillonnée puis les numéros « terrain ».

La base de données Fidéli est donc très riche en coordonnées facilitant les collectes d'enquêtes. Il existe des écarts sociaux dans la disponibilité de coordonnées de contact : les plus âgés, les plus pauvres, les personnes nées à l'étranger, les locataires ont moins fréquemment des coordonnées de contact (mail, numéro de téléphone) renseignées. De ce point de vue, l'enrichissement annuaire vient limiter ces écarts en apportant un numéro de téléphone plus fréquemment parmi les plus âgés et les personnes appartenant aux premiers déciles de niveau de vie. Beaucoup de numéros de téléphone n'ont pas permis de réaliser l'enquête par téléphone car ils ne conduisent pas à un contact avec les enquêtés. La partie suivante s'intéresse plus en détails à la rentabilité de ces informations de contact pendant la collecte.

## 2. Comment mobiliser ces données de contact dans différents contextes de collecte ?

### 2.1. Collecte sur Internet : multiplier les contacts pour convaincre

Dans cette partie, l'analyse porte sur les lots qui ont été ouverts uniquement à une collecte par internet (CAWI), soit les lots 5 à 20.

Les taux de réponse par internet sont très différents selon les moyens de contact à disposition pour

joindre les personnes échantillonnées (Tableau 5). Le fait de ne pas recevoir de courrier postal semble très pénalisant pour la collecte. Plus il y a de modes de contact, plus il y a de possibilités pour relancer les personnes (courrier de relance, mail, SMS), plus le taux de réponse est élevé. Le taux de réponse est plus élevé lorsque ces coordonnées de contact appartiennent à la personne échantillonnée plutôt qu'à une autre personne du ménage. Il y a tout de même une exception portant sur les numéros de mobile qui est sans doute liée au fait que les hommes ont plus fréquemment un numéro de mobile dans Fidéli et qu'ils répondent moins fréquemment à l'enquête. On remarque que 15,6 % des personnes qui n'ont reçu qu'un courrier postal ont répondu à l'enquête par internet en faisant l'effort de saisir l'adresse url sur internet et d'utiliser leurs identifiant et mot de passe. A l'inverse, sans courrier postal (PND) mais avec l'adresse mail de la personne échantillonnée et son numéro de mobile, le taux de réponse atteint 20 %.

Tableau 5 : Taux de réponse dans la collecte purement Internet en fonction des coordonnées de contact renseignées

Coordonnées disponibles	Rattachement des coordonnées	Taux de réponse (%)
<b>Courrier distribué</b>		
Mail uniquement	Personne échantillonnée	42,3
	Autre personne du ménage	38,5
Mobile uniquement	Personne échantillonnée	27,9
	Autre personne du ménage	33,9
Mail et mobile	Personne échantillonnée	53
	Autre personne du ménage	46,1
Ni mail, ni mobile	-	15,6
<b>Courrier non distribué (PND)</b>		
Mail uniquement	Personne échantillonnée	10,4
	Autre personne du ménage	6,4
Mobile uniquement	Personne échantillonnée	5,9
	Autre personne du ménage	3,9
Mail et mobile	Personne échantillonnée	20,3
	Autre personne du ménage	11,8
Ni mail, ni mobile	-	0,3
Ensemble		35,3

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 5 à 20, métropole.

*Lecture* : Le taux de réponse des personnes ayant reçu un courrier postal mais n'ayant pas d'adresse mail, ni de numéro de téléphone à disposition pour envoyer des SMS est 15,6 %. Il est plus faible que lorsqu'une adresse mail est en plus disponible pour la personne échantillonnée (42,3 %).

*Note* : La catégorie « mail et numéro de mobile de la personne échantillonnée » concerne le cas où il y a au moins un mail ou/et un mobile de la personne échantillonnée (il peut y en avoir un également pour une autre personne du ménage). La catégorie « mail et numéro de mobile d'une autre personne du ménage » concerne le cas où il y a ces coordonnées pour une autre personne du ménage mais pas pour la personne échantillonnée.

Ces résultats peuvent correspondre à l'efficacité d'un protocole de relance multi-support comme à un effet de sélection. Comme on l'a vu dans la partie précédente, la présence de coordonnées de contact est liée à des caractéristiques sociales qui sont elles-mêmes liées à des propensions à répondre différentes à l'enquête.

Pour essayer de contrôler les effets de sélection et estimer un effet « relance », une régression modélisant la probabilité de répondre en fonction des caractéristiques sociodémographiques des personnes et des modes de contact à disposition a été estimée. Les variables sociodémographiques introduites dans le modèle sont les mêmes que celles des Tableau 2 et Tableau 3. Le Tableau 6 reporte les coefficients estimés rattachés aux informations de contact. Cette régression confirme l'importance

du courrier postal avec une probabilité de réponse qui baisse de 20 points environ sans courrier postal, et cela de manière assez uniforme quels que soient les autres moyens de contact à disposition pour informer la personne de l'enquête et la relancer (mail, SMS). Par rapport au Tableau 5, il semblerait qu'une fois les caractéristiques sociodémographiques contrôlées, l'effet de l'absence de courrier postal soit prédominant sur les autres modes de contact. Son absence ne semble pas pouvoir être facilement compensée par un mail ou un SMS. Une fois le courrier postal reçu, les relances par mail semblent plus rentables en termes de collecte que les relances par SMS, tout en étant moins coûteuses. Les relances mail et SMS ont toutefois suivi des calendriers légèrement différents, ce qui peut générer des effets de sélection dans les populations relancées selon ces différents modes. En tous cas, la régression met bien en évidence l'importance de la pluralité des modes de contacts (mail et SMS). L'effet est d'autant plus fort que les coordonnées sont celles de la personne échantillonnée, mais il reste important lorsque ces coordonnées sont celles d'une autre personne du ménage, montrant bien l'intérêt de passer également par cet intermédiaire pour obtenir des réponses à l'enquête.

En ce qui concerne la relance par SMS, une fois les caractéristiques sociodémographiques contrôlées, le numéro d'une autre personne du ménage ne semble pas plus rentable que celui de la personne échantillonnée. Mais l'efficacité entre les deux est très proche, beaucoup plus qu'en ce qui concerne les relances par mail. Cela peut être dû à un usage plus répandu du SMS que du mail pour communiquer dans un ménage et donc une plus grande facilité à se transmettre l'information relative à l'enquête via SMS. L'effet estimé des moyens supplémentaires d'annonce et de relance est fort (+ 26,1 points de taux de réponse si on peut envoyer des mails d'annonce et de relance avec l'adresse de la personne échantillonnée) mais au-delà de l'effet relance, cela traduit aussi la simplification du processus de réponse. En effet, dans chaque mail et SMS était fourni un lien direct et personnalisé qui permettait d'accéder directement au questionnaire, ce qui est impossible avec un courrier postal qui nécessite de saisir l'URL, l'identifiant et le mot de passe du courrier pour se connecter au questionnaire Internet. Les variables de contrôle dans la régression étaient assez nombreuses du fait de la richesse de Fidéli, il n'est cependant pas exclu qu'il reste de la sélection sur d'autres caractéristiques qui sont inobservables dans l'estimation de l'effet « relance ».

Tableau 6 : Estimation de la probabilité de répondre par Internet en fonction des moyens de contact disponibles pour inciter à participer à l'enquête

Coordonnées disponibles	Rattachement des coordonnées	Probabilité de répondre
Constante		27,9***
<b>Courrier distribué</b>		
Mail uniquement	Personne échantillonnée	26,1***
	Autre personne du ménage	16,1***
Mobile uniquement	Personne échantillonnée	18,1***
	Autre personne du ménage	17,2***
Mail et mobile	Personne échantillonnée	37,7***
	Autre personne du ménage	26,4***
Aucune coordonnée	-	Réf.
<b>Courrier non distribué (PND)</b>		
Mail uniquement	Personne échantillonnée	-21,5***
	Autre personne du ménage	-22,8***
Mobile uniquement	Personne échantillonnée	-21,2***
	Autre personne du ménage	-24,3***
Mail et mobile	Personne échantillonnée	-18,4***
	Autre personne du ménage	-21,0***
Aucune coordonnée	-	-27,2***

Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 5 à 20, métropole.

**Lecture :** Au sein des lots 5 à 20, la probabilité de réponse est 27,9 % pour la situation de référence. Une fois les variables sociodémographiques contrôlées, disposer d'une adresse mail pour la personne échantillonnée sur laquelle la personne est relancée pour participer à l'enquête augmente la probabilité de répondre de 26,1 points par rapport à ne disposer que de son adresse postale pour l'informer de l'enquête et lui donner les identifiants et mot de passe de connexion. Cet effet est significatif au seuil de 0,1 %.

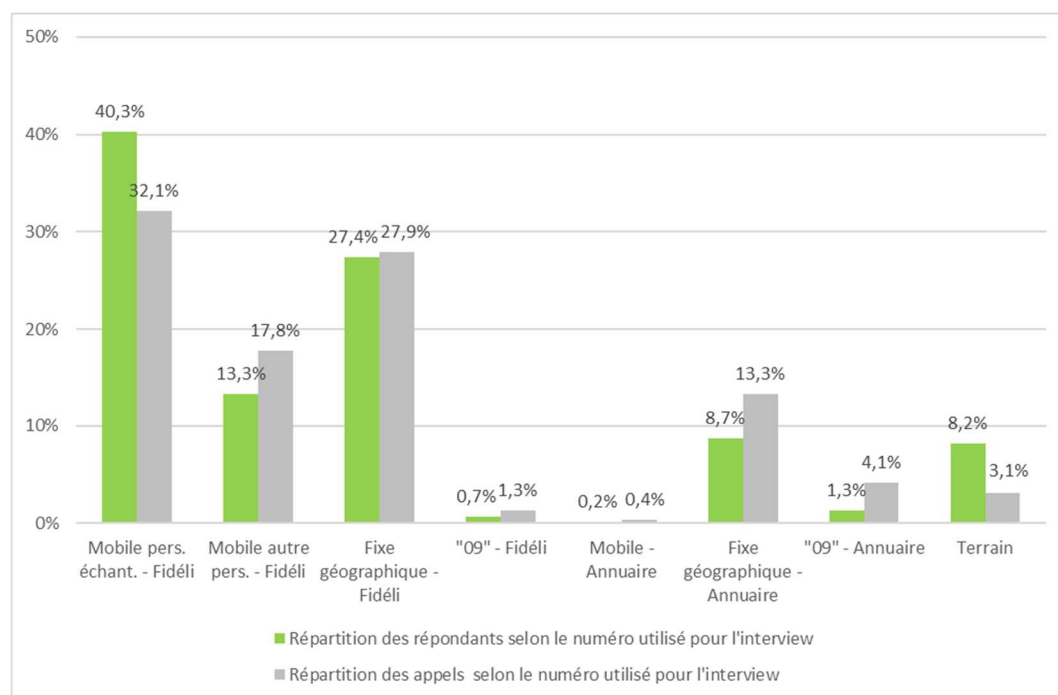
**Note :** \*\*\* : p-value < 0,01 ; les variables de contrôle et la situation de référence sont celles des Tableau 2 et Tableau 3.

## 2.2. Collecte par téléphone : mieux cibler pour limiter l'effort de collecte

L'analyse porte maintenant sur les lots 1 à 3 (collecte CAWI et CATI) dès le début du terrain et a pour objectif de tirer des enseignements des données concernant la collecte CATI.

La partie 1 a déjà mis en avant qu'une part importante des numéros de téléphone n'avait finalement pas été utile à la collecte car ceux-ci ne conduisaient pas à un contact avec le ménage. En miroir, on peut s'intéresser au nombre de questionnaires obtenus par type de numéro (Graphique 1). C'est à partir du mobile de la personne échantillonnée que le plus grand nombre d'interviews a été obtenu (40,3 %). Les fixes géographiques issus de Fidéli ont également permis d'avoir un grand nombre d'interviews (27,4 %). L'enrichissement annuaire permet d'augmenter de façon importante le nombre de répondants, tout comme les appels sur les mobiles d'une autre personne du ménage : 10,2 % des répondants ont été enquêtés via un numéro annuaire et 13,3 % via le mobile d'une autre personne du ménage.

Graphique 1: Répartition des répondants de la collecte téléphonique et des appels par type de numéro de téléphone



**Source :** Inserm-Drees, EpiCov, vague 1.

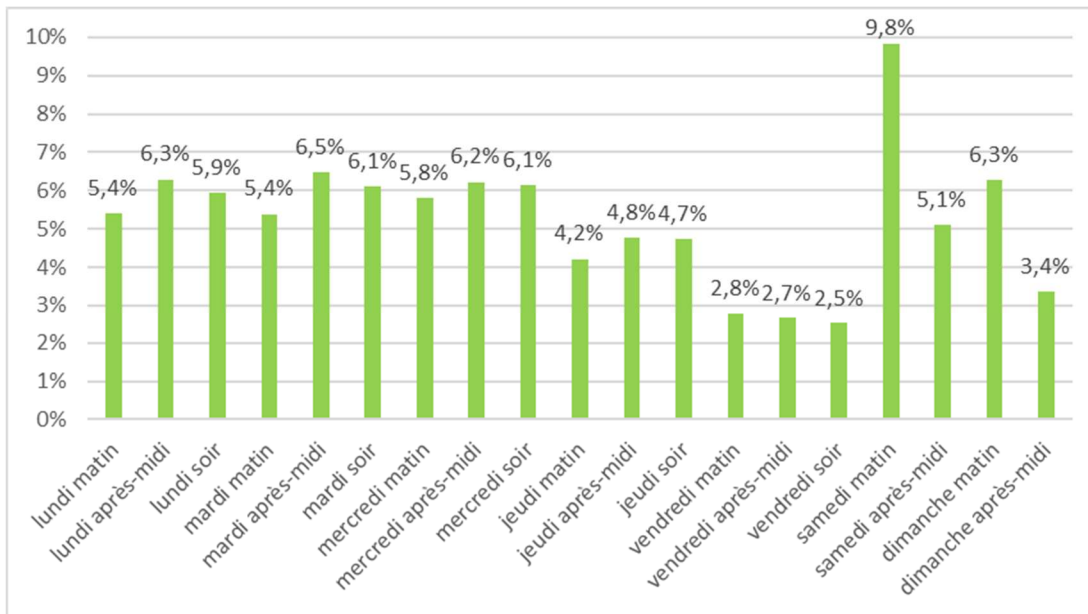
**Champ :** lots 1 à 3, métropole.

**Lecture :** Au sein des lots 1 à 3, 40,3 % des répondants par téléphone l'ont été par l'intermédiaire d'un numéro de téléphone mobile de la personne échantillonnée issu de Fidéli alors que 32,1 % des appels ont été faits sur ce type de numéro.

Concernant les moments des appels conduisant le plus fréquemment à réaliser un questionnaire, il s'agit du samedi matin (9,8 % des interviews). En début de semaine, la part des répondants selon les moments est assez similaire. Elle baisse le jeudi et le vendredi et remonte le week-end<sup>7</sup>.

<sup>7</sup> Dans le contexte de la vague 1 d'EpiCov, il avait été permis à Ipsos de faire travailler leurs enquêteurs le dimanche.

Graphique 2: Nombre de répondants par moment des appels



Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 3, métropole.

Lecture : Au sein des lots 1 à 3, 5,4 % des questionnaires collectés par téléphone l'ont été le lundi matin.

Note : matin : avant 14h ; après-midi : entre 14h et 17h ; soir : après 17h.

Les résultats des différents numéros de téléphone, en termes de collecte, dépendent de l'intensité avec laquelle ils ont été appelés et des moments de ces appels. L'exploitation de l'historique des appels apparié à l'échantillon Fidéli (encadré 2) permet d'étudier la rentabilité des appels par types de numéro en contrôlant le moment de ces appels et les caractéristiques sociodémographiques des personnes. Une régression modélise la probabilité de réaliser un questionnaire. Celle-ci est ensuite décomposée en deux régressions : la modélisation de la probabilité d'avoir un « décroché » et celle de la probabilité de réaliser un questionnaire sachant qu'il y a eu un « décroché ». Les résultats sont présentés dans le Tableau 7 (colonnes 2 à 4), seuls les coefficients associés aux types de numéro de téléphone et au retour en PND du courrier sont reportés.

Tableau 7 : Rentabilité des numéros de téléphone en termes de collecte

	Probabilité qu'un appel aboutisse à un "décroché"	Probabilité que l'appel aboutisse à une interview sachant qu'il y a eu un "décroché"	Probabilité que l'appel aboutisse à une interview	Probabilité que l'appel aboutisse à interview (modélisation biprobit)
Constante	16,9***	25,7***	4,3***	4,3***
<b>Type de numéro</b>				
Mobile pers. échant. - Fidéli	4,4***	9,5***	3,2***	3,0***
Mobile autre pers. - Fidéli	1,8***	-3,6***		
Fixe géographique - Fidéli	Réf.	Réf.	Réf.	Réf.
« 09 » - Fidéli	-5,8***		-1,4**	-1,3**
Mobile Annuaire	-3,8**		-1,7*	-1,6*
Fixe Annuaire	-1,3***	-5,5***	-1,3***	-1,2***
« 09 » Annuaire	-7,8***		-2,4***	-2,3***
Numéro « terrain »	12,0***	21,5***	9,8***	8,8***

PND				
Non	Réf.	Réf.	Réf.	Réf.
Oui	-6,5***	-7,2***	-2,4***	-2,4***

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 3, métropole.

*Lecture* : Au sein des lots 1 à 3, la probabilité qu'un appel aboutisse à un décroché est 16,9 % pour la situation de référence. Appeler sur un mobile Fidéli de la personne échantillonnée plutôt que sur un fixe géographique Fidéli augmente cette probabilité de 4,4 points. Le coefficient est significatif au seuil de 0,01 %.

*Note* : \*\*\* : p-value < 0,001 ; \*\* : p-value < 0,01 ; \* : p-value < 0,05 ; les variables de contrôle et la situation de référence sont celles des Tableau 2 et Tableau 3 avec en plus les moments d'appels tels que figurant sur le Graphique 2.

Lorsque le coefficient est manquant, c'est qu'il n'est pas significatif au seuil de 5 %.

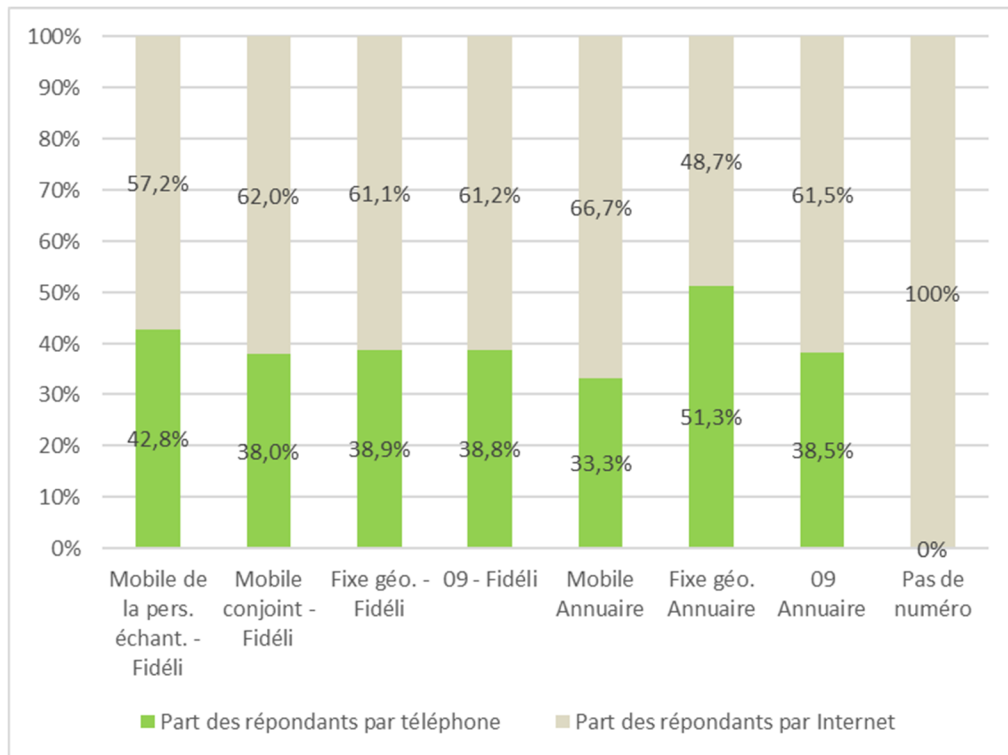
Les numéros de téléphone « terrain » apparaissent comme les plus rentables à tout point de vue. Ce sont des numéros récupérés suite à une démarche active des personnes, témoignant d'un accueil favorable à l'enquête et d'une certaine adhésion. En dehors de ce type de numéro, ce sont les appels sur les mobiles issus de Fidéli des personnes échantillonnées qui conduisent à la plus forte probabilité de réaliser un questionnaire. Cela est dû à la fois à une augmentation des « décrochés » par rapport aux fixes géographiques issus de Fidéli et à une hausse de l'acceptation de l'interview une fois qu'un contact a eu lieu, sans doute parce que l'enquêteur est directement en contact avec la personne à enquêter. À cet égard, l'écart d'efficacité en termes de collecte entre les mobiles des personnes échantillonnées et ceux d'une autre personne du ménage est important. La probabilité de réaliser un questionnaire sachant qu'il y eu un contact est plus faible avec les mobiles issus de Fidéli d'une autre personne du ménage qu'avec les fixes géographiques issus de Fidéli car l'enquêteur a, dans ce cas, toujours une autre personne au bout du fil, qui peut ne pas être avec la personne échantillonnée, faire barrage, etc. En ce qui concerne la différence de taux de « décroché », elle peut être mise en lien avec la réception de la lettre-avis qui était adressée à la personne échantillonnée spécifiquement et prévenait de l'appel. Finalement, les mobiles d'une autre personne du ménage ne sont pas plus rentables qu'un fixe géographique issu de Fidéli pour réaliser un questionnaire. Les numéros commençant par « 09 » conduisent moins fréquemment que les numéros fixes géographiques issus de Fidéli à un « décroché ». Cela corrobore l'idée que ces numéros sont sans doute moins utilisés et ne sont donc pas efficaces pour la collecte. En effet, seulement 2,0 % des questionnaires ont été réalisés avec ce type de numéro. Les fixes géographiques issus de l'enrichissement par annuaire sont moins rentables que les fixes géographiques issus de Fidéli à tout point de vue : « décroché » et questionnaire sachant qu'il y a eu un contact. Cela questionne la qualité de l'enrichissement. Il est probable que certains numéros récupérés par enrichissement ne mènent pas au bon ménage, car on observe une probabilité de questionnaire après contact largement inférieure au fixe géographique Fidéli. Ceci étant, leur apport quantitatif dans la collecte est important (8,7 % des interviews). Enfin l'impact d'un PND est également négatif dans une collecte téléphonique, mais plus faible que dans une collecte par Internet.

Ces modèles ont été estimés en gardant dans l'échantillon les répondants par Internet (les lots étaient multimodes) mais en les considérant comme non-répondants par téléphone (de fait, ils ont répondu par Internet). Les mêmes modèles ont été estimés en retirant de l'historique des appels les données des répondants par Internet, les résultats sont très proches. Un autre test de robustesse a été mené en modélisant, sur ces lots, conjointement la probabilité de répondre par internet et la probabilité de répondre par téléphone à l'aide d'un biprobit. Les résultats présentés également dans le Tableau 7 sont très proches des précédents.

Enfin, on peut plus largement s'interroger sur le choix du mode de collecte, en lien avec les informations disponibles dans l'échantillon. Le Graphique 3 présente la répartition des répondants selon le mode de réponse par type de numéro. On constate que les personnes disposant d'un fixe géographique issu de l'enrichissement annuaire sont en proportion plus nombreuses à répondre par téléphone alors même que ces numéros sont moins rentables que les fixes géographiques issus de Fidéli. Cela met à nouveau en avant l'intérêt de l'enrichissement annuaire. Il permet d'obtenir des questionnaires d'une population qui certainement n'aurait pas répondu par Internet et qui aurait ainsi été non-répondante à EpiCov.



Graphique 3: Répartition des répondants par mode selon le type de numéro de téléphone disponible



*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 3, métropole.

*Lecture* : Au sein des lots 1 à 3, les répondants à l'enquête disposant d'un numéro de fixe géographique Annuaire sont 51,3 % à avoir répondu par téléphone et 48,7 % à avoir répondu par internet.

Le courrier postal semble toujours très important pour inciter à la participation à l'enquête, même avec le développement des mails et des SMS. Cela est d'autant plus vrai pour une collecte par Internet où il n'y a pas d'enquêteur pour motiver à participer à l'enquête. Pour une collecte par Internet, solliciter selon différents canaux (courrier, mail, SMS) est de nature à augmenter le taux de réponse. Les relances auprès des autres personnes du ménage semblent, de ce point de vue, aider à la collecte par internet, notamment *via* les SMS.

Concernant la collecte CATI, les numéros issus de l'enrichissement annuaire permettent d'assurer une part importante de la collecte, environ 10 % des interviews. Ils permettent de toucher une population a priori moins encline à répondre par internet donc s'avèrent utiles dans une approche de complémentarité des modes de collecte. Les numéros de type « 09 » ne semblent pas très rentables pour la collecte. Ils sont donc à mobiliser uniquement si aucun autre type de numéro n'est disponible. Enfin, il serait intéressant de questionner la rentabilité des numéros de mobile (de la personne échantillonnée ou d'une autre personne du ménage) dans un autre contexte que celui des restrictions liées au premier confinement. Autrement dit, sont-ils aussi rentables dans une vie moins contrainte où le mobile sonnerait plus fréquemment alors que l'on n'est pas chez soi, dans les transports, etc. ?

### 3. Protocole monomode versus multimodes : l'apport d'un volet téléphonique

Au-delà de la disponibilité des données de contact dans Fideli et de leur mobilisation au sein d'une collecte par téléphone ou internet, se pose la question de l'impact du protocole de chacun des lots sur la participation à l'enquête et sur le profil des répondants. Plus spécifiquement, cette partie vise à mettre en évidence les différences en termes de taux de participation entre les lots ayant suivi un protocole monomode, les lots 5 à 20 dans lesquels les échantillonnés ne pouvaient répondre que par internet (CAWI), et les lots ayant suivi un protocole multimode, ayant pu également bénéficier d'une collecte par téléphone (CATI). La répartition aléatoire des individus entre les deux protocoles, indépendamment des coordonnées disponibles dans Fidéli, permet d'évaluer l'apport d'un protocole

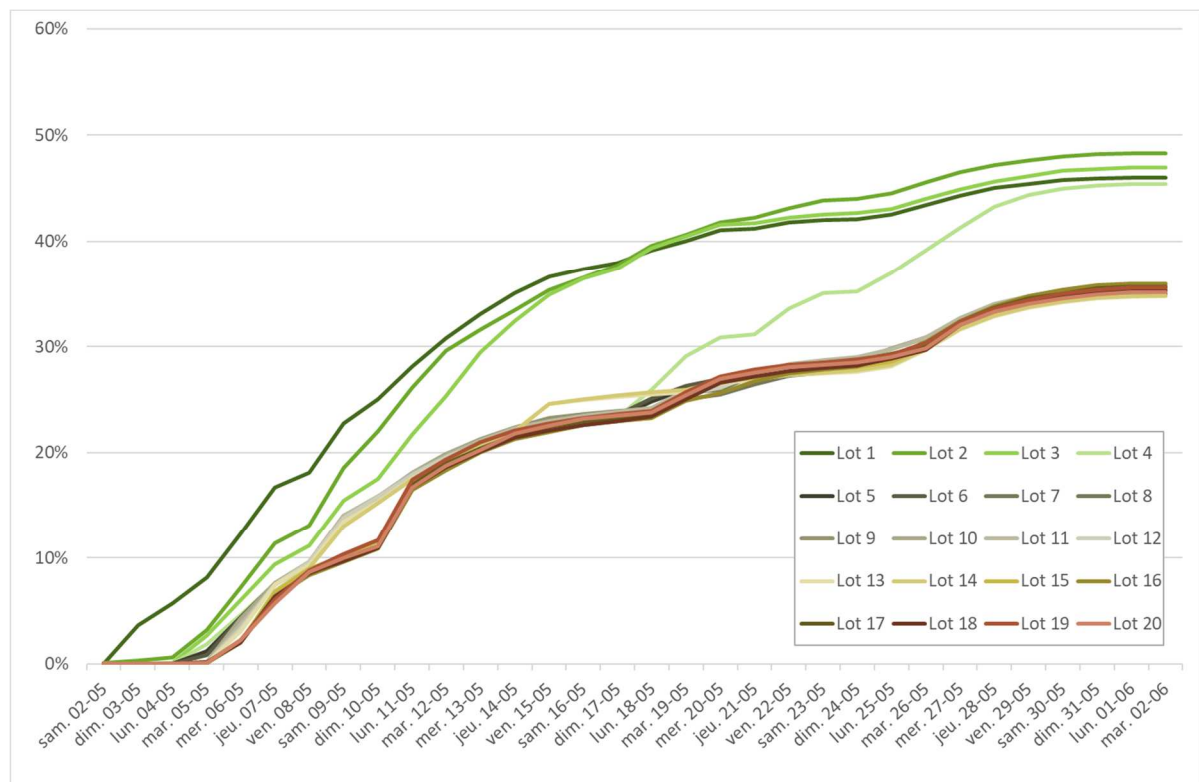
multimode. L'analyse de la structure des répondants dans chaque protocole montre que l'ajout du CATI permet de capter des répondants aux profils spécifiques.

Ces deux premières étapes permettent d'étudier le phénomène de non-réponse totale selon le type de protocole prévu. Dans un troisième temps le phénomène d'abandon en cours de collecte sera analysé, en se concentrant davantage sur le mode de réponse effectivement utilisé que sur le type de protocole. Ce sera l'occasion d'identifier les déterminants de l'abandon en cours de collecte selon le mode de réponse utilisé (internet ou téléphone) et participera à appuyer nos recommandations quant à l'intérêt d'introduire un volet téléphonique dans les enquêtes CAWI.

### 3.1. Taux de réponse

Le taux de participation pour l'enquête EpiCov est de 37,6 % pour l'ensemble des lots métropolitains, avec une nette différence selon le protocole mis en œuvre. Il atteint 46,7 % pour les individus des lots multimodes contre 35,3 % pour le reste de l'échantillon, soit une différence de plus de 10 points. Ce taux a suivi une évolution comparable au fil de la collecte entre les différents lots appartenant à un même protocole (Graphique 4), même si pour les lots multimodes, des différences de calendrier dans l'ouverture du téléphone ont causé des variations.

Graphique 4 : Évolution du taux de réponse par lot



Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 20, métropole.

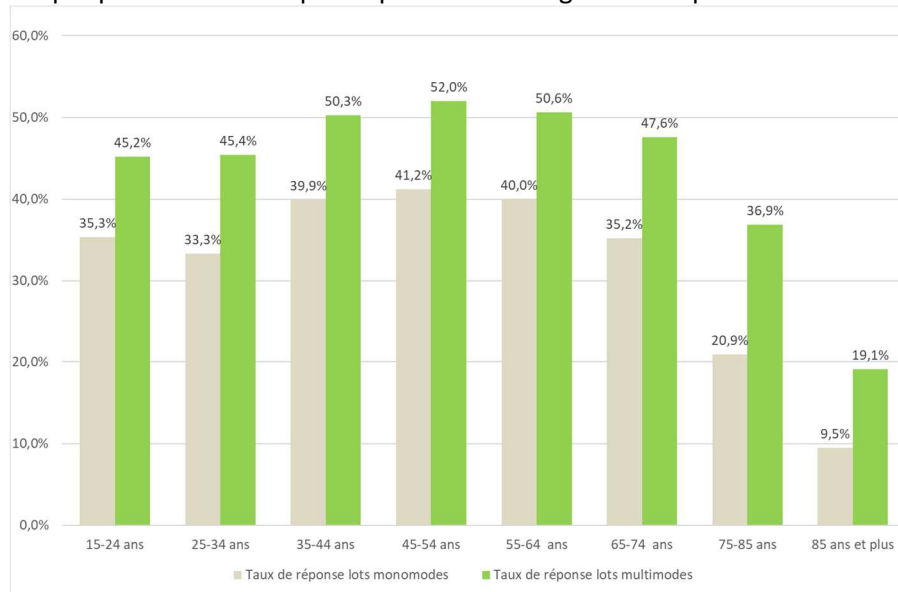
### 3.2. Profils des répondants

Pour évaluer si un protocole multimode CAWI-CATI et un protocole monomode CAWI conduisent à des profils de répondants différents, les taux de réponse de différentes sous-populations ont été comparés pour les deux protocoles. Autrement dit, le gain qu'a permis l'introduction du CATI dans les lots concernés en fonction des différentes caractéristiques de la population étudiée a été mesuré. Pour l'ensemble des lots, le taux de réponse des lots exploités en multimode est 32,2 % plus élevé que le taux de réponse des lots ayant suivi un protocole monomode.

En termes d'âge, les écarts les plus importants se retrouvent pour les tranches d'âges les plus élevées qui semblent avoir bénéficié le plus de l'ouverture du CATI (Graphique 6). L'introduction du téléphone a ainsi permis de multiplier par deux les taux de réponse des plus âgés, qui restent toutefois les plus bas de l'échantillon.

En termes de niveau de vie (Graphique 7), on constate que c'est parmi les premiers déciles que le gain apporté par le multimode est le plus conséquent. Le taux de participation est supérieur d'au moins 40 % pour les lots multimodes par rapport aux lots monomodes quand on considère les cinq premiers déciles de niveaux de vie. A partir du 8ème décile, si les lots multimodes connaissent toujours des meilleurs taux de réponse que les lots internet, ce gain est moindre que celui observé sur l'ensemble de l'échantillon.

Graphique 5 : Taux de réponse par tranche d'âge selon le protocole

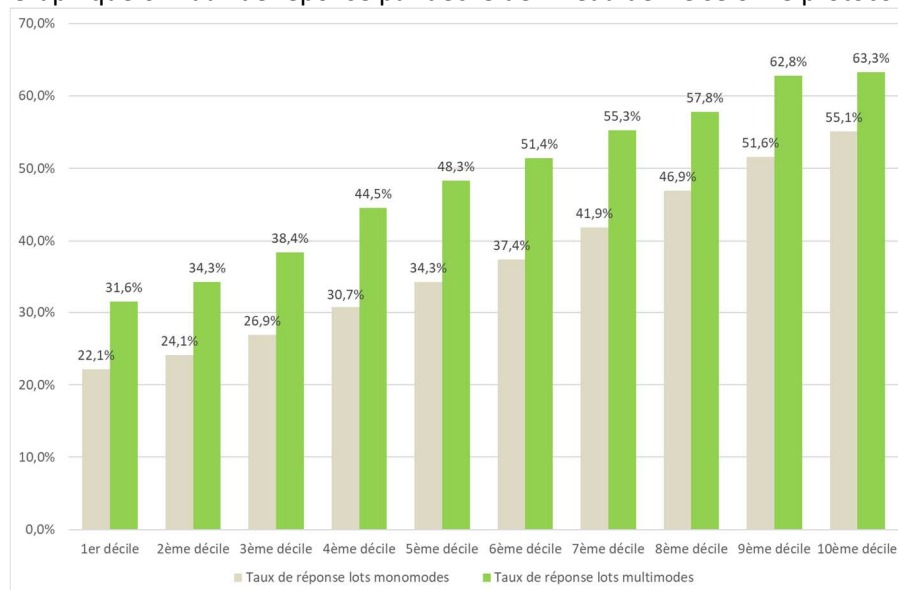


**Source :** Inserm-Drees, EpiCov, vague 1.

**Champ :** lots 1 à 20, métropole.

**Lecture :** Le taux de réponse pour les 15-24 ans des lots ayant suivi un protocole monomode est de 35,3%.

Graphique 6 : Taux de réponse par décile de niveau de vie selon le protocole



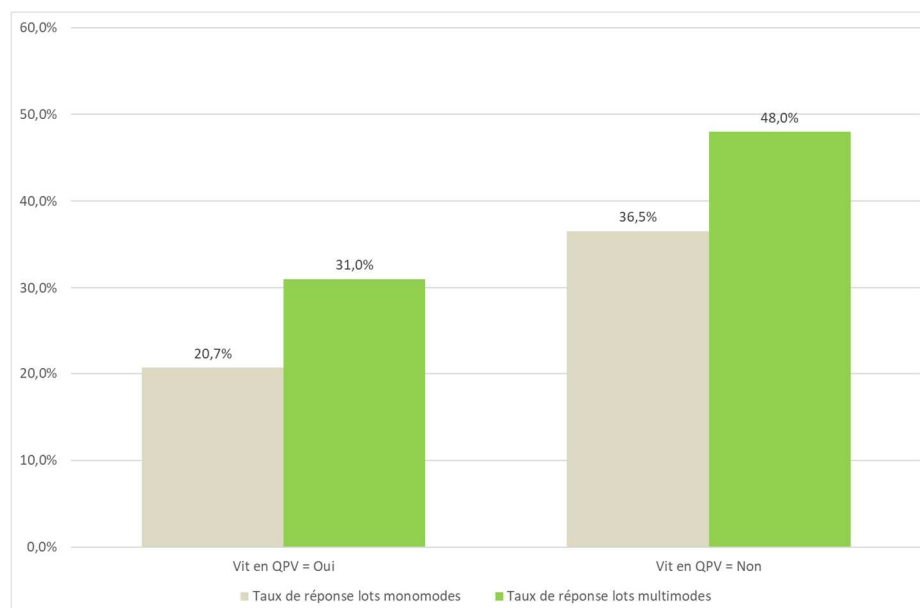
**Source :** Inserm-Drees, EpiCov, vague 1.

**Champ :** lots 1 à 20, métropole.

**Lecture :** Le taux de réponse pour les personnes appartenant au 1<sup>er</sup> décile de niveau de vie des lots ayant suivi un protocole monomode est de 22,1 %.

Pour les ménages vivant en quartier prioritaire de la politique de la ville (QPV), le taux de réponse est de 50 % supérieur dans les lots multimodes par rapport aux lots internet (Graphique 8). Ce résultat est cohérent avec celui observé sur les niveaux de vie dans le graphique précédent.

Graphique 7 : Taux de réponse et fait de vivre en QPV selon le protocole



*Source : Inserm-Drees, EpiCov, vague 1.*

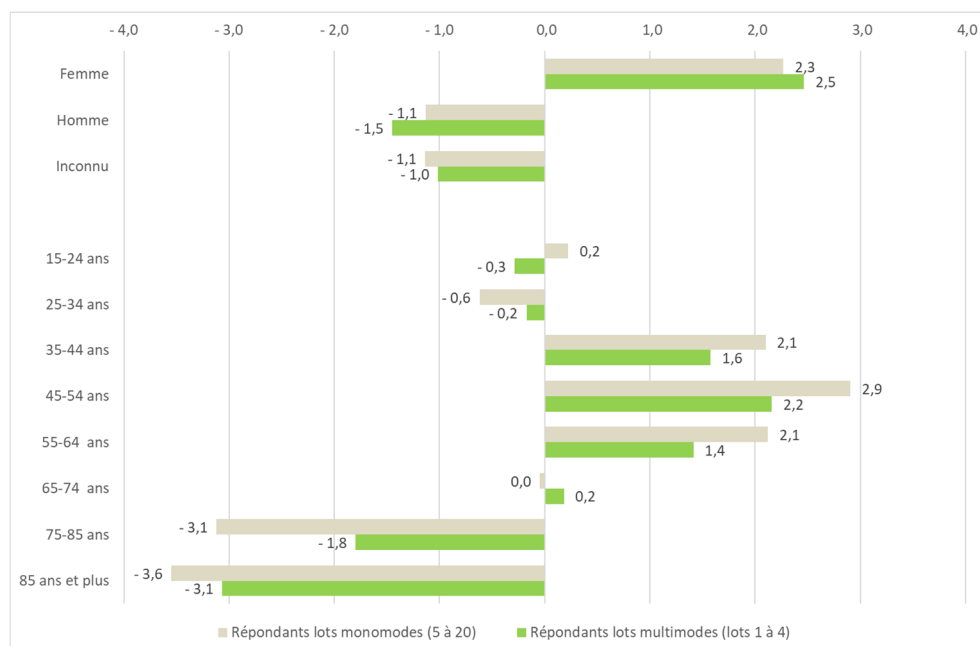
*Champ : lots 1 à 20, métropole.*

*Lecture : Le taux de réponse pour les personnes vivant en QPV des lots ayant suivi un protocole monomode est de 20,71 %.*

Il est également intéressant de souligner que les différences de taux de réponse entre les lots multimodes et les lots internet sont particulièrement marquées selon les coordonnées dont on disposait sur la personne échantillonnée dans le fichier Fidéli. En particulier et sans surprise, les personnes sans coordonnées (autre que l'adresse postale issue de Fidéli) ou ne disposant que d'une adresse mail connaissent des taux de participation équivalents dans les deux types de lots. Le taux de participation pour les personnes sans adresse mail ni numéro de téléphone est ainsi de 11,9 % dans les lots 1 à 4 et de 12,4 % dans les lots 5 à 20. Pour les individus disposant uniquement d'une adresse mail, les taux de réponse sont équivalents dans les lots multimodes (30,5 %) et dans les lots internet (31,7 %). Il est ainsi logique de constater que le gain apporté par l'introduction du volet téléphonique concerne uniquement les personnes qui disposent d'un numéro de téléphone dans Fidéli, particulièrement quand elles n'ont spécifiquement qu'un numéro de téléphone comme moyen de contact disponible, en plus de leur adresse postale ; le taux de réponse fait alors plus que doubler dans les lots multimodes. Le fait de disposer d'un numéro de téléphone dans Fidéli ou non, ajouté ou non d'une adresse mail, étant associé à des profils sociodémographiques particuliers, étudier la distribution des coordonnées de contact dans Fidéli comme c'est fait dans la première partie permet également d'anticiper l'apport de l'introduction du CATI.

C'est finalement parmi les échantillonnés dont le taux de réponse est le plus bas dans les lots monomodes internet que le gain de taux de réponse induit par l'ajout du CATI dans le protocole multimode est le plus important. Il en résulte une structure des répondants des lots multimodes plus proche de la structure de l'échantillon initial que ne l'est celle des répondants des lots monomodes, notamment pour l'âge où la surreprésentation des tranches d'âge intermédiaires (entre 35 et 64 ans) et la sous-représentations des tranches les plus élevés (au-delà de 75 ans) sont en partie corrigées dans le protocole multimode (Graphique 8). Les mêmes résultats s'observent pour les déciles de niveau de vie (en particulier sur le haut de la distribution), le fait de vivre en QPV et le lieu de naissance (Graphique 9), mais pas pour le sexe de la personne enquêtée.

Graphique 8 : Écarts en points de pourcentage entre la structure des répondants et celle de l'échantillon initial selon le sexe et l'âge

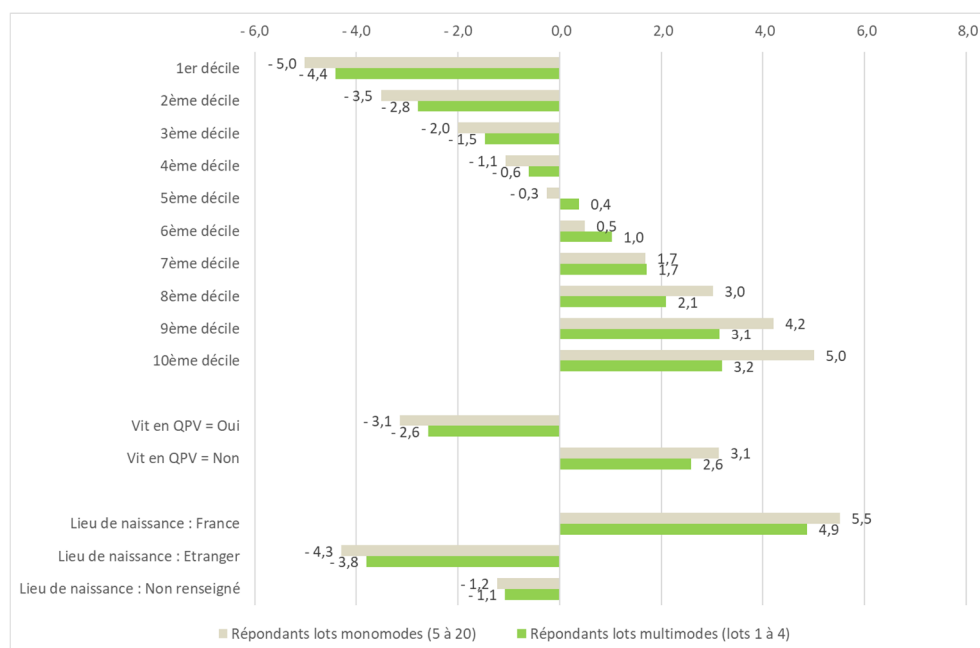


Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 20, métropole.

Lecture : La part des femmes dans les répondants des lots monomodes est supérieure de 2,3 points à leur part dans l'échantillon initial.

Graphique 9 : Écarts en points de pourcentage entre la structure des répondants et celle de l'échantillon initial selon le niveau de vie, le fait de vivre en QPV et lieu de naissance



Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 20, métropole.

Lecture : La part des répondants appartenant au 1<sup>er</sup> décile de niveau de vie dans les lots monomodes est inférieure de 5 points à leur part dans l'échantillon initial.

Ces résultats montrent que l'introduction d'une phase de collecte par téléphone permet d'améliorer la diversité des profils sociodémographiques des répondants par rapport à une collecte monomode Internet et améliore ainsi la représentativité de l'enquête.

### 3.3. Abandons

L'analyse suivante s'intéresse aux abandons selon les modes de collecte (Internet versus téléphone) sur la première page du questionnaire<sup>8</sup> ou plus loin dans le questionnaire. Comprendre les abandons est important car ils concernent des personnes qui ont fait l'effort de se connecter au site de collecte ou de passer la phase de contact avec l'enquêteur.

#### 3.3.1. État des lieux de l'abandon dans EpiCov

Selon notre définition extensive des abandons, ce sont 9,6 % des personnes échantillonnées dans EpiCov qui ont abandonné. Parmi elles, 25,0 % sont des abandons ayant eu lieu alors que le questionnaire était entamé et 75,0 % sont des abandons sur la première page (cf. Tableau 8).

Tableau 8 : Distribution des fiches adresses selon leur statut vis-à-vis de l'enquête

	Ensemble	
	Effectifs	Pourcentages
Abandon - questionnaire entamé	8 541	2,4 %
Abandon - questionnaire non commencé	25 146	7,2 %
Hors cible	5 320	1,5 %
Non répondant	181 486	51,9 %
Questionnaire complet	129 507	37 %
Total	350 000	100 %

Source : Inserm-Drees, EpiCov, les 20 lots.

Champ : Métropole.

Lecture : Le nombre de questionnaires complets est de 129 507, ce qui correspond à 37 % de l'échantillon.

Cette répartition est sensiblement différente si l'on ne s'intéresse qu'aux questionnaires remplis lors d'une collecte CATI : 51,4 % des abandons ont eu lieu en cours de questionnaire contre 48,6 % d'abandons sur la première page du questionnaire (Tableau 9). Une différence importante est l'intervention d'un enquêteur dans la collecte qui motive à participer à l'enquête et accompagner l'enquêté pendant le remplissage du questionnaire.

Se pose alors la question du profil des personnes concernées par chaque type d'abandon. Une analyse par régressions permet d'étudier, toutes choses égales par ailleurs, les caractéristiques sociodémographiques des personnes qui abandonnent le questionnaire. L'abandon est appréhendé ici comme un processus et deux modèles ont été estimés en conséquence : un premier modèle visant à estimer la probabilité d'abandonner sur la première page du questionnaire *versus* la probabilité de débiter le questionnaire (catégorie qui comprend les abandons en cours de questionnaire et les questionnaires complets, modèle 1a, Tableau 10), puis un second modèle qui permet d'estimer la probabilité d'abandonner en cours de questionnaire *versus* la probabilité d'y répondre en intégralité (modèle 1b).

Tableau 9 : Nombre d'abandons et de questionnaires complets selon le mode de passation

	Collecte par téléphone (CATI)		Collecte par Internet (CAWI)	
	Effectifs	Pourcentages	Effectifs	Pourcentages
Abandon - questionnaire entamé	560	4,3 %	7 981	5,3 %
Abandon - questionnaire non commencé	529	4,1 %	24 617	16,4 %
Questionnaire complet	11 945	91,6 %	117 562	78,3 %
Total	13 034	100 %	150 160	100 %

Source : Inserm-Drees, EpiCov, les 20 lots.

Champ : Métropole.

<sup>8</sup> Est qualifié d' « abandon sur la première page » tout abandon faisant suite à une connexion au questionnaire sans que le questionnaire n'ait été débuté (aucune réponse aux questions de la première page).

À la lumière de ces éléments, il apparaît que le fait d'être un homme, de vivre en QPV, d'être né à l'étranger, ou encore d'avoir répondu par Internet augmente de manière significative la probabilité d'abandonner que ce soit sur la première page ou en cours de questionnaire. L'effet du niveau de vie est quant à lui linéaire, les individus les plus aisés ont une moins grande propension à abandonner. De la même façon, l'âge a un effet quasi linéaire, plus les individus sont âgés, plus leur probabilité d'abandonner diminue. Quelques catégories d'âge font néanmoins exception aux deux extrêmes de la distribution. D'une part, nous avons les 15-24 ans qui, toutes choses égales par ailleurs, abandonnent plus que les 45-54 ans mais moins que 25-34 ans et, d'autre part, les 75 ans et plus qui abandonnent significativement plus que les 45-54 ans.

Les résultats des régressions mettent en évidence qu'il n'y a pas d'effet de la durée du questionnaire sur la probabilité d'abandonner. Que le questionnaire proposé ait été long (durée annoncée de 30 minutes, lots 1 et 20) ou court (20 minutes annoncées, les autres lots), il n'y a aucune différence significative sur le nombre d'abandons en cours de questionnaire ou sur la première page.

### **3.3.2. Abandons par téléphone (CATI), abandons par Internet (CAWI), quelles différences ?**

Ces résultats pourraient cependant masquer des effets de mode. C'est pourquoi trois modèles supplémentaires ont été estimés (le modèle 2b, relatif aux abandons en cours de questionnaire concernant le CATI et les modèles 3a et 3b relatifs aux abandons sur la première page et en cours de questionnaire pour le CAWI) afin d'analyser plus en détails les déterminants de l'abandon propre à chaque mode de passation, toujours en réalisant une ventilation par type d'abandon pour ce qui est des modèles 3<sup>9</sup>.

En outre, disposer de régressions spécifiques aux questionnaires collectés par Internet présente l'intérêt de pouvoir étudier le rôle joué par le matériel utilisé pour répondre.

En comparant le modèle 2b d'un côté et les modèles 3a et 3b de l'autre, il apparaît qu'une passation CATI, en plus de réduire la fréquence des abandons, permet aussi de limiter par rapport à une collecte par Internet, le poids des déterminants sociaux dans les abandons. En effet, le modèle 2b met en évidence moins d'effets significatifs que les deux derniers.

En CATI sont associés à une hausse de l'abandon en cours de questionnaire : le fait d'être un homme, d'être âgé de 65 ans ou plus, et le fait d'être né à l'étranger. L'appartenance aux déciles de niveaux de vie les plus élevés reste, quant à elle, associée à une baisse de l'abandon, quel qu'il soit. Il ne nous est toutefois plus possible, contrairement aux modèles précédents, de mettre en évidence une corrélation entre l'appartenance aux déciles les moins élevés et la propension à abandonner, quelle que soit la forme d'abandon considérée. Cela montre que cet effet tenait à la collecte CAWI.

En CAWI, les résultats obtenus sont proches de ce qui a déjà pu être observé sur les modèles 1, sur l'ensemble de l'échantillon.

Les variables spécifiques à Internet montrent que l'ordinateur demeure, l'option à privilégier pour ce qui est de limiter les abandons en cours de questionnaire. Ce résultat rejoint un constat globalement partagé par la littérature scientifique même si les raisons pour lesquelles l'ordinateur semble s'imposer comme le meilleur périphérique pour répondre à un questionnaire sont multiples et aucune ne fait

---

<sup>9</sup> Aucune régression n'a été réalisée pour modéliser la probabilité d'abandonner sur la première page du questionnaire plutôt que de débiter ce dernier en CATI, les différences entre ces abandons et les refus étant trop ténues.

consensus : vitesse et qualité de la connexion internet, plus grande taille d'écran, confort d'utilisation, etc. (Couper et Peterson, *ibid*).

L'usage de la tablette conduit quant à lui à moins d'abandons par rapport au mobile (Android ou iOS), quel que soit le type d'abandon considéré. Cela peut s'expliquer par le fait que la tablette vient pallier, au moins en partie, les défauts du mobile : les individus sont certainement moins mobiles avec une tablette qu'avec un téléphone, moins enclins à vaquer à d'autres tâches parallèlement au remplissage du questionnaire ; l'effort de « scrolling » est moindre car l'écran est plus grand et peut afficher plus d'informations, le fait d'avoir un écran plus grand limite par ailleurs le risque de faire un erreur dans le questionnaire et offre un confort de remplissage plus important, etc. (Couper et Peterson, 2016). En cela, l'usage de la tablette se rapproche de celui de l'ordinateur

En ce qui concerne les différences entre types de mobile, il apparaît d'une part, que les mobiles pour lesquels l'affichage du questionnaire n'a pas pu être optimisé<sup>10</sup>, conduisent beaucoup plus fréquemment à des abandons de tous types. Ces mobiles utilisent le plus souvent des systèmes d'exploitation différents d'Android ou iOS qui sont aujourd'hui les plus fréquemment répandus. D'autre part, en regardant de plus près les « user agent<sup>11</sup> » des personnes qui ont abandonné avec leur mobile, on constate qu'une part relativement importante s'est connectée au questionnaire depuis un lien réceptionné via WhatsApp ou Messenger<sup>12</sup>. Le questionnaire s'ouvre alors sur un navigateur internet interne à ces logiciels, souvent peu adapté pour une navigation de longue durée ou nécessitant beaucoup d'opérations à cette fin.

Tableau 10 : Modélisations de la probabilité d'abandonner

	(1a)	(1b)	(2b)	(3a)	(3b)
	Effet marginal	Effet marginal	Effet marginal	Effet marginal	Effet marginal
Constante	11,8***	4,3***	3***	9,6***	6,7***
Sexe					
Femmes	Réf.	Réf.	Réf.	Réf.	Réf.
Hommes	2,4***	0,7***	0,6*	1,8***	1,3***
Inconnu		5,6***			7,9***
Âge					
15 - 24 ans	2,4***	0,6**		2,2***	
25 - 34 ans	8,5***	2,0***		8,0***	1,5***
35 - 44 ans	3,3***	0,9***		3,2***	0,6*
45 - 54 ans	Réf.	Réf.	Réf.	Réf.	Réf.
55 - 64 ans	-0,8**			-0,8***	
65 - 74 ans	-0,6*		1,4*	-0,8***	
75 - 84 ans	2,5***	1,3***	2,6***	1,7***	2,9***
85 ans et plus	12,9***	8,8***	11,0***	10,6***	13,6***
Déciles de niveau de vie					
1er décile	5,3***	1,7***		4,7***	2,4***
2ème décile	3,4***	0,9***		3,1***	1,1**
3ème décile	2,5***	0,8**		2,1***	1,1*
4ème décile	0,8*	0,7**		0,7*	1,0*

<sup>10</sup> Parmi les paradonnées de la collecte Internet, une variable précisait si l'affichage avait pu ou non être optimisé sur l'écran du mobile.

<sup>11</sup> Le *user agent* est un code envoyé par le navigateur de l'utilisateur à un site web lorsqu'il tente de s'y connecter. Ce code permet, entre autres, de savoir quel est le navigateur ou encore le système d'exploitation utilisé par l'internaute.

<sup>12</sup> Il est aujourd'hui possible, sur la plupart des téléphones mobiles, de laisser des applications telles que WhatsApp ou Messenger gérer la réception et l'envoi de SMS. Ainsi certaines ont pu recevoir les SMS de relance contenant un lien les menant au questionnaire directement sur ces applications.



5ème décile	Réf.	Réf.	Réf.	Réf.	Réf.
6ème décile					
7ème décile	-1,9***	-0,5**		-1,7***	
8ème décile	-2,9***	-0,5**		-2,5***	
9ème décile	-4,0***	-1,1***	-1,1*	-3,5***	-1,4***
10ème décile	-4,9***	-1,2***	-1,2*	-4,2***	-1,4***
Le ménage vit en QPV					
Oui	3,0***	1,5***		2,9***	1,5***
Non	Réf.	Réf.		Réf.	Réf.
Lieu de naissance					
France	Réf.	Réf.	Réf.	Réf.	Réf.
Etranger	6,1***	3,8***	2,5***	5,6***	4,9***
Inconnu	5,0**			4,2**	
Mode de passation					
CATI	-9,2***	-1,5***			
CAWI	Réf.	Réf.			
Type de questionnaire					
Court			Réf.	Réf.	Réf.
Long	Réf.	Réf.		-0,5*	
Périphérique utilisé					
Mobile (Android ou iOS)				Réf.	Réf.
Mobile (autres)				18,9***	20,4***
Tablette				-3,4***	-3,2***
Ordinateur				3,6***	-3,9***

Source : Inserm-Drees, EpiCov, les 20 lots.

Lecture : Toutes choses égales par ailleurs, la probabilité d'abandonner le questionnaire en cours plutôt que d'y répondre en intégralité diminue de 1,5 % quand on répond en CATI plutôt qu'en CAWI (modèle 1b).

Significativité : \*\*\*  $\leq 0,001$  ; \*\*  $\leq 0,01$  ; \*  $\leq 0,5$

Note : Pour chaque modèle le préfixe désigne le champ sur lequel a été réalisé la régression (1 : ensemble des passations ; 2 : passations CATI ; 3 : passations CAWI). Le suffixe quant à lui désigne le phénomène que nous cherchons à modéliser (a : probabilité d'abandonner sur la première page plutôt que de débiter le questionnaire ; b : probabilité d'abandonner en cours de passation plutôt que de répondre au questionnaire en intégralité).

Les valeurs non significatives ne sont pas présentées.

Enfin, il est un dernier effet qui semble contre-intuitif : le fait qu'un questionnaire soit annoncé comme long est à l'origine d'une baisse de la probabilité d'abandonner le questionnaire sur la première page plutôt que de poursuivre la passation. Un effet analogue a été observé par Crawford *et al.* qui notaient dans un de leur article que de diminuer la durée de passation annoncée en début de questionnaire était associé à une hausse du taux d'abandon. D'après ces mêmes auteurs, il s'agit alors d'un « retour de flamme » du fait d'avoir sous-estimé la durée de passation. Or, cette hypothèse semble peu crédible ici étant donné que la diminution du taux d'abandon ne concerne que les abandons ayant eu lieu sur la première page du questionnaire et non les abandons en cours de passation. Ici, il semblerait que cette baisse de l'abandon soit due au fait que le questionnaire long ait été soumis uniquement à deux lots (lots 1 et 20) dont un suivait un protocole multimode (lot 1). Nous avons alors en proportion beaucoup plus de personnes contactées en CATI parmi les questionnaires longs que parmi les questionnaires courts, or comme nous l'avons précédemment été interrogé en CATI est associé à une baisse de l'abandon.

En conclusion, nous pouvons dire que les abandons CATI sont de nature différente des abandons CAWI, qui sont eux bien plus corrélés aux caractéristiques sociodémographiques. Cet écart peut très certainement s'expliquer par le fait que l'enquêteur accompagne l'enquêté dans une collecte par téléphone et crée une situation d'interaction favorisant la bonne compréhension du questionnaire et dans laquelle abandonner s'avère plus difficile.

### 3.3.3. A quel moment surviennent les abandons ?

En réalisant des courbes de survie des abandons en cours de questionnaire, nous pouvons observer que les abandons CATI et CAWI suivent la même tendance même si les abandons CATI se produisent généralement un peu plus tôt.

La majorité des abandons surviennent dans les premières minutes de l'entretien puisqu'environ 70 % des abandons ont lieu au cours du module A du questionnaire (le détail des modules du questionnaire est disponible en annexes, tableau 18) qui est un module introductif assez court, portant sur les caractéristiques sociodémographiques des individus, d'une durée médiane de 3 minutes. Cela peut certainement s'expliquer par une certaine incompréhension de la part des répondants ne voyant pas de rapport entre les questions posées et ce qui était annoncé dans la lettre-avis qui mettait en avant le fait que le questionnaire porterait sur le Covid-19. Ces questions sont également les plus intrusives en matière d'identification du répondant. Cette proportion d'abandons est par ailleurs très similaire à celle observée dans l'enquête VIRAGE (Charrance et Legleye, 2021) qui est une enquête dans laquelle le module sociodémographique placé en début de questionnaire était aussi relativement long et vraisemblablement généré de l'abandon.

Au moment où débute le module B une phase de décroissance plus lente s'amorce durant laquelle les écarts entre modes de passation se creusent.

La fin de ce module B est ponctuée par un décrochage relativement important en comparaison avec la décroissance lente qui caractérisait les abandons de ce même module. L'essentiel de ces abandons se produit après la question B36 qui est une question filtre amenant les répondants soit au module C directement soit, pour 13,2 % de l'échantillon<sup>13</sup> (faute de matériel d'autoprélèvement disponible pour tous les enquêtés), vers un sous-module « auto-prélèvement » avant d'aller au module C.

Ce sous-module s'ouvre sur une question demandant si les personnes sont ou non d'accord pour réaliser un auto-prélèvement sanguin. Le caractère particulièrement sensible de cette question est à l'origine d'une partie de ces abandons mais pas de la majorité. Ainsi, environ 27,0 % des abandons survenus à ce moment du questionnaire sont le fruit de personnes éligibles à la question « auto-prélèvement ».

Ce décrochage en fin de module B semble donc dû à un effet conjugué de l'apparition de ce sous-module « auto-prélèvement » qui a pu rebuter certains et aussi très certainement de la une lassitude en voyant s'ouvrir un nouveau module (après le module B qui est aussi le plus long du questionnaire). Le module C était un module portant sur la situation dans l'emploi et non plus sur le COVID-19. Certaines personnes ont alors pu croire que les informations qui intéressent les concepteurs de l'enquête avaient déjà été communiquées et qu'il n'était donc pas nécessaire de poursuivre.

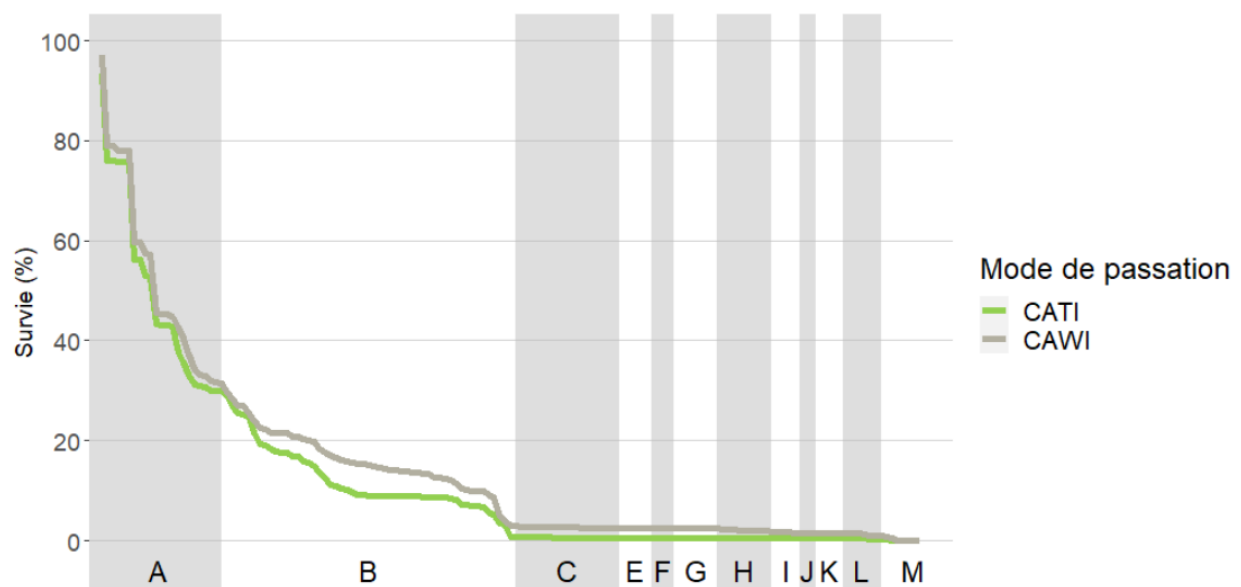
Ces différents événements mis bout à bout font que seuls 3 % des individus ayant abandonné le questionnaire EpiCov en cours sont encore présents dans le questionnaire à l'issue du module B. Les quelques abandons survenant après cela sont épars et ne semblent suivre aucune logique particulière, ce qui se traduit par une décroissance très lente jusqu'à la fin du questionnaire.

Le mode de passation joue un rôle relativement important dans la propension des individus à abandonner, les répondants CATI abandonnant moins que les répondants en CAWI. Par ailleurs, la collecte CATI semble prémunir d'une trop grande distorsion de l'échantillon, et cela notamment parce qu'elle tend à atténuer le poids des déterminants sociaux (âge, niveau de vie, lieu de naissance) au regard de l'abandon.

---

<sup>13</sup> La sélection des individus dépendant du département de résidence et du lot auquel ils ont été attribués.

Figure 1 : Courbes de survie des abandons par mode de collecte



Source : Inserm-Drees, EpiCov, les 20 lots.

Champ : Métropole.

Lecture : Environ 30 % des personnes ayant abandonné le questionnaire EpiCov en CATI l'ont fait après que le module B ait débuté.

Là où certains pouvaient voir dans le mobile un outil révolutionnaire en matière de collecte de données pour les questionnaires auto-administrés (Link et al., 2014), cette analyse des abandons ne nous permet pas d'être aussi enthousiastes. Le type d'appareil utilisé pour répondre au questionnaire par Internet n'est pas sans impact sur la propension à abandonner, le mobile étant l'appareil pour lequel le taux d'abandon est le plus fort, suivi par la tablette et enfin par l'ordinateur. En l'espèce, cette analyse plaide en faveur d'une conception de questionnaire orientée « mobile first » (Beck et al., 2022) permettant un affichage optimisé du questionnaire quel que soit le système d'exploitation, les taux d'abandons sur mobile étant particulièrement forts sur d'autres systèmes qu'Android ou iOS. Ils montrent également toute l'importance des « outils » pour la collecte, par exemple la nécessité que ceux-ci soient « responsive design » et donnent ainsi un rendu agréable du questionnaire sur tous les supports, sur tous les systèmes d'exploitation, etc.

En adoptant une définition extensive de l'abandon, recouvrant aussi bien les abandons survenant suite à une simple connexion que les abandons en cours de questionnaire, cela croisé au mode de passation, il apparaît que les déterminants de l'abandon ne sont pas strictement identiques selon les contextes. Si ces différents types d'abandons ne suivent pas strictement les mêmes logiques, cela invite à réfléchir à des remèdes différents permettant de limiter les abandons

#### 4. Protocoles multimodes : Comment articuler les différents modes proposés ?

Dans cette dernière partie, une analyse est menée sur les quatre lots de métropole exploités selon des protocoles multimodes. Pour trois d'entre eux, la collecte s'est déroulée selon un protocole multimode purement concurrentiel, les collectes internet et téléphone ont été ouvertes simultanément (ou quasi pour le lot 3). Pour le quatrième lot, on ne peut parler de multimode séquentiel au sens strict du terme car le premier mode (CAWI) est resté ouvert pendant toute la durée du terrain, mais le second mode (CATI) n'a été ouvert qu'à mi-terrain (J+14).

L'objectif ici est d'étudier l'impact des différences de calendrier (en l'occurrence, l'ouverture plus tardive du CATI pour le lot 4) afin de tirer des enseignements sur la manière d'exploiter au mieux la complémentarité des deux modes de collecte.

#### 4.1. Les différences de calendriers des lots multimodes

Quelques éléments ont déjà été indiqués sur les différences de calendrier des lots multimodes mais revenons plus en détail sur ce point (cf Annexe 1).

Sur le lot 1, l'ouverture du CATI et l'envoi du mail annonce ont tous deux eu lieu le 2 mai 2020. Les deux modes de collecte sont donc disponibles sur toute la durée du terrain. Sur le lot 2, le mail d'annonce a été envoyé le 5 mai 2020. Du lancement (2 mai 2020) jusqu'à cette date, seule la participation par téléphone était donc possible, les personnes n'étant pas encore informées de la possibilité de participer en ligne (le courrier étant parti le 4 mai + délai d'acheminement). Pour le lot 3, l'exploitation téléphonique a débuté le 4 mai. Cependant, à cette date, les personnes n'avaient encore reçu aucun des supports annonce (lettre annonce envoyée le 4 mai, mail annonce envoyé le 5 mai et SMS annonce envoyé le 8 mai). On peut aussi considérer ce lot comme du multimode concurrentiel. Le lot 4 se distingue plus nettement des autres, avec une ouverture du CATI le 16 mai (soit 14 jours après le lancement de l'enquête). Jusqu'à cette date, les personnes avaient uniquement la possibilité de répondre en ligne.

#### 4.2. Des taux de participation très proches selon les lots

En termes de participation, on observe assez peu de différences entre les différents lots. La participation varie de 45,4 % sur le lot 4 à 48,3 % sur le lot 2, soit une différence de 3 points de pourcentage.

Tableau 11 : Taux de réponse par lot

	Répondants	Taille du lot	Taux de participation
Lot 1	7 867	17 092	46,0 %
Lot 2	8 265	17 109	48,3 %
Lot 3	8 053	17 147	47,0 %
Lot 4	7 798	17 168	45,4 %
Total (lots multi-modes)	31 983	68 516	46,7 %

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 4, métropole.

*Lecture* : 46% des personnes échantillonnées et non identifiées comme hors cible appartenant au premier lot ont participé à l'enquête.

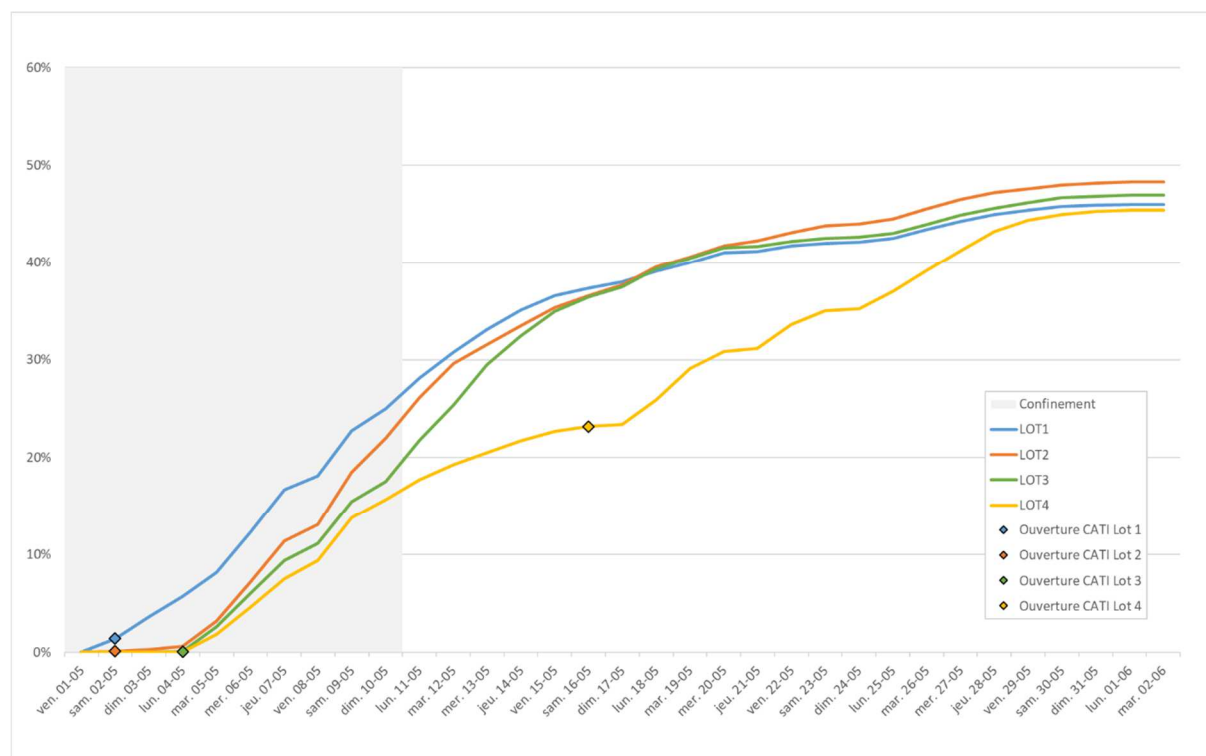
*Note* : le taille du lot exclut les personnes classées « hors champ ».

Lorsque les deux modes de réponse sont proposés dès le départ (lots 1 à 3), le taux de réponse semble légèrement meilleur. Par ailleurs, on note que le taux de réponse du lot 1 est inférieur à celui des lots 2 et 3. Cette différence est très certainement due au fait que le lot 1, contrairement aux lots 2 à 4, a été soumis à un questionnaire long (environ 30 minutes vs 20 minutes pour les autres lots), information délivrée lors de la prise de contact et pouvant jouer sur la décision de participer ou non.

#### 4.3. Un rythme de collecte ralenti sur le lot avec ouverture différée du CATI

Nous nous sommes également intéressés à l'impact des différences de calendrier sur le rythme de collecte. Le Graphique 10 montre un rythme de collecte assez similaire sur les trois premiers lots (aux protocoles assez proches) et un rythme plus lent et régulier sur le lot 4. L'ouverture plus tardive du lot 4 ralentit la collecte mais permet d'atteindre un taux de participation malgré tout semblable aux autres lots.

Graphique 10 : Taux de réponse cumulé par lot



Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 4, métropole.

#### 4.4. Une part de questionnaires CAWI plus importante lorsque le CATI est ouvert tardivement

Le recours à un protocole multimode est souvent motivé par la volonté ou même la nécessité de réduire les coûts de collecte. Dans le cas du multimode séquentiel, on propose dans un premier temps le mode de collecte le moins coûteux afin de limiter les coûts de collecte, puis on recourt à des modes de collecte plus coûteux dans un second temps afin de capter les non-répondants. Nous faisons donc l'hypothèse, dans le cas d'EpiCov, que le lot 4 permettrait d'avoir une part de questionnaires CAWI plus importante que les autres lots exploités selon un protocole multimode concurrentiel pur.

Sans surprise et conformément à l'hypothèse formulée, le lot 4 compte, parmi ses répondants, près de trois quarts de questionnaires complétés en CAWI contre 54,7 % à 62,1 % sur les autres lots. La part de questionnaires réalisés sur les deux modes (questionnaire débuté en CAWI puis finalisé avec un enquêteur ou l'inverse) est quant à elle assez faible pour les 4 lots (et minimale pour le lot 4).

Tableau 12 : Répartition des questionnaires selon le mode de collecte par lot

	CAWI		CATI		CAWI-CATI		Total	
Lot 1	4 305	54,7 %	3 314	42,1 %	248	3,2 %	7 867	100,0 %
Lot 2	5 006	60,6 %	3 153	38,1 %	106	1,3 %	8 265	100,0 %
Lot 3	5 004	62,1 %	2 860	35,5 %	189	2,3 %	8 053	100,0 %
Lot 4	5 740	73,6 %	1 992	25,5 %	66	0,8 %	7 798	100,0 %
Total	20 055	62,7 %	11 319	35,4 %	609	1,9 %	31 983	100,0 %

Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 4, métropole.

Lecture : 54,7% des répondants du lot 1 ont participé en CAWI (en auto-administré, sur internet), cela concerne 4305 personnes

#### 4.5. Une exploitation téléphonique allégée lorsque celle-ci se fait plus tardivement

Le recours à un multimode davantage séquentiel que concurrentiel permet a priori de réduire le

volume de numéros à appeler, en éliminant les personnes ayant déjà répondu en CAWI, et de ce fait, permet de réduire les coûts de collecte. Nous avons donc mobilisé les données des quatre lots multimodes pour confirmer ou infirmer cette hypothèse. Mais nous nous sommes également demandés si les personnes appelées du lot 4 (sollicités depuis 2 semaines pour une participation CAWI) étaient plus difficiles à joindre que les personnes appelées des autres lots (non-réponse par manque de temps par exemple), et nécessitaient donc davantage de tentatives d'appels.

L'exploitation téléphonique est conditionnée par la présence d'un numéro de téléphone dans le fichier. On dispose d'au moins un numéro de téléphone pour 83,3 à 83,5 % des individus des lots 1 à 3 et 79,4 % pour le lot 4. Cette différence s'explique par le fait que le lot 4 était initialement destiné à une exploitation purement CAWI. L'enrichissement téléphonique réalisé par Ipsos a donc été fait tardivement, uniquement sur les numéros non-répondants (en CAWI) lors de la prise de décision d'ouvrir ce lot au CATI et non sur l'intégralité du lot.

Sur les lots 1 et 2, près de 94 % des individus avec numéro de téléphone ont été appelés. Sur le lot 3, avec une exploitation téléphonique débutant deux jours après le lancement de l'enquête, 85,2 % des individus avec numéro de téléphone ont été appelés (soit 8 points de moins que pour les lots 1 et 2). Sur le lot 4, avec une ouverture bien plus tardive du CATI, 14 jours après le lancement de l'enquête, seules 71,4 % des personnes avec numéros ont été appelées (si l'enrichissement avait été fait sur le lot complet, cette proportion serait encore plus basse, a priori). Le volume de numéros appelés est donc nettement réduit lorsque l'exploitation téléphonique se fait plus tardivement.

Par ailleurs, l'analyse des données révèle que les personnes appelées appartenant au lot 4 n'ont pas été plus difficiles à joindre que ceux des autres lots. On enregistre, en moyenne, 6,7 appels par personne appelée sur le lot 4, versus 5,3 pour le lot 1, 9,1 pour le lot 2 et 7,7 pour le lot 3.

Tableau 13 : Nombre de numéros appelés et nombre d'appels passés par lot

	Nombre de numéros appelés	% de numéros appelés parmi les appelables	Nombre d'appels passés	Nombre moyen d'appels passés par numéro appelé
Lot 1	13 714	93,9 %	73 168	5,3
Lot 2	13 653	93,6 %	123 614	9,1
Lot 3	12 414	85,2 %	95 485	7,7
Lot 4	9 917	71,4 %	66 155	6,7
Total	49 698	86,2 %	358 422	7,2

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 4, métropole.

*Lecture* : 93,9% des individus appelables (présence d'un numéro de téléphone) du lot 1 ont été appelés, soit 13 714 individus.

Nous avons également cherché à savoir si la non-participation à l'enquête des individus du lot 4 jusqu'alors était due au fait que le mode de réponse proposé ne leur convenait pas ou plutôt à un refus ferme de participer à l'enquête (quel que soit le mode de réponse).

Pour cela, nous nous sommes appuyés sur les taux de réponse observés sur les appelés, et comme suspecté, le taux de réponse sur les appelés du lot 4 est moindre par rapport à celui des autres lots : il est d'environ 50 % sur les lots 1 et 2, 44,8 % sur le lot 3 et 32,9 % sur le lot 4. La non-participation après 15 jours de sollicitation pour répondre en CAWI est bien, pour une partie, un refus de participer à l'enquête, quel que soit le mode de réponse.

Parmi les appelés du lot 4, 17,2 % s'étaient connectés au moins une fois au questionnaire avant d'être appelés contre moins de 9 % pour les appelés des lots 1 et 2 et 15,4 % du lot 3. Le faible écart entre les lots 3 et 4 semble montrer que les personnes, quand elles se connectent au CAWI, le font assez rapidement après l'invitation.

Nous avons étudié le mode de réponse finalement utilisé par les répondants appelés. Lors de l'appel

téléphonique, les personnes pouvaient indiquer à l'enquêteur préférer répondre en CAWI plutôt qu'en CATI (n'étant pas dans un protocole multimode séquentiel). Parmi les personnes appelées, 45,8 % d'entre elles ont finalement participé à l'enquête en CAWI, ce qui n'est pas du tout négligeable.

Tableau 14 : Répartition des questionnaires des individus appelés (exploitation téléphonique) selon le mode de collecte par lot

	CAWI		CATI		CAWI-CATI		Total	
Lot 1	3 014	45,8 %	3 314	50,4 %	248	3,8 %	6 576	100,0 %
Lot 2	3 559	52,2 %	3 153	46,2 %	106	1,6 %	6 818	100,0 %
Lot 3	2 385	43,9 %	2 860	52,6 %	189	3,5 %	5 434	100,0 %
Lot 4	1 128	35,4 %	1 991	62,5 %	66	2,1 %	3 185	100,0 %
Total	10 086	45,8 %	11 318	51,4 %	609	2,8 %	22 013	100,0 %

Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 4, métropole.

Lecture : 45,8 % des répondants ayant été appelés ont répondu en CAWI (auto-administré sur Internet), soit 3 014 individus.

Sur le lot 4, on note une part plus importante de répondants CATI parmi les personnes appelées. Ce résultat n'est pas surprenant. Pendant les 15 jours précédant l'appel, les personnes avaient la possibilité de répondre en ligne. Si elles ne l'ont pas fait, c'est qu'a priori, ce mode de réponse ne leur convenait pas et elles ont donc été plus nombreuses à choisir le téléphone lorsque celui-ci leur a été proposé. Cependant, il est également important de noter qu'il reste pour ce lot 35,4 % de personnes appelées qui ont répondu en CAWI après une relance téléphonique, cela montre que le fait d'appeler permet également de relancer la collecte CAWI.

#### 4.6. Des biais de réponse relativement similaires dans les deux protocoles

L'objectif de cette partie est d'identifier si l'ouverture tardive du CATI sur le lot 4 a un impact sur la structure de l'échantillon collecté. Pour cela, nous avons mobilisé quelques variables sociodémographiques disponibles dans le fichier Fideli. Nous avons également regardé si les différences de protocoles avaient un impact sur les taux de réponse en fonction des coordonnées disponibles.

Le Tableau 15 montre que les taux de réponse des personnes sans coordonnées téléphoniques (Aucune coordonnée et Mail uniquement) sont supérieurs sur le lot 4 que sur les autres lots (écarts plus faibles avec le taux de réponse moyen que sur les autres lots). Malheureusement, nous ne disposons pas de pistes d'explications pour ces écarts, hormis le fait qu'une partie de ces personnes seraient potentiellement classées dans une autre catégorie (Mail + Téléphone ou Téléphone uniquement) si l'enrichissement téléphonique avait été réalisé sur la totalité du lot 4. Des analyses complémentaires ont cependant été menées dans ce sens et ont montré que les écarts se maintiennent.

Tableau 15 : Ecarts au taux de réponse moyen du lot en points de pourcentage

	Lot 1	Lot 2	Lot 3	Lot 4
<b>Coordonnées disponibles</b>				
Aucune coordonnée	-35,3	-37,5	-36,7	-30,3
Mail + Téléphone	11,1	11,6	11,3	11,3
Mail uniquement	-18,3	-18,7	-19,7	-9,5
Téléphone uniquement	-12,3	-12,6	-11,5	-15,0
<b>Sexe</b>				
Homme	-1,3	-1,8	-1,2	-1,2
Femme	2,3	2,5	1,8	2,3
Non renseigné	-9,7	-8,4	-7,2	-11,1

	Lot 1	Lot 2	Lot 3	Lot 4
<b>Classe d'âge</b>				
15 - 24 ans	-2,3	-0,2	-1,0	-1,1
25 - 34 ans	-0,9	-2,6	-3,0	-1,5
35 - 44 ans	2,2	4,0	3,6	3,7
45 - 54 ans	5,1	3,8	6,5	5,3
55 - 64 ans	4,0	3,5	4,1	4,8
65 - 74 ans	1,9	3,0	0,3	0,5
75 - 84 ans	-6,0	-8,4	-9,4	-9,2
85 ans et plus	-26,8	-26,6	-24,2	-27,9
<b>Lieu de naissance</b>				
France	3,1	2,9	2,6	2,9
Etranger	-12,4	-12,8	-11,2	-11,4
Non renseigné	-9,7	-8,5	-7,2	-10,8
<b>Décile de niveau de vie</b>				
1er décile	-15,3	-16,2	-14,9	-13,9
2ème décile	-13,0	-11,4	-12,4	-12,7
3ème décile	-9,0	-8,1	-8,3	-7,7
4ème décile	-1,1	-2,1	-1,6	-3,8
5ème décile	1,6	2,7	0,7	1,4
6ème décile	4,3	5,5	4,7	4,4
7ème décile	8,5	9,1	8,3	8,4
8ème décile	10,3	11,0	12,2	11,1
9ème décile	17,3	15,4	16,2	15,5
10ème décile	16,6	15,8	16,3	17,8
Inconnu	-17,9	-21,1	-19,8	-21,6
<b>Le ménage vit en QPV</b>				
Non	1,4	1,2	1,3	1,3
Oui	-16,4	-15,6	-15,4	-15,2
<b>Indicateur de pauvreté</b>				
Non	3,7	3,6	3,6	3,5
Oui	-14,6	-14,3	-14,4	-13,9

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 4, métropole.

*Lecture* : Le taux de réponse observé chez les individus du lot 1 sans aucune coordonnée de contact est inférieur de 35,3 points à celui observé au sein du lot 1

Sur les variables sociodémographiques considérées, on observe que les biais sont relativement identiques sur les quatre lots : une sous-représentation des individus au sexe non renseigné (personnes à charge de moins de 18 ans), des 75 ans et plus, des personnes nées à l'étranger ou dont le pays de naissance n'est pas renseigné, des ménages au niveaux de vie les plus faibles (trois premiers déciles), des personnes résidant en QPV ou encore des personnes vivant en dessous du seuil de pauvreté. Concernant l'ampleur de ces biais entre les lots 1 à 3 et le lot 4, on ne peut conclure à un effet significativement différent du protocole sur la structure des répondants.

En dernier lieu, nous nous sommes intéressés à la comparaison des échantillons par mode selon l'appartenance au lot. Nous avons, cette fois-ci, agrégé les lots 1 à 3 pour éviter de multiplier les résultats. Globalement, les différences CATI/CAWI sont les mêmes dans les échantillons 1-3 et 4. En mobilisant un test du Khi<sup>2</sup>, sur le lot 4, on observe davantage de disparités entre CATI et CAWI sur le sexe et moins de différences sur le lieu de naissance ou le fait d'habiter en QPV.



Tableau 16 : Distribution des variables sociodémographiques dans les échantillons CATI &amp; CAWI

	Lots 1 à 3			Lot 4		
	CAWI	CATI	P(Khi <sup>2</sup> )	CAWI	CATI	P(Khi <sup>2</sup> )
<b>Sexe</b>			0,4367			0,0457
Homme	44,3	43,9		43,7	46,7	
Femme	51,5	51,6		52,3	49,1	
Non renseigné	4,2	4,5		4,0	4,2	
<b>Classe d'âge</b>			<,0001			<,0001
15 - 24 ans	14,3	11,7		14,4	11,3	
25 - 34 ans	13,2	13,0		13,2	14,2	
35 - 44 ans	17,1	15,2		17,0	15,4	
45 - 54 ans	19,3	16,3		18,9	15,2	
55 - 64 ans	17,5	16,0		17,8	15,2	
65 - 74 ans	13,2	15,3		13,1	16,4	
75 - 84 ans	4,3	9,4		4,5	9,5	
85 ans et plus	1,2	3,2		1,2	2,9	
<b>Lieu de naissance</b>			<,0001			0,0265
Non renseigné	4,4	4,8		4,3	4,5	
France	85,1	83,1		85,2	82,8	
étranger	10,4	12,1		10,5	12,7	
Non renseigné	4,4	4,8		4,3	4,5	
<b>Décile de niveau de vie</b>			<,0001			<,0001
1er décile	8,2	10,5		8,6	11,7	
2ème décile	7,4	9,6		7,0	9,9	
3ème décile	6,0	8,0		6,5	8,9	
4ème décile	7,4	8,9		7,2	8,6	
5ème décile	8,9	9,5		8,2	10,3	
6ème décile	9,8	9,8		10,1	9,8	
7ème décile	11,0	10,4		10,8	10,4	
8ème décile	11,6	10		11,9	10,4	
9ème décile	13,4	10,4		13,7	8,6	
10ème décile	13,7	9,5		13,7	8,0	
Inconnu	2,7	3,4		2,4	3,3	
<b>Le ménage vit en QPV</b>			<,0001			0,0042
Non	95,6	94,0		95,3	93,7	
Oui	4,4	6,0		4,7	6,3	
<b>Indicateur de pauvreté</b>			<,0001			<,0001
Non	87,5	83,8		87,4	82,6	
Oui	12,5	16,2		12,6	17,4	

*Source* : Inserm-Drees, EpiCov, vague 1.

*Champ* : lots 1 à 4, métropole.

*Lecture* : Sur les lots 1 à 3, les hommes représentent 44,3 % de l'échantillon CAWI et 43,9 % de l'échantillon CATI.

L'objectif de cette dernière partie était d'évaluer l'impact d'une ouverture tardive du CATI en comparant les lots 1 à 3 (ouverture simultanée de deux modes de collecte) au lot 4 (ouverture du CATI 2 semaines après le lancement du CAWI).

Le protocole avec ouverture tardive du CATI présente un rythme de collecte plus lent mais ce « retard » est finalement compensé, conduisant à un taux de réponse quasi similaire aux lots avec ouverture simultanée des deux modes. Par ailleurs, les analyses ont montré que ce protocole permet

d'augmenter la part de questionnaires CAWI, avantage non négligeable en matière de coûts. Il permet également de réduire le volume d'appels et de numéros appelés, impactant là aussi les coûts de collecte. L'exploitation téléphonique, quant à elle, offre des résultats moins satisfaisants lorsqu'elle a lieu plus tardivement (moindre taux de réponse) mais permet cependant de relancer la collecte CAWI. Nous avons d'ailleurs tenté de synthétiser les éléments présentés dans cette partie pour calculer un coût par questionnaire. Pour cela, nous avons associé à chaque appel passé un coût (mimant la durée de la communication) :

- un appel « contact » coûte 3 unités (soit 3 minutes),
- un appel « passation complète » coûte 30 unités pour le lot 1, 20 pour les lots 2 à 4 (le lot 1 étant soumis à un questionnaire long),
- un appel « passation partielle » coûte 15 unités pour le lot 1 et 10 unités pour les lots 2 à 4.

Par ailleurs, nous avons fait l'hypothèse qu'il a fallu un appel « passation complète » par répondant CATI et un appel « passation partielle » par répondant CATI-CAWI. Le coût d'un questionnaire réalisé intégralement en CAWI est quant à lui, fixé à 0 unité.

Tableau 17 : Estimation des coûts pour chaque lot

	Lot 1	Lot 2	Lot 3	Lot 4	Total
Nb de répondants	7 867	8 265	8 053	7 798	31 983
Nb de répondants CATI	3 314	3 153	2 860	1 992	11 319
Nb de répondants CATI-CAWI	248	106	189	66	609
Nb d'appels total	73 168	123 614	95 485	66 155	358422
Nb appels "Contact"	69 606	120 355	92 436	64 097	346 494
Nb appels "Passation Complète"	3 314	3 153	2 860	1 992	11 319
Nb appels "Passation partielle"	248	106	189	66	609
Coûts totaux	311 958	425 185	336 398	232 791	1 306 332
Coûts / questio	40	51	42	30	41

Source : Inserm-Drees, EpiCov, vague 1.

Champ : lots 1 à 4, métropole.

La dernière colonne du Tableau 17 montre bien que le lot 4 permet un coût par questionnaire plus faible que les lots 1 à 3. Cependant, Ipsos nous a indiqué avoir eu des difficultés en matière de remontées d'information sur les lots 1 et 4, conduisant certainement à une sous-estimation des appels sur ces deux lots. Il convient donc de considérer ces résultats avec précaution, mais la comparaison des lots 1 et 4 (y compris en tenant compte de la différence de temps de passation du questionnaire avec un coût de 35 unités/questionnaire) montre tout de même la plus-value du protocole du lot 4 par rapport au lot 1.

Sur le profil des répondants, ce décalage ne semble pas avoir d'impact. Les caractéristiques de l'échantillon collecté au sein du lot 4 ne sont pas significativement différentes de celles des échantillons des lots 1 à 3. De même, les différences de profil entre répondants CATI et répondants CAWI semblent être à peu de choses près les mêmes dans les différents lots.

A ce stade, et au vu des éléments présentés dans ce document, l'ouverture tardive du CATI semble donc permettre de réduire les coûts sans pour autant biaiser davantage l'échantillon collecté.

## Conclusion

L'étude de la collecte de la première vague de l'enquête EpiCov nous offre des perspectives sur les bonnes pratiques pour les enquêtes ménages CATI et/ou CAWI. Si d'autres analyses pourraient être menées à partir des parodonnées de la collecte d'Epicov (notamment sur d'autres vagues de l'enquête pour évaluer les liens avec l'attrition, mais aussi spécifiquement sur les DROM), de cet article, nous retenons les enseignements suivants :

- Enseignements généraux :

La multiplication des relances sous des formes différentes (courrier postal, mail, SMS) favorise la collecte. Il convient donc de pouvoir mobiliser tous ces modes de contact pour animer la collecte. Le courrier postal reste, même avec le développement du numérique, incontournable pour annoncer une enquête.

- Concernant la collecte CATI :

Mener un enrichissement annuaire des numéros de téléphone reste important à ce jour, malgré la richesse de Fidéli en matière de numéros de téléphone et l'augmentation de l'inscription sur les listes rouges. En effet, près de 10 % des questionnaires CATI ont été renseignés via un numéro issu de l'enrichissement et les profils de ces personnes diffèrent de ceux pour lesquels un numéro est disponible dans Fidéli.

Tous les numéros de téléphone ne permettent pas également d'obtenir un questionnaire. Les numéros « 09 » sont par exemple très peu rentables pour la collecte. En cas de surcharge globale en numéros de téléphone (au regard par exemple de la durée de la collecte) ou au niveau individuel, il est préférable de les exclure de l'exploitation téléphonique. Les numéros des autres personnes du ménage sont utiles pour obtenir des questionnaires, il est donc conseillé de les utiliser même pour une enquête « individuelle » (et non « ménage »).

Enfin, la répartition des appels sur différents créneaux reste cruciale ; le samedi est notamment un jour permettant de nombreux contacts avec les enquêtés.

- Concernant la collecte CAWI

L'effet du courrier postal est plus fort pour une collecte CAWI que CATI car aucun contact avec un enquêteur n'a lieu. Les abandons en cours de collecte sont plus fréquents sur mobile que sur ordinateur alors que les enquêtés ont été plus nombreux à répondre par mobile. Cela met bien en évidence la nécessité d'avoir des outils de collecte adaptés à tous les mobiles. Comme les outils, la conception du questionnaire gagne à être adaptée au CAWI et même spécifiquement aux mobiles. Le fait que les abandons soient plus liés aux caractéristiques sociodémographiques lors d'un remplissage CAWI par rapport à une collecte CATI met sans doute en évidence une difficulté ou une lassitude face au questionnaire que le travail de l'enquêteur permet d'atténuer. L'analyse n'a pas mis en évidence d'effet de la longueur du questionnaire (entre 20 minutes et 30 minutes en moyenne) sur le taux d'abandon par Internet.

- Sur le multimode

Si la multiplication des modes complexifie l'organisation d'une enquête, la collecte multimode permet, d'une part, un taux de réponse plus élevé et, d'autre part, de rapprocher la structure des répondants de l'échantillon initial, limitant ainsi les biais de sélection.

Sur le plan pratique, il semble plus avantageux d'ouvrir la collecte CATI une fois la collecte CAWI lancée, limitant le coût de la collecte téléphonique sans que cela ne nuise à la représentativité des échantillons de répondants obtenus. Cela réduit la période de mobilisation des enquêteurs mais nécessite d'en avoir un nombre suffisant sur une période réduite pour assurer la collecte CATI.

Plus généralement, cet article regroupe un ensemble d'analyses qui ont pu être menées à partir des parodonnées recueillies dans les systèmes d'information d'Ipsos lors de la collecte de l'enquête EpiCov. Il tend à montrer la richesse des analyses qui peuvent être menées avec ce type de données dont on a moins l'habitude de tirer des résultats qu'avec les données relatives au questionnaire. L'intérêt des

résultats invite à renforcer l'exploitation de ces données pour capitaliser sur un « savoir-commun » en matière de problématiques à aborder avec ces données et de manière de traiter statistiquement ces données. Elle invite aussi à favoriser le recueil de ces parodonnées dans les outils de collecte. C'est à partir d'un recueil de plus en plus riche que les analyses pourront affûter leur angle d'attaque et leurs méthodes et obtenir les résultats les plus pertinentes possibles pour améliorer les protocoles de collecte. À terme, l'analyse des parodonnées pourrait guider en temps réel la collecte.

## Bibliographie

- [1] ANTOUN, Christopher, COUPER, Mick et CONRAD, Frederick. Effects of Mobile versus PC Web on Survey Response Quality: A Crossover Experiment in a Probability Web Panel. *Public Opinion Quarterly*. 1 avril 2017. Vol. 81, n° 51, pp. 280-306. DOI 10.1093/poq/nfw088.
- [2] BECK, François, CASTELL, Laura, LEGLEYE, Stéphane et SCHREIBER, Amandine. Le multimode dans les enquêtes auprès des ménages : une collecte modernisée, un processus complexifié. *Courrier des statistiques*. 2022. Vol. 7, pp. 7-28.
- [3] CHARRANCE, Géraldine, LEGLEYE, Stéphane. Les abandons dans l'enquête Virage : quel impact sur les estimations de violence ? 2021. 11<sup>e</sup> Colloque francophone sur les sondages, Bruxelles, Belgique
- [4] LINK, Michael, MURPHY, Joe, SCHOBER, Michael, BUSKIRK, Trent, HUNTER CHILDS, Jennifer et LANGER TESFAYE, Casey. Mobile technologies for conducting, augmenting and potentially replacing survey. *Public Opinion Quarterly*. 2014. Vol. 78, n° 4, pp. 779-787. DOI doi:10.1093/poq/nfu054.
- [5] CRAWFORD, Scott, COUPER, Mick et LAMIAS, Mark. Web Surveys: Perceptions of Burden. *Social Science Computer Review*. 2001. Vol. 19, n° 2, pp. 146-162. DOI 10.1177/089443930101900202.
- [6] COUPER, Mick et PETERSON, Gregg. Why do web surveys take longer on smartphones ? *Social Science Computer Review*. 2016. pp. 1-21. DOI 10.1177/0894439316629932.
- [7] WARSZAWSKI, Josiane et al., A national mixed-mode seroprevalence random population-based cohort on SARS-CoV-2 epidemic in France: the socio-epidemiological EpiCov study, 2021, MedRxiv.

## Annexes

Figure 1 : Calendrier d'exploitation des lots en métropole

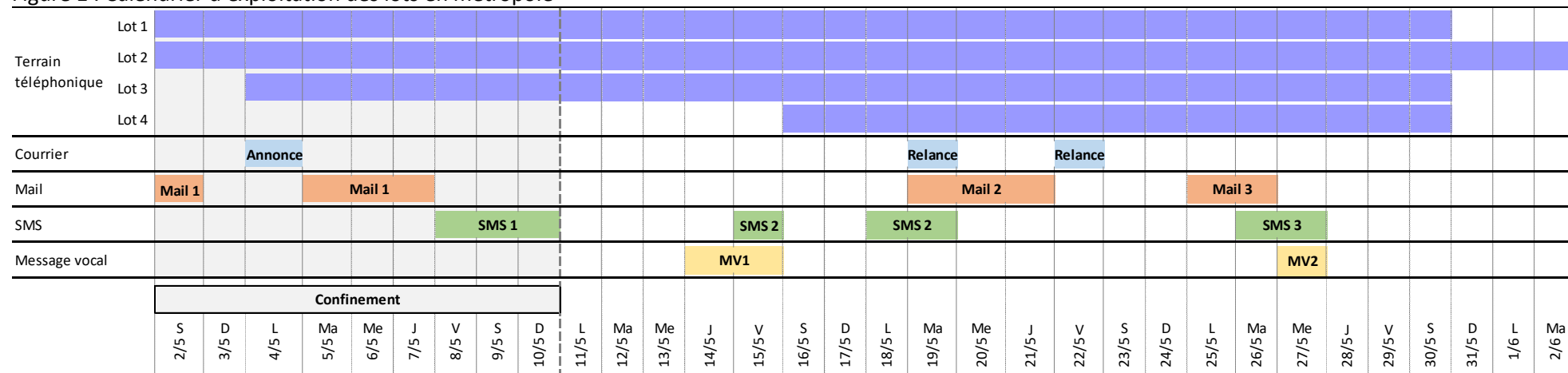


Tableau 18 : Détails des modules du questionnaire

A	Situation actuelle
B	Santé
C	Emploi
E*	Vie quotidienne*
F	Enfants
G*	Vie affective*
H	Sorties
I	Tabac et alcool
J*	Ressenti sur la situation*
K*	Opinion sur l'épidémie*
L	Parents
M	Téléphone/internet

\* Modules ajoutés au questionnaire long

