Journées de méthodologie statistique de l'Insee 2022

Le traitement de la non réponse endogène : une application à l'enquête post formation

Alejandra Arbelaez (*), Anne Bucher(*), Melchior Fosse(**) Pauline Givord(*)

(*) Dares (**) Insee

alejandra.arbelaez@travail.gouv.fr- anne.bucher@travail.gouv.fr-
melchior.fosse@insee.fr- pauline.givord@travail.gouv.fr

Mots-clés. : Échantillonnage, calage, collecte, équilibrage, multimode, séries temporelles.

Domaines. Multimode, Redressement

Résumé

Cette étude s'intéresse au redressement de la non réponse endogène dans l'enquête Post-Formation. Cette enquête a été mise en place par la Dares en 2019, auprès des anciens stagiaires de la formation professionnelle.L'enquête est menée chaque trimestre auprès de 25000 à 30000 personnes qui ont suivi une formation professionnelle au titre de leur recherche d'emploi, six mois après la fin de leur formation. Les questions portent sur le déroulement de la formation, le ressenti de la qualité de celle-ci ainsi que sur la situation actuelle des stagiaires. L'enquête est auto-administrée, les personnes interrogées pouvant répondre en ligne ou avec un questionnaire papier. Les taux de réponse sont peu élevés (34% en moyenne). Ce faible taux de réponse peut être un problème en cas de non réponse endogène, c'est-à-dire si la propension à répondre à l'enquête est corrélée avec les variables d'intérêt.

Suivant Castell et Sillard (2021), nous proposons de redresser ces biais de non réponse endogène en appliquant un modèle de sélection endogène d'Heckman (Tobit II). Plus précisément, nous mobilisons deux vagues de l'enquête pour lesquelles un sur-échantillon aléatoire a bénéficié de relances téléphoniques. Ces relances téléphoniques augmentent nettement le taux de réponse dans le sur-échantillon. On modélise simultanément la variable d'intérêt et le comportement de réponse à l'enquête, à partir d'un modèle de sélection où le fait d'appartenir à l'échantillon avec relances est utilisé comme instrument.

Introduction

L'enquête Post-Formation est menée chaque trimestre depuis 2019 auprès de personnes ayant été en recherche d'emploi et ayant, à ce titre, suivi une formation professionnelle. Elle cible les personnes entre 6 et 9 mois après leur sortie de formation, y compris celles qui l'ont interrompue avant la fin. L'échantillon est constitué à partir des fichiers administratifs des stagiaires de la formation professionnelle (base Brest, voir section 1), avec un objectif de 25 000 à 30 000 répondants par trimestre et d'une représentativité régionale.

L'enquête est auto-administrée, avec un recueil sur internet ou sur papier. Les taux de réponse sont faibles (34% en moyenne, avec une tendance à la baisse). Ce faible taux de réponse interroge sur les méthodes de redressement à adopter, en particulier si les comportements de non réponse peuvent être en partie corrélés avec les variables d'intérêt qu'on souhaite mesurer (on parle de non réponse endogène). Par exemple, les personnes qui ont retrouvé un emploi depuis la formation et sont occupées au moment de l'enquête sont moins disponibles pour répondre. Les personnes qui ont suivies une formation à distance ont aussi plus de chances d'avoir une connexion internet, et donc de répondre à une enquête en ligne. On peut aussi supposer que le fait d'avoir été satisfait, ou à l'inverse déçu par la formation peut aussi avoir une incidence sur la probabilité de répondre à une enquête s'y rapportant. Cette corrélation entre variables d'intérêt et comportement de réponse peut persister même en conditionnant par des variables auxiliaires disponibles, et les méthodes de correction de la non réponse classique peuvent donc être insuffisantes pour obtenir des estimateurs sans biais.

Comme montré récemment par Castell et Sillard (2021), une solution pour redresser de la non réponse endogène est d'utiliser des modèles de sélection d'Heckman (Tobit II). Ce modèle consiste à modéliser simultanément la variable d'intérêt et le comportement de réponse à l'enquête, à la condition de disposer d'une variable qui affecte le comportement de réponse mais pas la variable d'intérêt (un "instrument"). Ce type de variable est disponible dans certaines vagues de l'enquête Post formation. Dans ces vagues, un protocole de collecte distinct a été utilisé pour un sur-échantillon tiré aléatoirement. Pour ces sur-échantillons, des relances téléphoniques ont été utilisées afin d'augmenter le taux de réponse. On peut alors utiliser l'appartenance à ce sur-échantillon comme un instrument : l'effort supplémentaire de collecte peut permettre d'atteindre des personnes qui ne répondraient pas dans le protocole classique. Cependant, le fait que le sur-échantillon a été défini aléatoirement assure que l'instrument soit indépendant des variables qu'on souhaite estimer.

Cette étude se propose donc d'appliquer cette méthode de redressement de la non réponse endogène sur les deux vagues de l'enquête Post formation pour lesquelles un sur-échantillonage avec relance téléphonique a été mis en place. Les résultats suggèrent que certains indicateurs issus de l'enquête pourraient être biaisés. Les conséquences opérationnelles de l'utilisation de cette méthode sont ensuite posées. En principe, les modèles de non réponse doivent être utilisées pour chaque variable prise isolément ce qui est coûteux en temps. L'étude s'interroge d'une part sur les critères qui peuvent permettre d'identifier la présence de non réponse endogène, et d'autre part sur la possibilité d'utiliser des modèles pour certains groupes de variables.

Après avoir présenté l'enquête (section 1) et le modèle (section 2.1), nous identifions les variables pour lesquelles un comportement de non réponse endogène est suspecté et comparons les estimations obtenues en corrigeant de ce biais par la méthode d'Heckman (section 2.2). Les conséquences opérationnelles des résultats obtenus sont ensuite discutées.

1 Présentation de l'enquête Post-Formation

L'enquête Post-Formation a été mise en place mi-2019 dans le cadre de l'évaluation du Plan d'investissement dans les compétences. Cette enquête est réalisée chaque trimestre auprès des stagiaires de la formation professionnelle. Ces derniers sont interrogés 6 à 9 mois après leur sortie de formation telle que prévue initialement, y compris si la formation a été interrompue ou abandonnée. Ainsi, la première vague de collecte a eu lieu au cours du troisième trimestre 2019, auprès des stagiaires de la formation professionnelle dont la formation devait s'achever au cours du quatrième trimestre 2018. En régime permanent, la collecte sur les sortants de formation d'un trimestre T_i se fait sur au cours du trimestre T_{i+3}.

L'enquête vise à rendre compte de la perception des stagiaires sur la formation suivie, son déroulé, ainsi que sur ce qu'elle leur a apporté. Plus précisément, le questionnaire comprend une partie sur les obstacles rencontrés pour accéder à la formation, et le cas échéant les raisons d'abandon, le déroulement de la formation (préparation, présence effective de période en entreprise, perception d'avoir été accompagné, mise en contact avec des employeurs, etc.), la qualité et les apports de la formation (compétences acquises, obtention d'une certification, poursuite en formation après une formation d'insertion sociale et professionnelle, etc.), ainsi que sur la situation vis-à-vis de l'emploi (ou l'accès à la formation) au moment de l'enquête (voir détails en table A.1 en annexe).

Le champ de l'enquête couvre l'ensemble du territoire, mais depuis la quatrième vague, la collecte dans les DOM se fait à un rythme semestriel (les stagiaires sont donc interrogés 6 à 12 mois après leur sortie). Chaque trimestre, un échantillon d'environ 75 000 stagiaires de la formation professionnelle sont interrogés pour la France Métropolitaine, auquel s'ajoute entre 15 000 et 20 000 stagiaires pour les territoires ultra-marins. L'échantillon est constitué de manière à être représentatif au niveau régional, à partir de la base Brest [1]. Pour permettre une représentativité au niveau régional, l'échantillon est stratifié par région (et à l'intérieur des régions, par âge, objectif du stage et niveau de diplôme le plus élevé obtenu). Dans les DOM et la Corse, tous les stagiaires concernés sont enquêtés. Dans les autres régions, le taux de sondage est variable de manière à assurer une cible de 2 000 personnes répondant par région.L'enquête a comme objectif initial de fournir des indicateurs pour le Plan d'investissement dans les compétences : les peu diplômés sont donc sur-échantillonné, en visant une proportion de 65% au moins de personnes de niveau de diplôme inférieur au baccalauréat.

En principe, le questionnaire est auto-administré. Les enquêtés reçoivent une lettre-avis accompagnée du questionnaire par voie postale et peuvent remplir le questionnaire papier ou répondre sur internet. Différents type de coordonnées des stagiaires sont disponibles dans la base de sondage et sont mobilisées pour contacter les personnes. Deux vagues de relances sont prévues (par courrier, e-mail et SMS) et la collecte s'étale sur trois mois. Les taux de réponse sont faibles : en moyenne, 34% par vague, avec une tendance à la baisse (voir tableau A.2 en annexe). Les données sont redressées par une méthode classique de calage sur marge (en utilisant les variables de stratification, soit la région, la tranche d'âge, le sexe, l'objectif de la formation et le niveau de diplôme), à partir de la macro SAS Calmar développée par l'Insee.

Sur certaines vagues, des relances téléphoniques ont été mises en place sur une partie de l'échantillon compte tenu du faible taux de collecte. Pour un échantillon supplémentaire, les non répondants 12 jours après le début de l'enquête ont été contactées par téléphone. C'est ce sur-échantillonnage qui

1. La base régionalisée des stagiaires de la formation professionnelle (Brest), construite à partir des fichiers de rémunération des stagiaires, contient les entrées en formation de personnes en recherche d'emploi bénéficiant à minima de la protection sociale.

est utilisé ici pour étudier l'effet de la non réponse endogène : les résultats présentés sont issus de l'exploitation des vagues 3 (menée auprès de sortants de formation au deuxième trimestre 2019) et 10 de l'enquête (sortants au premier trimestre de 2021).

2 Correction de la non réponse endogène : principe et application

2.1 Modèle

La non réponse pose deux problèmes : elle induit d'une part un défaut de couverture de la population totale et, d'autre part un biais de sélection si les personnes qui répondent à l'enquête sont spécifiques. Les méthodes classiques de redressement de la non réponse (calage sur marge, groupes de réponse homogènes) permettent en principe de corriger ces biais. Cependant, elles reposent sur une hypothèse d'indépendance conditionnelle, c'est-à-dire sur le fait que le comportement de non réponse est indépendant des variables d'intérêt, conditionnellement à des variables auxiliaires disponibles dans la base d'échantillonnage. Cette condition peut ne pas être vérifiée. Par exemple il est possible que les personnes qui ont abandonné leur formation avant la fin se sentent moins concernées par une enquête qui porte sur cette formation et qu'elles aient donc une propension plus faible à y répondre. La variable d'intérêt - dans cet exemple l'abandon de la formation - serait alors un déterminant de la non réponse. Il n'est cependant pas possible d'estimer directement cette dépendance, puisque par définition cette variable d'intérêt n'est observée que pour les répondants.

Comme proposé par Castell et Sillard (2021), une solution pour contourner ce problème est de s'appuyer sur le modèle d'Heckman (Heckman, 1979), aussi appelé *Tobit II* (Wooldridge, 2010; Cameron et Trivedi, 2005). Le principe est de modéliser explicitement ce problème d'observation. On suppose l'existence d'une variable latente (ici, la propension à répondre r_i^*), qui va déterminer si une personne répond ou non à l'enquête. Par convention, la personne est un répondant $r_i = 1$ si cette propension est positive $r_i = \mathbf{1}(r_i^* \geqslant 0)$. On modélise alors simultanément la variable d'intérêt y_i et r_i^*, en tenant compte de l'existence d'une corrélation.

Dans le cas d'une variable y_i dichotomique, on peut utiliser un biprobit, qui tient compte de la corrélation entre la variable d'intérêt et la propension à répondre. Plus précisément, en reprenant les notations de Castell et Sillard (2021), on peut écrire le modèle sous la forme suivante :

$$\begin{cases} \text{(i)} & y_i^* &=& c^1 + \mathbf{z}_i\chi + \epsilon_i \\ \text{(ii)} & y_i &=& \mathbf{1}(y_i^* \geqslant 0) \\ \text{(iii)} & r_i^* &=& c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi + \nu_i \\ \text{(iv)} & r_i &=& \mathbf{1}(r_i^* \geqslant 0) \end{cases} \tag{1}$$

où les deux variables y_i^* et r_i^* sont les variables latentes correspondant respectivement au fait d'abandonner y_i et de répondre à l'enquête r_i (y_i^* peut s'interpréter comme une "propension" plus ou moins forte à abandonner une formation en cours, r_i^* comme une propension à répondre à l'enquête) ; z_i correspond aux variables disponibles dans la base d'échantillonnage (et donc observées pour l'ensemble de la population) et w_i est un instrument, c'est-à-dire une variable qui a un effet sur la réponse à l'enquête r_i (Eq. 1-(iii)) mais pas sur le fait d'abandonner une formation en cours y_i (Eq. 1-(i)). Cet instrument est ici le fait d'appartenir au sur-échantillon qui a bénéficié de relances téléphoniques : ces relances sont susceptibles d'augmenter la participation, mais sont indépendantes du

fait de poursuivre ou non une formation jusqu'au bout. Formellement, on peut écrire les conditions d'identifications ainsi :

$$\begin{cases} \mathbb{E}\left(\begin{pmatrix}\nu_i\\ \epsilon_i\end{pmatrix}\middle|\, \mathbf{z}_i, \mathbf{w}_i\right) = 0 \\ \begin{pmatrix}\nu_i\\ \epsilon_i\end{pmatrix} \hookrightarrow \mathcal{N}\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \Sigma\right) \\ \Sigma = \begin{pmatrix}1 & \varrho\\ \varrho & 1\end{pmatrix} \end{cases} \tag{2}$$

où $\mathcal{N}$ est la loi normale bivariée.

Deux méthodes, reposant soit sur l'imputation, soit sur la repondération, sont ensuite possibles pour disposer d'un estimateur sans biais de la variable d'intérêt.

L'estimation par imputation repose entièrement sur le modèle : elle consiste à imputer à partir du modèle les valeurs de la variable d'intérêt pour les non répondants. Plus précisément, on peut estimer par le modèle la propension à abandonner en cours de formation pour les non répondants :

$$\begin{aligned} \mathbb{P}(y_i = 1|\mathbf{z}_i, \mathbf{w}_i, r_i = 0) &= \frac{\mathbb{P}(y_i^* \geqslant 0, r_i^* \leqslant 0|\mathbf{z}_i, \mathbf{w}_i)}{\mathbb{P}(r_i^* \leqslant 0|\mathbf{z}_i, \mathbf{w}_i)} \\ &= \frac{\Phi_2(-(\overset{\delta}{c}^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi), c^1 + \mathbf{z}_i\chi; -\varrho)}{\Phi(-(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi)} \end{aligned}$$

avec $\Phi_2(x, y; \varrho)$ correspondant à la loi normale bivariée de variance unitaire et de corrélation ϱ, estimée au point (x, y). A partir de cette estimation de la probabilité, on peut ensuite imputer, pour chaque non répondant, son comportement d'abandon en tirant dans une loi de Bernouilli de paramètre $\mathbb{P}(y_i = 1|\mathbf{z}_i, \mathbf{w}_i, r_i = 0)$. L'estimation de la moyenne est alors obtenue simplement à partir de la moyenne de ces valeurs, pondérées par les poids initiaux de sondage (voir Castell et Sillard, 2021).

Une autre solution est de repondérer simplement les répondants, à partir d'une repondération de type $(\pi_i\hat{r}_i)^{-1}$, où $\hat{r}_i$ est un estimateur convergent de r_i. Cet estimateur peut être obtenu à partir du modèle. Plus spécifiquement, on a :

$$\mathbb{P}(r_i = 1|\mathbf{z}_i, \mathbf{w}_i, y_i = 1) = \frac{\Phi_2(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi, c^1 + \mathbf{z}_i\chi; \varrho)}{\Phi(c^1 + \mathbf{z}_i\chi)}$$

et

$$\mathbb{P}(r_i = 1|\mathbf{z}_i, \mathbf{w}_i, y_i = 0) = \frac{\Phi_2(c^0 + \mathbf{z}_i\beta + \mathbf{w}_i\psi, -(c^1 + \mathbf{z}_i\chi); -\varrho)}{\Phi(-(c^1 + \mathbf{z}_i\chi))}$$

L'estimation de y repose sur une modélisation jointe de (y_i, r_i). Il n'est donc pas possible d'utiliser les résultats obtenus pour une variable d'intérêt aux autres variables. Il est immédiat de remarquer qu'il est *a priori* faux d'utiliser le modèle 1 correspond à une variable y_1 pour imputer les valeurs manquantes d'une autre variable y_2. Cela revient à supposer que ces deux variables auraient exactement la même corrélation avec le comportement de non réponse, hypothèse qui a de fait peu de chances d'être vérifiée. Mais cette question se pose également pour l'estimateur par repondération, dès lors que la variable y_2 est également corrélée avec le comportement de réponse. En effet, l'estimateur obtenu à partir de la modélisation jointe (y_1, r) ne sera plus convergent conditionnellement à y_2.

Par ailleurs, ce modèle est valable pour une variable qui est posée à l'ensemble des répondants à l'enquête. Certaines variables sont filtrées (par exemple, dans l'enquête Post formation la question sur l'obtention d'un diplôme n'est posée qu'aux personnes ayant déclaré le préparer). En principe, il faudrait donc tenir compte de cette structure en utilisant un modèle emboîté.

Remarques

- En principe, le modèle (1) est identifié même si l'instrument $\mathbf{w}$ n'a pas d'impact sur le comportement de réponse (dans le modèle, si $\psi = 0$). Cependant, dans ce cas, le modèle est identifié uniquement sur la forme fonctionnelle, et donc plus dépendant d'une mauvaise spécification (par exemple, si les distributions sous-jacentes des deux variables latentes y^* et r^* sont éloignées d'une loi normale). En pratique, plus ψ est élevé (i.e. plus l'instrument est fort), et plus les estimateurs seront précisément estimés.
- Si le taux de non réponse est élevé, et si celui-ci est endogène à la variable d'intérêt, alors l'estimation finale reposera de manière importante sur le fait que la non réponse est bien approximée par le modèle joint (1). Celui-ci détermine implicitement une fonction $y^*(r^*)$, à partir de laquelle on peut imputer les valeurs manquantes (ou alternativement une fonction $r^*(y^*)$ qui permet d'estimer la probabilité de non réponse pour les répondants, dans l'estimateur de Hajek). Cependant, cette dépendance n'est estimée qu'à partir de deux points - le fait d'appartenir à l'un ou l'autre des deux échantillons (voir Castell et Sillard, 2021 pour une discussion). Si la forme fonctionnelle est très éloignée d'une forme linéaire, il y a donc un risque que le redressement de la non réponse soit de mauvaise qualité, ce qui est d'autant plus problématique que la non réponse est élevée.
- Une autre manière de discuter cette question est de se référer au cadre utilisée en économétrie. En utilisant la terminologie d'Imbens et Angrist (1994), l'estimation précédente permet d'estimer la corrélation entre la non réponse et la variable d'intérêt uniquement pour les "compliers", c'est-à-dire les personnes pour qui l'instrument modifie effectivement le comportement de réponse : ils ne répondent pas avec le protocole de collecte classique (en auto-administré), mais le font uniquement lorsqu'ils sont contactés par téléphone. Il n'est pas possible en revanche d'estimer cette dépendance pour les "never takers", c'est-à-dire les personnes qui ne répondront pas quelque soit le protocole de collecte. De manière générale, il n'est pas possible de caractériser cette population : pour cela, il faudrait pouvoir l'observer en même temps mais avec deux protocoles de collecte distincts, ce qui est bien sûr impossible [2]. Ici, par construction, on peut identifier *ex ante* certains "never takers". En effet, les personnes pour lesquelles on ne disposent pas de coordonnées téléphoniques ne peuvent pas être contactées pour les relances.

2.2 Application à l'enquête Post-Formation

2.2.1 Validité de l'instrument

Dans le cadre de l'enquête Post-Formation, on sélectionne deux échantillons tirés de manière indépendante dans la même base de sondage. L'instrument $\mathbf{w}_i$ est ici le fait d'appartenir à l'échantillon

2. On peut estimer en revanche la proportion des compliers dans la population totale. Sous l'hypothèse qu'il n'existe pas de personnes qui répondraient dans le protocole classique mais pas dans le cas d'une relance, et comme les deux échantillons sont indépendants, la proportion de "compliers" correspond à la différence de taux de réponse dans l'échantillon avec relance et celui avec le protocole classique. La proportion des never takers correspond au taux de non réponse dans l'échantillon sans réponse

pour lequel une relance téléphonique est prévue. Le fait que les deux échantillons soient tirés de manière indépendante garantit la validité de l'instrument, c'est-à-dire le fait qu'il ne soit pas corrélé avec les comportements que l'on souhaite mesurer. L'ensemble des caractéristiques des deux échantillons sont très proches (voir tableau A.3 en annexe), ce qui est conforme au fait que la sélection dans l'un ou l'autre des échantillons est aléatoire.

La deuxième condition à vérifier est que l'instrument a bien un effet sur le comportement de réponse. Une première indication est donnée simplement par la comparaison des taux de réponses observés dans les deux échantillons (celui uniquement avec collecte auto-administrée et celui avec relances téléphoniques). Le tableau 1 présente les taux de réponse observés pour les deux vagues de l'enquête pour lesquelles un sur-échantillonnage a été mis en place. Pour la vague 3, les taux de réponses de l'échantillon avec relances téléphoniques sont supérieurs de près de 20 points de pourcentage aux taux de réponse de l'échantillon avec le protocole de collecte classique : il est de 34,5 % dans l'échantillon principal, et 54,8% dans l'échantillon avec relances téléphoniques. L'écart est un peu plus faible pour la vague 10 (13 points), ce qui reflète une érosion générale des taux de réponse observée sur l'enquête : le taux de réponse est de 31,9% dans l'échantillon principal, et de 44,9% dans l'échantillon avec relances.

Le tableau 1 présente également des résultats détaillés selon les groupes de répondants. Ces groupes sont définis par le type de coordonnées disponibles (adresse postale, mail, numéro de téléphone mobile). Les taux de réponse sont similaires dans les deux échantillons pour les groupes de personnes dont les coordonnées téléphoniques n'étaient pas disponibles dans l'échantillon (qui représentent un peu plus d'un quart des personnes échantillonnées dans les deux échantillons).. En revanche, lorsque les coordonnées téléphoniques étaient disponibles, les taux de réponses sont nettement plus élevés pour l'échantillon bénéficiant de relances téléphoniques : pour la vague 3, les écarts sont respectivement de 27 points de pourcentage pour les personnes dont toutes les coordonnées étaient disponibles (environ deux tiers de l'échantillon) et 32 points de pourcentage pour ceux dont l'adresse postale et coordonnées téléphoniques étaient disponibles (environ 10% de l'échantillon). Pour la vague 10, ces écarts sont un peu plus faible mais reste élevés : respectivement 17 et 28 points.

2.2.2 Formes réduites : écarts observés dans les deux échantillons

Le tableau 2 présente les moyennes observées dans les deux échantillons pour les principaux indicateurs calculés à partir de l'enquête (cf. tableau A.1). Ces indicateurs correspondent à des informations sur la description du parcours de formation, son déroulé, la satisfaction retirée et la situation au moment de l'enquête, six mois après la formation.

Les moyennes de plusieurs indicateurs sont différentes entre les deux échantillons, ce qui est un indice qu'ils sont corrélés avec la propension à répondre à l'enquête.

Ces écarts peuvent aller dans des sens opposés. Ainsi, le taux d'abandon est plus élevé dans l'échantillon avec relance. C'est également le cas pour le fait de déclarer être en emploi au moment de l'enquête (six mois après la formation), ce qui peut s'expliquer par le fait que les personnes en emploi sont souvent moins disponibles et demandent donc un effort de collecte supplémentaire pour répondre à l'enquête. Le fait d'avoir connu une période en entreprise ou de déclarer des délais courts (moins de trois mois) entre les premières démarches et l'entrée en formation sont également plus fort dans l'échantillon avec relance.

A l'inverse, certaines variables semblent positivement corrélées avec la propension à répondre. En

TABLE 1 – Effet de la relance téléphonique sur le taux de réponse (par groupe de répondants)

Groupe initial	**Echantillon principal**			**Echantillon téléphonique**		
	Part (%)	Taux de réponse (%)	SD	Part (%)	Taux de réponse (%)	SD
Vague 3						
Ensemble	100	34,5	0,5	100	54,8	0,5
G1 : Adr. postale + Mail + Tél.	62,4	37,0	0,5	62,8	63,2	0,5
G2 : Adr. postale + Mail	18,1	33,1	0,5	18,5	34,1	0,5
G3 : Adr. postale + Tél.	11,7	30,2	0,5	11,1	61,8	0,5
G4 : Adresse postale	7,7	24,0	0,4	7,6	25,8	0,4
Vague 10						
Ensemble	100	31,9	0,5	100	44,9	0,5
G1 : Adr. postale + Mail + Tél.	65,6	34,1	0,5	66,1	51,1	0,5
G2 : Adr. postale + Mail	21,9	32,0	0,5	22,3	30,6	0,5
G3 : Adr. postale + Tél.	7,5	20,7	0,4	6,7	48,7	0,5
G4 : Adresse postale	5,0	18,7	0,4	4,9	20,5	0,4

Note : les groupes initiaux sont définis par le type de coordonnées (adresse postale, mail, numéro de téléphone mobile) disponible initialement dans la base de sondage.
Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

comparaison avec l'échantillon principal, on observe en présence de relances une plus faible proportion de stagiaires déclarant avoir bénéficié d'une évaluation de compétences à l'entrée (test utilisé pour permettre une individualisation de la formation), d'avoir du réorganiser leur vie personnelle du fait de la formation, ou d'avoir reçu une aide financière pour la formation. ces variables peuvent être liées à un engagement important de la part des stagiaires, ce qui qui pourrait expliquer qu'ils sont aussi plus enclins ensuite à répondre à des questions s'y rapportant.

En revanche, les moyennes observées dans les deux échantillons sont très proches pour plusieurs variables. Alors qu'on pourrait penser que la satisfaction retirée de l'enquête est un déterminant de la non réponse, les moyennes de la plupart des variables sur la perception subjective de sa qualité et son utilité sont identiques dans les deux échantillons : c'est le cas de l'appréciation de l'utilité de la formation, de la perception d'avoir été suffisamment informé sur la formation, mis en contact avec des employeurs ou encore d'avoir reçu des opportunités d'emploi grâce à la formation.

Pour plusieurs variables, les écarts sont observés pour le groupe de répondants dont les coordonnées téléphoniques n'étaient initialement pas disponibles (voir tableau A.4 en annexe). L'institut de sondage procède à un enrichissement des coordonnées, notamment téléphoniques, pour les personnes dont seule adresse postale était disponible. Des personnes appartenant au groupe initial de répondants "sans téléphone" peuvent avoir finalement été contactées par ce mode. Par ailleurs, le fait d'être plus ou moins équipé peut être corrélé avec l'indicateur que l'on cherche à mesurer : par exemple, le fait de disposer d'une adresse mail signifie que la personne dispose d'une connexion internet, et est donc en capacité de suivre une formation à distance. Les personnes qui ont suivi des formations à distance sont plus susceptibles de répondre seules en ligne au questionnaire : elles représentent une plus faible proportion de l'échantillon avec relances téléphoniques.

TABLE 2 – Moyenne classique des principales variables

Variable	**Vague 3**		**Vague 10**	
	Echantillon principal (%)	**Echantillon téléphonique** (%)	**Echantillon principal** (%)	**Echantillon téléphonique** (%)
Déroulé du parcours de formation				
Abandon	7,0	8,6	8,3	8,9
Obtention d'une certification	85,4	83,8	81,6	81,7
Délais d'entrée en formation court	72,7	77,1	80,1	83,9
Test de compétences	61,4	57,4	59,4	56,3
Modalités de la formation et de son suivi				
Période en entreprise	42,2	46,5	32,8	33,5
Formation à distance	7,0	5,0	46,0	44,1
Réorganisation de la vie personnelle	-	-	36,7	31,9
Aide financière	-	-	16,0	14,0
Satisfaction				
Information suffisante	82,7	83,6	82,5	82,4
Accompagnement suffisant	41,4	40,3	40,9	38,2
Contacts avec des employeurs	25,3	25,0	20,8	20,0
Opportunités d'emploi	51,1	50,3	44,7	44,4
Formation utile	85,7	85,8	83,9	83,6
Suites de la formation				
Poursuite en formation	11,3	10,7	10,5	12,0
En emploi	49,3	54,1	50,7	56,7

Notes : Variables indicatrices reconstruite à partir des réponses à l'enquête post formation. La moyenne classique correspondent ici au valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).
Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête post formation.

2.2.3 Redressement de la non réponse endogène

On applique alors la méthode décrite dans la section précédente aux variables de l'enquête.

Un modèle distinct est estimé pour chaque variable d'intérêt. A titre d'illustration, les estimations du biprobit obtenu en modélisant simultanément le fait d'avoir abandonné la formation avant la fin et le fait d'avoir répondu à l'enquête sont présentées en annexe (A.5 et A.6). Comme suggéré par les taux de réponse observés dans l'échantillon (table 1), le fait d'être inclus dans l'échantillon de relances téléphoniques a un effet très significatif sur le fait de répondre. Par ailleurs, le coefficient de corrélation ρ est significativement négatif et assez élevé pour la vague 3 (-0,23). En revanche, il est plus faible pour la vague 10 (-0,12), et même non significatif. Le signe de la corrélation est conforme avec les écarts observés plus haut des moyennes de l'indicateur dans les deux échantillons (tableau 2). Il peut s'expliquer par l'intuition que les personnes qui ont abandonné une formation avant la fin se sentiraient moins concernées par l'enquête et ont donc une propension à répondre plus faible.

Comme décrit plus haut, ces estimations peuvent être ensuite utilisées pour redresser le biais de non réponse endogène. Le tableau 3 présente les résultats obtenus pour les principaux indicateurs issus de l'enquête. Un modèle spécifique doit être utilisé pour chaque indicateur. Le tableau présente la moyenne de l'indicateur tel qu'obtenu en utilisant une méthode de redressement sur observables, la corrélation estimée par le biprobit entre la variable considérée et la non réponse, et finalement les taux estimés en redressant de la non réponse par imputation [3].

Comme attendu, les corrélations estimées avec la propension à répondre à l'enquête sont faibles et non significatives pour les indicateurs pour lesquels les moyennes calculées entre les deux échantillons sont proches. Pour ces variables, le biais lié à une non réponse endogène peut être négligé, et les différentes méthodes d'estimation fournissent des résultats très proches. En revanche, les écarts sont nettement plus élevées pour les variables pour lesquelles un comportement endogène de réponse a été mis en évidence. En redressant de la non réponse endogène, le taux d'abandon estimé est de 11% pour les deux vagues, alors que l'estimation simple est de 7% (vague 3) ou 8% (vague 10). Une étape supplémentaire sera de vérifier si ces différences sont significatives, ce qui demande de calculer la précision de ces estimateurs (par bootstrap).

La part estimée de formation à distance passe de 6% à 3% après correction de la non réponse endogène pour la vague 3 de l'enquête (qui portent sur des formations qui se sont déroulées avant le premier confinement), et de 45% à 38% pour la vague 10 (sur des formations s'étant achevés au premier trimestre 2021). Après correction de la non réponse endogène, la part des personnes qui auraient passé un test de compétences avant l'entrée en formation passe de 60% à 52% pour la vague 3, et de 59% à 43% pour la vague 10. La part de personnes qui auraient bénéficié de délais courts d'entrée en formation augmente de 74% à 80% (et de 81% à 89% en vague 10).

Pour les deux tiers des 15 indicateurs mesurés sur les deux vagues, les corrélations vont dans un sens identique entre les deux vagues. Cependant, pour 4 variables (abandon, obtention d'une certification, période en entreprise, contacts avec des employeurs), les corrélations significatives observées avec le comportement de réponse en vague 3 ne le sont plus pour la vague 10. L'inverse est observé pour une variable (poursuite en formation). Des différences sur les conditions de collecte (la vague 3 a été prolongée du fait du confinement, la vague 10 porte sur des formations s'étant déroulées en partie pendant), ou de champs (la vague 10 ne porte que sur la France métropolitaine, alors que

3. Pour l'ensemble des variables, les deux méthodes fournissent des valeurs très proches (voir tableau A.7 en annexe). Pour ne pas surcharger la présentation, on choisit donc de ne présenter qu'une seule estimation.

TABLE 3 – Moyenne classique et par repondération après troncature des principales variables

Variable	Moyenne classique (%)	rho	Valeur p	Reponderation après troncature (%)
	Vague 3			
Déroulé du parcours de formation				
Abandon	7,4	-0,23	0,00	11,4
Obtention d'une certification	85,0	0,17	0,01	79,3
Délais d'entrée en formation court	73,8	-0,25	0,00	80,1
Test de compétences	60,4	0,22	0,00	52,2
Modalités de la formation et de son suivi				
Période en entreprise	43,2	-0,25	0,00	52,5
Formation à distance	6,5	0,52	0,00	3,1
Réorganisation de la vie personnelle	-	-	-	-
Aide financière	-	-	-	-
Satisfaction				
Information suffisante	83,0	-0,09	0,10	85,1
Accompagnement suffisant	41,1	0,12	0,01	36,8
Contacts avec des employeurs	25,2	0,04	0,41	23,8
Opportunités d'emploi	50,9	0,06	0,22	48,4
Formation utile	85,7	-0,03	0,57	86,4
Suites de la formation				
Poursuite en formation	11,2	0,03	0,68	10,8
En emploi	50,5	-0,30	0,00	60,8
	Vague 10			
Déroulé du parcours de formation				
Abandon	8,4	-0,12	0,31	10,6
Obtention d'une certification	81,6	-0,11	0,33	84,1
Délais d'entrée en formation court	80,9	-0,37	0,00	88,7
Test de compétences	58,8	0,37	0,00	43,3
Modalités de la formation et de son suivi				
Période en entreprise	32,9	-0,06	0,52	35,0
Formation à distance	45,6	0,16	0,06	39,2
Réorganisation de la vie personnelle	35,8	0,43	0,00	21,5
Aide financière	15,6	0,44	0,03	6,0
Satisfaction				
Information suffisante	82,5	0,04	0,68	81,2
Accompagnement suffisant	40,3	0,24	0,00	31,0
Contacts avec des employeurs	20,6	0,18	0,07	15,8
Opportunités d'emploi	44,6	0,04	0,66	42,9
Formation utile	83,8	0,09	0,35	81,2
Suites de la formation				
Poursuite en formation	10,8	-0,23	0,02	16,0
En emploi	51,9	-0,53	0,00	71,8

Notes : Variables indicatrices reconstruites à partir des réponses à l'enquête Post-Formation. Les groupes avec ou sans téléphone correspondent ici aux personnes pour lesquelles un numéro de téléphone mobile était ou non disponible dans la base de sondage. La moyenne classique correspond ici à la valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Le rho correspond ici au coefficient de corrélation entre les résidus de l'équation d'outcome et de sélection. Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).
Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

la vague 3 inclut les DROM) pourraient constituer des éléments d'explication, mais qui restent à confirmer. L'exploitation de vagues ultérieures de l'enquête, sur lesquelles ce même protocole a été utilisé, pourrait permettre de répondre de manière plus complète.

Les estimations obtenues en redressant la non réponse peut s'éloigner sensiblement des estimateurs classiques. C'est par exemple le cas du taux de retour à l'emploi, pour lequel on observe un écart de près de 20 points entre l'estimateur par pondération sur observables et l'estimateur redressant de la non réponse endogène, pour la vague 10. Ces écarts peuvent être notamment importants lorsqu'on calcule des estimateurs sur des sous-population (par caractéristiques), comme illustré dans le tableau 4. Ces écarts peuvent s'expliquer par le fait que les effectifs sont plus faibles (et donc les estimations moins précises), mais également par le fait que la corrélation entre propension à répondre et variable d'intérêt pourrait varier d'une catégorie à l'autre (ce que ne prend pas en compte ces estimations, qui reposent sur une corrélation "moyenne" sur l'ensemble de la population). Comme discuté plus haut, la qualité du redressement obtenu avec cet estimateur paramétrique dépends de manière cruciale de l'adéquation du modèle paramétrique au "vrai" modèle (inobservé) - et ce d'autant plus que le taux de réponse est faible. Comme discuté par une étude récente de Dutz *et al.* (2021), l'utilisation d'une spécification plus riche, et moins paramétrique, pourrait permettre d'obtenir des estimateurs plus stables.

Par ailleurs, la correction de la non réponse endogène nécessite de faire une modélisation pour chacune des variables d'intérêt. Comme discuté dans la partie **??**, le modèle estimé pour une variable ne s'applique pas pour une autre. Le tableau **??** en fournit une illustration : elle permet de comparer pour l'ensemble des indicateurs considérés la valeur obtenue en utilisant soit le jeu de pondérations correspondant au modèle estimé pour redresser le taux d'abandon, soit celui obtenu avec le modèle spécifique à la variable. Pour la plupart des variables, les écarts sont très importants entre les deux modèles.

Modéliser autant de variables qu'indicateurs peut être problématique notamment pour le calcul de la précision des estimateurs : compte tenu de la complexité de la méthode utilisée, une méthode par bootstrap peut être envisagée mais suppose donc de nombreuses itérations. En termes opérationnels, comme discuté plus haut il est cependant possible d'identifier les variables pour lesquelles la non réponse peut être considérée comme ignorable. Ces variables peuvent être identifiées à partir des statistiques descriptives, en comparant les moyennes de l'indicateur dans les deux échantillons. Le modèle de non réponse peut alors être appliqué uniquement pour les variables pour lesquelles cette analyse en forme réduite suggère qu'il existe un risque de non réponse endogène.

TABLE 4 – Taux d'abandon estimé, par caractéristiques individuelles

Variable	Moyenne classique (%)	Reponderation avant troncature (%)	Reponderation après troncature (%)	Imputation (%)
Vague 3				
Age				
Moins de 25 ans	10,9	18,0	16,9	14,2
25-50 ans	6,0	9,7	9,6	11,5
Plus de 50 ans	6,2	8,7	8,7	10,4
Niveau de diplôme				
Aucun diplôme obtenu	10,8	17,1	16,5	13,9
CEP OU BEPC ou CAP / BAC non obtenu	7,5	12,2	11,9	12,6
Diplôme niveau BAC ou plus	5,4	8,6	8,5	10,9
Niveau non renseigné	11,7	20,5	18,7	14,3
Objectif de stage				
Création d'entreprise	1,1	2,6	2,3	10,3
Formations qualifiantes / professionnalisantes	4,8	7,9	7,8	11,0
Formations préparatoires	10,1	15,8	15,4	13,1
Adaptation au poste (AFPR-POEI)	11,6	17,9	17,5	14,5
Objectif non renseigné	11,1	17,3	16,7	13,6
Sexe				
Femmes	7,5	11,4	11,3	11,9
Hommes	7,2	12,2	11,6	12,4
Vague 10				
Age				
Moins de 25 ans	11,4	15,0	14,6	11,9
25-50 ans	7,4	9,6	9,6	10,6
Plus de 50 ans	7,9	9,7	9,7	10,1
Niveau de diplôme				
Aucun diplôme obtenu (niveau VI)	10,3	13,5	13,2	11,9
CEP OU BEPC ou CAP / BAC non obtenu	9,5	12,4	12,4	11,3
Diplôme niveau BAC ou plus	6,6	8,3	8,3	9,9
Niveau non renseigné	11,9	17,8	16,1	12,6
Objectif de stage				
Création d'entreprise	4,2	6,1	6,1	8,4
Formations qualifiantes / professionnalisantes	7,6	9,9	9,8	10,6
Formations préparatoires	10,3	13,2	13,1	11,4
Adaptation au poste (AFPR-POEI)	9,9	12,6	12,7	11,2
Objectif non renseigné	11,1	14,5	13,9	12,1
Sexe				
Femmes	9,0	11,4	11,3	10,8
Hommes	7,8	10,4	10,2	10,8

Notes : La taux d'abandon est une variable indicatrice reconstruite à partir des réponses à l'enquête Post-Fformation.
Champ : sortants de formation de la France entière au T2 2019 (vague 3) et de la France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

TABLE 5 – Estimateur d'Heckman par repondération après troncature avec le modèle propre et le modèle d'abandon

Variable	Vague 3		Vague 10	
	Modèle propre (%)	**Modèle abandon** (%)	**Modèle propre** (%)	**Modèle abandon** (%)
Déroulé du parcours de formation				
Abandon	**11,4**	**11,4**	**10,7**	**10,7**
Obtention d'une certification	53,1	69,0	84,1	85,5
Délais d'entrée en formation court	80,1	82,1	88,7	84,5
Test de compétences	52,2	72,1	43,1	65,3
Modalités de la formation et de son suivi				
Période en entreprise	52,8	54,8	34,7	39,1
Formation à distance	3,1	10,7	39,7	52,3
Réorganisation de la vie personnelle	-	-	21,5	42,4
Aide financière	-	-	6,0	19,6
Satisfaction				
Information suffisante	85,1	89,2	81,2	86,0
Accompagnement suffisant	37,5	52,1	30,8	47,1
Contacts avec des employeurs	23,8	36,2	15,6	25,5
Opportunités d'emploi	48,4	63,7	42,6	51,4
Formation utile	86,4	91,1	81,2	87,2
Suites de la formation				
Poursuite en formation	10,8	17,5	16,0	13,5
En emploi	60,8	63,0	71,8	58,4

Notes : Variables indicatrices reconstruites à partir des réponses à l'enquête Post-Formation. Les groupes avec ou sans téléphone correspondent ici aux personnes pour lesquelles un numéro de téléphone mobile était ou non disponible dans la base de sondage. La moyenne classique correspond ici à la valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Le rho correspond ici au coefficient de corrélation entre les résidus de l'équation d'outcome et de sélection. Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).
Champ : sortants de formation de France entière au T2 2019 (vague 3) et de France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

Discussion

Cette étude s'intéresse à la détection et à la correction de la non réponse endogène dans une enquête menée auprès des sortants de stages de la formation professionnelle. Elle illustre que pour plusieurs variables, il existe une forte présomption de réponse endogène, ce qui signifie que les estimateurs obtenus pourraient être biaisés. Les redressements opérés, en mobilisant le modèle de sélection d'Heckman (adaptée ici à une variable d'intérêt binaire), mettent en évidence des écarts parfois importants pour ces variables.

Cette méthode de redressement de la non réponse endogène a plusieurs avantages. Les relances téléphoniques permettent d'obtenir un taux de réponse nettement plus élevé, mais ont un coût également nettement plus important. Le protocole, qui repose sur le fait de mettre en oeuvre une collecte renforcée par téléphone que sur un sous-échantillon de taille réduite, peut donc permettre de redresser la non réponse tout en limitant le budget nécessaire. Les estimateurs reposant sur la modélisation peuvent être cependant moins précis, ce qui peut conduire à un arbitrage pour déterminer la taille de l'échantillon de relance. Cependant, comme discuté plus haut, les estimations finales reposeront sur une estimation, et la confiance dans l'estimateur sera d'autant plus importante que le taux de réponse est élevé. Comme discuté par Dutz *et al.* (2021), l'utilisation de méthodes paramétriques peut soulever des questions, d'autant que la propension à répondre peut être lié à plusieurs dimensions, que ne capte pas l'instrument actuel. Ce problème de spécification est d'autant plus aigu pour les variables filtrées. Enfin, de manière pratique, la modélisation par variable, si elle n'est pas insurmontable, peut être chronophage en particulier car le calcul de précision nécessite d'utiliser des méthodes de bootstrap.

Par ailleurs, on peut s'interroger sur le fait que le protocole de collecte utilisé ici, pour augmenter le taux de réponse, repose sur des relances téléphoniques. L'hypothèse sous-jacente est qu'il n'existe pas d'"effet de mode". Cette question est débattue : par exemple, un biais de désirabilité sociale peut conduire certaines personnes à répondre différemment lorsqu'elles sont interrogées par un enquêteur, ou lorsqu'elles répondent seules à un questionnaire. La médiation d'un enquêteur peut également permettre de mieux comprendre certaines questions, qui seraient sinon survolées ou auxquelles des personnes auraient tendance à répondre de manière mécanique. Le protocole utilisé ici permet "d'aller chercher", par des relances téléphoniques, des personnes qui ont une trop faible propension à répondre pour le faire lorsqu'ils reçoivent simplement des sollicitations à remplir seul un questionnaire. Mais cela peut aussi induire un biais, si certaines de ces personnes ont aussi tendance à répondre différemment au questionnaire qu'ils ne l'auraient fait seuls.

Il n'est pas possible ici d'identifier séparément ces deux effets avec le protocole utilisé ici [4]. Une adaptation du protocole pourrait être envisagée. En plus de l'échantillon pour lequel une collecte auto-administrée est prévue, on pourrait prévoir cette fois deux sur-échantillons distincts : outre l'échantillon utilisant des relances téléphoniques en complément de la collecte auto-administrée (comme dans l'étude ici), un autre échantillon indépendant pourrait bénéficié uniquement d'une collecte par téléphone [5].

4. Typiquement, un estimateur naïf qui chercherait à mesurer l'effet de la collecte par téléphone en introduisant une indicatrice capte non seulement l'effet du mode, mais aussi le fait que les personnes qui répondent par téléphone sont aussi celles qui ont une propension différente à répondre, propension qui est elle-même lié à la valeur de la variable.

5. Ici, on capturerait l'effet du mode en introduisant une indicatrice d'appartenance à l'échantillon "pur téléphone". Pour tenir compte du fait que cette variable capte aussi une plus forte propension à répondre, elle serait également introduite dans l'équation de participation, avec toujours comme variable d'exclusion

Références

Cameron, A. C. et Trivedi, P. K. (2005). *Microeconometrics : methods and applications*. Cambridge university press.

Castell, L. et Sillard, P. (2021). Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman. Rapport technique M2021-02, Institut National de la Statistique et des Études Économiques.

Dutz, D., Huitfeldt, I., Lacouture, S., Mogstad, M., Torgovitsky, A. et van Dijk, W. (2021). Selection in surveys. Working Paper 29549, National Bureau of Economic Research.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica : Journal of the econometric society*, pages 153–161.

Imbens, G. W. et Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

une indicatrice d'appartenance au mode de collecte mixte.

Annexe A : tables complémentaires

TABLE A.1 – Description des principales variables

Variable indicatrice	**Définition**
	Déroulé du parcours de formation
Abandon	Le sortant a abandonné sa formation avant la fin
Obtention d'une certification*	Le sortant a obtenu en intégralité ou en partie une certification
Délais d'entrée en formation court	Le sortant a déclaré un délai de 3 mois ou moins entre les démarches et le début de la formation
Test de compétences	Le sortant a fait une évaluation des compétences (test de positionnement) avant le début de la formation
	Modalités de la formation et de son suivi
Période en entreprise	Le sortant a déclaré qu'au moins une partie de la formation se déroulait en entreprise
Formation à distance	Le sortant a déclaré que la formation se déroulait au moins en partie à distance
Réorganisation de la vie personnelle	Le sortant a déclaré avoir du réorganiser sa vie personnelle pour suivre la formation
Aide financière*	Le sortant a déclaré avoir reçu une aide financière pour réorganiser sa vie personnelle
	Satisfaction
Information suffisante	Le sortant considère avoir été bien informé sur le contenu de la formation
Accompagnement suffisant	Le sortant considère avoir été accompagné pendant la formation pour préparer sa recherche d'emploi
Contacts avec des employeurs	Le sortant a déclaré avoir été mis en contact par l'organisme de formation avec des employeurs potentiels
Opportunités d'emploi	Le sortant a déclaré avoir obtenu des opportunités d'emploi grâce à la formation
Formation utile	Le sortant considère que la formation était utile
	Suites de la formation
Poursuite en formation	Le sortant a déclaré être de nouveau en formation à la date d'enquête
En emploi	Le sortant a déclaré être, à la date d'enquête, travailleur indépendant, salarié d'une entreprise, salarié d'un ou de particulier(s), fonctionnaire ou avoir une promesse d'embauche.

Notes : Variables indicatrices reconstruite à partir des réponses à l'enquête post formation. *Ces variables sont conditionnelles.
Source : Dares, exploitation des vagues 3 et 10 de l'enquête post formation.

TABLE A.2 – Présentation de l'enquête Post-Formation

Vague	Sortants	Champ	Collecte	Taux de sondage (effectifs échantillons)	Taux de réponse (effectifs enquêtés)
1	T4 2018	France entière	T3 2019	43% (75 000)	35% (26 000)
2	T1 2019	France entière	T4 2019	51% (90 000)	33% (30 000)
3	T2 2019	France entière	T1 2020	37% (90 000)	38% (34 000)
4	T3 2019	France métro.	T2 2020	45% (75 000)	36% (27 000)
5	T4 2019	France métro.	T3 2020	42% (90 000)	35% (32 000)
	S2 2019	DOM			
6	T1 2020	France métro.	T4 2020	47% (71 500)	37% (27 000)
7	T2 2020	France métro.	T1 2021	48% (86 000)	37% (32 000)
	S1 2020	DOM			
8	T3 2020	France métro.	T2 2021	41% (71 500)	32% (23 000)
9	T4 2020	France métro.	T3 2021	35% (91 500)	28% (26 000)
	S2 2020	DOM			
10	T1 2021	France métro.	T4 2021	29% (71 500)	34% (24 000)

Source : Dares.

TABLE A.3 – Caractéristiques initiales des deux échantillons

Variable	Vague 3				Vague 10			
	Echantillon principal		Echantillon téléphonique		Echantillon principal		Echantillon téléphonique	
	Effectifs	%	Effectifs	%	Effectifs	%	Effectifs	%
Ensemble	74498	100%	14902	100%	60736	100%	10721	100%
Par sexe								
Femmes	35596	47,8	7070	47,4	28059	46,2	5007	46,7
Hommes	38902	52,2	7833	52,6	32677	53,8	5714	53,3
Par âge								
Moins de 25 ans	20527	27,6	4095	27,5	14414	23,7	2537	23,7
25-50 ans	44854	60,2	8984	60,3	37826	62,3	6685	62,4
Plus de 50 ans	9117	12,2	1824	12,2	8496	14,0	1499	14,0
Par niveau de diplôme								
Infra BAC (ou niveau BAC non obtenu)	42681	57,3	8546	57,3	37201	61,3	6575	61,3
Diplôme niveau BAC ou plus	26826	36,0	5363	36,0	21400	35,2	3783	35,3
Niveau non renseigné	4991	6,7	994	6,7	2135	3,5	363	3,4
Par objectif du stage								
Création d'entreprise	2226	3,0	437	2,9	2340	3,9	411	3,8
Formations qualifiantes /professionnalisantes	39419	52,9	7881	52,9	37680	62,0	6654	62,1
Formations préparatoires	8864	11,9	1779	11,9	11870	19,5	2105	19,6
Adaptation au poste (AFPR-POEI)	6524	8,8	1304	8,7	3989	6,6	700	6,5
Objectif non renseigné	17465	23,4	3502	23,5	4857	8,0	851	7,9
Par région de résidence								
Auvergne Rhone Alpes	5484	7,4	1094	7,3	4984	8,2	882	8,2
Bourgogne Franche Comte	5484	7,4	1096	7,4	4980	8,2	881	8,2
Bretagne	5460	7,3	1091	7,3	4976	8,2	872	8,1
Centre Val De Loire	5473	7,3	1099	7,4	4976	8,2	871	8,1
Corse	1146	1,5	231	1,6	981	1,6	174	1,6
Grand Est	5477	7,4	1095	7,3	4981	8,2	879	8,2
Guadeloupe	1056	1,4	204	1,4	-	-	-	-
Guyane	646	0,9	125	0,8	-	-	-	-
Hauts De France	5479	7,4	1094	7,3	4984	8,2	878	8,2
Ile De France	5483	7,4	1106	7,4	4975	8,2	885	8,3
La Reunion	3815	5,1	762	5,1	-	-	-	-
Martinique	1753	2,4	351	2,4	-	-	-	-
Mayotte	458	0,6	86	0,6	-	-	-	-
Normandie	5443	7,3	1094	7,3	4980	8,2	878	8,2
Nouvelle Aquitaine	5481	7,4	1100	7,4	4981	8,2	879	8,2
Occitanie	5483	7,4	1100	7,4	4978	8,2	887	8,3
Pays De La Loire	5450	7,3	1085	7,3	4981	8,2	881	8,2
Provence Alpes Cote D Azur	5427	7,3	1090	7,3	4979	8,2	874	8,2
Par groupe de disponibilité des coordonnées								
Gr. 1 : Adr. postale + Mail + Tél.	46519	62,4	9356	62,8	39848	65,6	7087	66,1
Gr. 2 : Adr. postale + Mail	13475	18,1	2754	18,5	13284	21,9	2395	22,3
Gr. 3 : Adr. postale + Tél.	8735	11,7	1657	11,1	4558	7,5	716	6,7
Gr. 4 : Adresse postale	5769	7,7	1135	7,6	3046	5,0	523	4,9

Lecture : L'ensemble des caractéristiques initiales des deux échantillons sont bien identiques, ce qui est conforme au fait que la sélection dans l'un ou l'autre des échantillons est aléatoire.
Champ : sortants de formation de la France entière au T2 2019 (vague 3) et de la France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

TABLE A.4 – Moyenne classique des principales variables

Groupe initial	Vague 3		Vague 10	
	Echantillon principal (%)	Echantillon téléphonique (%)	Echantillon principal (%)	Echantillon téléphonique (%)
Déroulé du parcours de formation				
Abandon				
Avec téléphone (G1 & G3)	6,6	8,7	8	8,7
Sans téléphone (G2 & G4)	8	8,2	9	9,5
Obtention d'une certification				
Avec téléphone (G1 & G3)	85,9	84	82,2	82,2
Sans téléphone (G2 & G4)	83,7	82,8	79,8	79,4
Délais d'entrée en formation court				
Avec téléphone (G1 & G3)	73,4	77,9	80,5	84,4
Sans téléphone (G2 & G4)	70,5	72,7	79,2	81,2
Test de compétences				
Avec téléphone (G1 & G3)	61,2	57,4	59,6	56
Sans téléphone (G2 & G4)	61,7	57,6	58,8	57,7
Modalités de la formation et de son suivi				
Période en entreprise				
Avec téléphone (G1 & G3)	41,8	46,2	32,8	33,2
Sans téléphone (G2 & G4)	43,3	47,9	32,7	35,4
Formation à distance				
Avec téléphone (G1 & G3)	7,2	5,1	45,8	43,2
Sans téléphone (G2 & G4)	6,5	4,3	46,6	48,4
Réorganisation de la vie personnelle				
Avec téléphone (G1 & G3)	-	-	36,4	31
Sans téléphone (G2 & G4)	-	-	37,6	36,3
Aide financière				
Avec téléphone (G1 & G3)	-	-	16,6	14,9
Sans téléphone (G2 & G4)	-	-	14,3	10,4
Satisfaction				
Information suffisante				
Avec téléphone (G1 & G3)	82,6	83,9	83,1	82,8
Sans téléphone (G2 & G4)	83,1	81,9	80,8	80,4
Accompagnement suffisant				
Avec téléphone (G1 & G3)	41,3	40,4	40,7	37,4
Sans téléphone (G2 & G4)	41,7	39,9	41,3	42,2
Contacts avec des employeurs				
Avec téléphone (G1 & G3)	25	25,1	20,5	19,7
Sans téléphone (G2 & G4)	26,1	24,7	21,6	21,5
Opportunités d'emploi				
Avec téléphone (G1 & G3)	51,2	50,1	44,5	44,4
Sans téléphone (G2 & G4)	50,9	51,1	45,1	44,2
Formation utile				
Avec téléphone (G1 & G3)	85,8	85,9	84,2	83,7
Sans téléphone (G2 & G4)	85,4	85,5	82,9	83
Suites de la formation				
Poursuite en formation				
Avec téléphone (G1 & G3)	10,9	10,7	10,5	11,9
Sans téléphone (G2 & G4)	12,7	10,9	10,7	12,7
En emploi				
Avec téléphone (G1 & G3)	49,6	54,9	50,3	57,3
Sans téléphone (G2 & G4)	48,3	49,7	52	53,7

Notes : Variables indicatrices reconstruite à partir des réponses à l'enquête Post-Formation. Les groupes avec ou sans téléphone correspondent ici aux personnes pour lesquelles un numéro de téléphone mobile était ou non disponible dans la base de sondage. La moyenne classique correspond ici à la valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).
Champ : sortants de formation de la France entière au T2 2019 (vague 3) et de la France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.

TABLE A.5 – Coefficients du modèle probit bivarié pour le taux d'abandon-Vague 3

Variable	Coefficient	SD	Valeur t	Valeur p	Coefficient	SD	Valeur t	Valeur p
	Équation de sélection				Équation d'outcome			
Intercept	-0,75	0,03	-23,31	0,00	-1,75	0,16	-10,73	0,00
Echantillon Relance	0,53	0,01	45,61	0				
Formations qualifiantes / professionnalisantes	0,26	0,03	9,74	0,00	0,50	0,12	4,03	0,00
Formations préparatoires	0,18	0,03	6,14	0,00	0,87	0,13	6,95	0,00
Adaptation au poste (AFPR-POEI)	0,05	0,03	1,62	0,10	1,01	0,13	8,03	0,00
Objectif non renseigné	0,28	0,03	9,78	0,00	0,82	0,13	6,57	0,00
25-50 ans	0,34	0,01	32,04	0,00	-0,28	0,03	-10,08	0,00
Plus de 50 ans	0,69	0,02	44,97	0	-0,34	0,04	-7,82	0,00
Hommes	-0,26	0,01	-28,80	0,00	-0,01	0,02	-0,40	0,69
Niveau non renseigné	-0,01	0,02	-0,72	0,47	0,10	0,04	2,37	0,02
Diplôme niveau BAC ou plus	0,18	0,01	18,91	0,00	-0,20	0,02	-8,41	0,00
Groupe2	-0,18	0,01	-15,53	0,00	0,07	0,03	2,11	0,04
Groupe3	-0,21	0,01	-14,93	0,00	0,15	0,03	4,69	0,00
Groupe4	-0,47	0,02	-25,23	0,00	0,23	0,05	4,26	0,00
Bourgogne Franche Comte	0,01	0,02	0,57	0,57	0,23	0,05	4,25	0,00
Bretagne	0,00	0,02	0,15	0,88	0,20	0,06	3,56	0,00
Centre Val De Loire	0,01	0,02	0,48	0,63	0,08	0,06	1,37	0,17
Corse	-0,29	0,04	-7,09	0,00	0,22	0,10	2,26	0,02
Grand Est	0,01	0,02	0,39	0,70	0,20	0,05	3,61	0,00
Guadeloupe	-0,19	0,04	-4,75	0,00	-0,02	0,11	-0,19	0,85
Guyane	-0,30	0,05	-5,85	0,00	0,24	0,12	1,94	0,05
Hauts De France	-0,03	0,02	-1,11	0,27	0,11	0,06	1,96	0,05
Ile De France	-0,07	0,02	-3,12	0,00	0,07	0,06	1,21	0,22
La Reunion	-0,17	0,03	-6,49	0,00	0,10	0,07	1,47	0,14
Martinique	-0,40	0,03	-11,85	0,00	0,11	0,09	1,19	0,23
Mayotte	-0,10	0,06	-1,61	0,11	0,11	0,14	0,82	0,41
Normandie	0,03	0,02	1,43	0,15	0,12	0,06	2,19	0,03
Nouvelle Aquitaine	0,03	0,02	1,33	0,18	0,05	0,06	0,94	0,35
Occitanie	-0,05	0,02	-2,23	0,03	0,10	0,06	1,83	0,07
Pays De La Loire	0,08	0,02	3,40	0,00	0,14	0,06	2,54	0,01
Provence Alpes Cote D Azur	-0,06	0,02	-2,55	0,01	0,15	0,06	2,52	0,01
Rho					-0,23	0,06	-3,81	0,00

Champ : sortants de formation au T2 2019 en France entière.
Source : Dares, exploitation de la vague 3 de l'enquête Post-Formation.

TABLE A.6 – Coefficients du modèle probit bivarié pour le taux d'abandon - Vague 10

Variable	Coefficient	SD	Valeur t	Valeur p	Coefficient	SD	Valeur t	Valeur p
	Équation de sélection				Équation d'outcome			
Intercept	-0,81	0,03	-24,27	0,00	-1,28	0,19	-6,85	0,00
Echantillon Relance	0,36	0,01	26,45	0,00				
Formations qualifiantes / professionnalisantes	0,12	0,03	4,40	0,00	0,26	0,08	3,39	0,00
Formations préparatoires	0,11	0,03	3,98	0,00	0,42	0,08	5,18	0,00
Adaptation au poste (AFPR-POEI)	0,04	0,03	1,19	0,23	0,42	0,09	4,85	0,00
Objectif non renseigné	0,22	0,03	6,60	0,00	0,36	0,09	3,97	0,00
25-50 ans	0,32	0,01	25,69	0,00	-0,22	0,04	-5,45	0,00
Plus de 50 ans	0,71	0,02	42,03	0	-0,23	0,07	-3,43	0,00
Hommes	-0,20	0,01	-19,84	0,00	-0,10	0,03	-3,51	0,00
Niveau non renseigné	-0,09	0,03	-2,80	0,01	0,03	0,08	0,32	0,75
Diplôme niveau BAC ou plus	0,25	0,01	23,54	0,00	-0,18	0,03	-5,64	0,00
Groupe2	-0,12	0,01	-9,93	0,00	0,07	0,03	2,31	0,02
Groupe3	-0,40	0,02	-20,21	0,00	0,09	0,06	1,45	0,15
Groupe4	-0,56	0,03	-20,54	0,00	0,01	0,10	0,11	0,92
Bourgogne Franche Comte	0,06	0,02	2,40	0,02	0,06	0,06	1,05	0,29
Bretagne	0,02	0,02	0,95	0,34	-0,01	0,06	-0,18	0,86
Centre Val De Loire	0,04	0,02	1,45	0,15	-0,04	0,06	-0,75	0,45
Corse	-0,25	0,04	-5,61	0,00	0,13	0,10	1,21	0,23
Grand Est	0,02	0,02	0,82	0,41	0,00	0,06	0,02	0,99
Guadeloupe								
Guyane								
Hauts De France	-0,02	0,02	-0,95	0,34	0,02	0,06	0,27	0,78
Ile De France	-0,13	0,02	-5,43	0,00	0,01	0,06	0,09	0,93
La Reunion								
Martinique								
Mayotte								
Normandie	0,02	0,02	0,65	0,51	-0,01	0,06	-0,21	0,83
Nouvelle Aquitaine	0,02	0,02	0,89	0,38	0,00	0,06	0,07	0,94
Occitanie	0,00	0,02	-0,11	0,92	0,01	0,06	0,09	0,93
Pays De La Loire	0,11	0,02	4,36	0,00	-0,03	0,06	-0,44	0,66
Provence Alpes Cote D Azur	-0,05	0,02	-2,10	0,04	-0,05	0,06	-0,92	0,36
Rho					-0,12	0,11	-1,02	0,31

Champ : sortants de formation au T1 2021 en France métropolitaine.
Source : Dares, exploitation de la vague 10 de l'enquête Post-Formation.

TABLE A.7 – Moyenne classique et par repondération après troncature des principales variables

Variable	Moyenne classique (%)	rho	SD	Valeur p	Reponderation avant troncature (%)	Reponderation après troncature (%)	Imputation (%)
				Vague 3			
Déroulé du parcours de formation							
Abandon	7,4	-0,23	0,06	0,00	11,8	11,4	12,1
Obtention d'une certification	85,0	0,17	0,06	0,01	78,3	79,3	78,1
Délais d'entrée en formation court	73,8	-0,25	0,05	0,00	80,1	80,1	80,4
Test de compétences	60,4	0,22	0,05	0,00	52,0	52,2	51,9
Modalités de la formation et de son suivi							
Période en entreprise	43,2	-0,25	0,04	0,00	52,7	52,5	53,2
Formation à distance	6,5	0,52	0,10	0,00	3,1	3,1	3,0
Réorganisation de la vie personnelle	-	-	-	-	-	-	-
Aide financière	-	-	-	-	-	-	-
Satisfaction							
Information suffisante	83,0	-0,09	0,06	0,10	85,1	85,1	85,3
Accompagnement suffisant	41,1	0,12	0,05	0,01	36,7	36,8	37,0
Contacts avec des employeurs	25,2	0,04	0,05	0,41	23,7	23,8	23,3
Opportunités d'emploi	50,9	0,06	0,05	0,22	48,3	48,4	47,2
Formation utile	85,7	-0,03	0,06	0,57	86,4	86,4	86,3
Suites de la formation							
En emploi	50,5	-0,30	0,04	0,00	60,9	60,8	59,8
Poursuite en formation	11,2	0,03	0,07	0,68	10,8	10,8	10,8
				Vague 10			
Déroulé du parcours de formation							
Abandon	8,4	-0,12	0,11	0,31	10,7	10,6	10,7
Obtention d'une certification	81,6	-0,11	0,11	0,33	84,3	84,1	84,5
Délais d'entrée en formation court	80,9	-0,37	0,10	0,00	88,8	88,7	88,7
Test de compétences	58,8	0,37	0,07	0,00	42,9	43,3	42,6
Modalités de la formation et de son suivi							
Période en entreprise	32,9	-0,06	0,09	0,52	35,0	35,0	35,5
Formation à distance	45,6	0,16	0,08	0,06	39,1	39,2	38,4
Réorganisation de la vie personnelle	35,8	0,43	0,08	0,00	21,3	21,5	21,2
Aide financière	15,6	0,44	0,20	0,03	5,4	6,0	5,5
Satisfaction							
Information suffisante	82,5	0,04	0,10	0,68	81,2	81,2	81,5
Accompagnement suffisant	40,3	0,24	0,08	0,00	30,8	31,0	30,8
Contacts avec des employeurs	20,6	0,18	0,10	0,07	15,7	15,8	15,7
Opportunités d'emploi	44,6	0,04	0,08	0,66	42,8	42,9	42,7
Formation utile	83,8	0,09	0,10	0,35	81,2	81,2	81,0
Suites de la formation							
Poursuite en formation	10,8	-0,23	0,10	0,02	16,3	16,0	16,0
En emploi	51,9	-0,53	0,06	0,00	72,2	71,8	72,1

Notes : Variables indicatrices reconstruites à partir des réponses à l'enquête Post-Formation. Les groupes avec ou sans téléphone correspondent ici aux personnes pour lesquelles un numéro de téléphone mobile était ou non disponible dans la base de sondage. La moyenne classique correspond ici à la valeur moyenne pondérée par les poids issus de la macro SAS CALMAR (Insee). Le rho correspond ici au coefficient de corrélation entre les résidus de l'équation d'outcome et de sélection.Pas d'information pour réorganisation de la vie personnelle et aide financière en vague 3, car une partie des répondants n'a pas pu répondre à ces deux questions. Par erreur ils ont répondu au questionnaire de la vague 2 (la question a été ajoutée en vague 3).
Champ : sortants de formation de la France entière au T2 2019 (vague 3) et de la France métropolitaine au T1 2021 (vague 10).
Source : Dares, exploitation des vagues 3 et 10 de l'enquête Post-Formation.