Présentation de l'enquête de cheminement tous niveaux de sortie du système éducatif

Génération 2017 (1ère interrogation)

Sommaire

1. Contexte de l'enquête 3
1.1 - Objectifs détaillés et principaux thèmes abordés 3
1.2 – Pilotage de l'enquête et calendrier 11
2. Bilan d'exécution de l'enquête et des résultats produits 12
3. Méthodologie statistique 16
3.1 - Champ, unités enquêtées 16
3.2 - Paramètres d'intérêt de l'enquête 17
3.3 - Description du sondage 19
3.4 - Traitements statistiques 20

1. Contexte de l'enquête

1.1 - Objectifs détaillés et principaux thèmes abordés

Les enquêtes « Génération » s'intéressent à l'insertion et au cheminement des sortants du système éducatif lors de leurs premières années de vie active. Elles ont pour objectifs principaux de produire des indicateurs d'insertion (taux d'emploi, taux de chômage, taux d'emploi à durée indéterminée, etc.), selon les niveaux de formation, les filières, les spécialités, à destination des acteurs publics et sociaux. Elles proposent ainsi des informations qui contribuent à la compréhension des processus d'insertion et des différenciations des parcours en début de carrière.

L'enquête réalisée en 2020 auprès de la Génération 2017 a permis d'analyser l'insertion des jeunes sortis de formation initiale durant l'année scolaire 2016-2017 sur leurs trois premières années de vie active.

- **Génération 2017, une enquête rénovée**

La première enquête Génération a porté sur la Génération 1992, interrogée 5 ans après sa sortie d'études. A partir de la Génération 1998, le Céreq a mis en place un dispositif d'enquêtes régulier afin d'étudier tous les trois ans l'accès à l'emploi et les premières mobilités professionnelles des jeunes à l'issue de leur formation initiale. Le dispositif était alors structuré en deux types d'opérations en alternance :

- Une enquête Génération panélisée (socle de 30 000 à 60 000 questionnaires) comprenant une première interrogation 3 ans après la sortie du système éducatif et deux ré-interrogations 5 et 7 ans après la sortie. La Génération 1998 a bénéficié d'une 4ème interrogation 10 ans après la fin de la formation initiale.
- Une enquête sur une taille plus réduite (socle de 25 000 questionnaires environ) comprenant seulement une interrogation à trois ans. L'enquête Génération 2013 est de ce type.

Suite à la Génération 2013, le Céreq a rénové le dispositif Génération. L'enquête Génération 2017 est la première enquête entrant dans ce nouveau dispositif. Voici les principales évolutions opérées :

- Modification de l'architecture du dispositif :
  Passage de 3 à 4 années d'intervalle entre deux Générations
  Toutes les Générations sont interrogées à 3 et 6 ans après la sortie des études (au lieu d'alterner des Générations « pleines » interrogées à 3-5-7 ans, et des Générations « légères » interrogées une seule fois à 3 ans)

- Le champ de l'enquête : le champ demeure celui des primo-sortants de formation initiale de tous niveaux de formation. Trois changements sont introduits : la redéfinition des périodes de césure (de 12 à 16 mois), l'inclusion des sortants de contrats de professionnalisation dans le champ (au même titre que les sortants de contrats d'apprentissage), et l'inclusion des résidents à l'étranger au moment de l'enquête.

  Le mode de collecte : passage d'une collecte monomode (téléphone) à une collecte multimode (internet et téléphone). Le mode de collecte est séquentiel et concurrentiel : les enquêtés sont, dans un premier temps, incités à répondre par internet, puis ils sont relancés par téléphone, le mode internet restant disponible.

- **Objectifs et contenu du questionnaire**

Le tableau 1 présente les principales évolutions des modules thématiques entre l'enquête de 2013 auprès de la Génération 2010 et l'enquête de 2020 auprès de la Génération 2017 (hors thématiques financées par des partenaires de l'enquête). La comparaison n'est pas faite avec l'enquête de 2016 auprès de la Génération 2013 car il s'agit d'une enquête dite « légère » où le volume de répondants et la durée du questionnaire sont réduits.
Au cours du processus de rénovation, des choix ont été faits visant à définir un « tronc commun » de questionnement, constitutif du cœur de l'enquête. Les objectifs de ce tronc commun sont d'identifier les informations qui apparaissent nécessaires pour reconstruire les parcours des jeunes et qui représentent des « forces » de l'enquête pour la réalisation d'études. Ainsi, l'architecture de ce tronc commun s'appuie sur deux piliers : d'une part, le parcours scolaire, d'autre part, le calendrier professionnel. La richesse des analyses produites dépend à la fois de la richesse des informations collectées et de leur pertinence dans l'éclairage des processus d'insertion.

<u>En amont : le parcours scolaire</u>

Celui-ci vise à identifier le plus haut diplôme atteint, clé de voute des analyses produites par l'enquête.

Les derniers travaux sur la Génération 2010 pointent que, au-delà du plus haut niveau de diplôme atteint, le parcours scolaire antérieur peut influer sur la réussite du parcours professionnel ultérieur. Il s'agit donc de saisir l'hétérogénéité des différents cursus (en même temps que leur richesse) en récapitulant l'ensemble des diplômes obtenus.

Cependant, il s'agit aussi de comprendre les processus d'orientation qui contribuent à dessiner ces parcours scolaires. Les processus de sélection internes au système éducatif représentent des indications de performance des individus en son sein en même temps qu'ils renseignent sur le caractère plus ou moins contraint de ces orientations en cours de scolarité. Outre l'éclairage qu'elles apportent sur les processus d'orientation proprement dits, ces informations aident à capter une part de l'hétérogénéité non observée entre individus, et sont donc nécessaires pour faire des analyses économétriques.

Enfin, les conditions matérielles et économiques qui permettent - ou non - aux jeunes de suivre des études, constituent un enjeu sociétal et de politique publique de plus en plus aigu (cf. infra le module sur le logement en cours d'études, sollicité par la DGESCO, le SDES et la DHUP). L'enquête de 2020 auprès de la Génération 2017 développera donc un module sur ce thème, qui complètera le module sur les expériences de travail en cours d'études.

<u>En aval : le parcours professionnel</u>

Le calendrier d'activité mensuel représente le socle des enquêtes Génération depuis leur création. Il est à la fois la source de l'information longitudinale produite et la cheville ouvrière du recueil d'information sur les différentes situations vécues par la personne enquêtée. Pour cette nouvelle enquête, il connait quelques aménagements. D'une part, il est enrichi par l'ouverture de pop-up en cours de saisie, qui simplifieront la suite du déroulé du questionnaire. D'autre part, sa nouvelle

mouture entérine le fait que la description de toutes les situations de non-emploi vécues au cours des trois dernières années peut constituer un exercice difficile pour un certain nombre de jeunes. Aussi, il a été acté de supprimer un descriptif détaillé de ces situations révolues pour se restreindre à un petit nombre de questions ayant principalement pour fonction de s'assurer un classement correct des individus dans les différentes situations.

Concernant le détail des situations recueillies dans le calendrier, quelques évolutions sont à noter.

Dans le prolongement des recommandations du rapport Gazier sur la diversité des formes d'emploi, un effort est fait pour identifier les personnes travaillant avec un statut d'auto-entrepreneur. De même, des questions sur la multi-activité (au moment de l'enquête) sont introduites. Enfin, l'enquête Génération met en œuvre la PCS 2020 dans l'identification de la profession exercée.

Concernant les questions liées à la formation et l'obtention de nouveaux diplômes après la fin de la formation initiale, les diplômes préparés lors des séquences de reprises d'études à temps plein ou en alternance sont collectés.

Un module sur les intermédiaires du marché du travail est introduit. Il a pour vocation à synthétiser l'information auparavant recueillie de façon éparse dans les différentes situations de non-emploi et présente l'avantage d'être aussi posé désormais aux jeunes en emploi.

Informations sociodémographiques sur la personne

Un tableau des mobilités résidentielles permet de mieux suivre les mobilités géographiques de la personne enquêtée (adresses renseignées pour certaines étapes du parcours comme l'entrée en 6ème et le baccalauréat).

Tableau 1 – Évolutions du questionnaire « tronc commun » (hors extensions) de l'enquête Génération 2017 interrogée trois ans après leur sortie de formation initiale, par rapport à l'enquête Génération 2010 à 3 ans

	THÉMATIQUES MAINTENUES	NOUVELLES THÉMATIQUES
Parcours de formation initiale	Arrêt des études	Financement des études
	Diplôme de sortie, plus haut diplôme obtenu, diplôme(s) intermédiaire(s), baccalauréat	Impact de la crise sanitaire sur les parcours
	Orientation (après la troisième, après le bac selon le niveau d'études atteint)	
	Expérience(s) de travail en cours d'études	
Parcours professionnel	Calendrier d'activité mensuel (Description de tous les emplois)	Auto-entrepreneuriat Multi-activité (indépendant, auto-entrepreneuriat, intermittent du spectacle, etc.) *PCS 2020*
	Opinion sur l'emploi	Intermédiaires du marché du travail
Caractéristiques de l'individu	Origine sociale, pays de naissance des parents et langues parlées à la maison	
	Ressenti de discrimination à l'embauche	
	Perspectives professionnelles	
	Calendrier du mode de cohabitation	

- **Objectifs et contenus du questionnaire des extensions**

Les extensions d'échantillon (pour obtenir davantage de questionnaires) répondent aux demandes émanant d'acteurs publics (régions, ministères, acteurs du champ de la relation formation-emploi), qui souhaitent disposer d'indicateurs comparables à ceux de l'exploitation nationale pour des champs particuliers. Les extensions de questionnement visent, elles, à développer ou approfondir une thématique spécifique éclairant le processus d'insertion professionnelle pour tout ou partie des jeunes enquêtés.

Le tableau 2 présente les différentes extensions de l'enquête Génération 2017 interrogée en 2020.

Tableau 2 – Extensions de questionnement, de champ et d'échantillon de l'enquête Génération 2017

EXTENSION	THÉMATIQUE(S) DE L'EXTENSION	COMMANDITAIRE(S)
Extension de questionnement (questions supplémentaires)		
Logement	Conditions de logement au cours de la dernière année d'études : Distances logement / établissement de formation, logement / entreprise (de stage ou d'alternance), difficultés dans la poursuite d'études liées au logement	DGESCO DHUP (Direction de l'habitat, de l'urbanisme et des paysages) SDES-CGDD Injep
Séjour à l'étranger en cours d'études	Séjours à l'étranger effectués à l'étranger en cours d'études et description plus détaillée des conditions du séjour le plus long	Erasmus + DGESIP Injep OFQJ (L'Office franco-québécois pour la Jeunesse)
Service civique	Identification des jeunes ayant effectué un service civique et datation dans le parcours	Agence du service civique
Risques physiques et chimiques	Formation : sensibilisation ? Emploi : prévention par l'employeur ? Équipement spécifique ?	DARES
Formations environnementales	Formation et emploi en lien avec l'environnement ? (perception de l'enquêté)	CGDD

Attractivité de la Fonction publique	Perception de la fonction publique : - pour ceux qui n'y ont jamais travaillé - pour ceux qui y travaillent ou y ont travaillé transitoirement Lien (parents) avec la Fonction publique	DGAFP
Extension d'échantillon (nombre de répondants supplémentaires sur certains champs)		
Formations de l'enseignement supérieur	Insertion des jeunes sortant de l'enseignement supérieur (tous diplômes)	DGESIP
Formations environnementales	Insertion des jeunes sortant de formations environnementales	CGDD
Formations automobiles	Insertion des jeunes de la filière (CAP, Bac pro, BTS)	ANFA
Jeunes de QPV	Insertion des jeunes qui résidaient dans un QPV à la fin de leurs études	CGET devenu ANCT
Sport	Insertion des jeunes sortant des formations des secteurs du sport et de l'animation. L'extension comprend les post-initiaux, habituellement hors-champ Céreq. De plus, le module multi-activité est davantage développé pour ce public (activité en lien avec le secteur ?).	Ministère des Sports Injep
Bretagne	Insertion des jeunes ayant terminé leurs études en Bretagne	Conseil régional de Bretagne
Pays de la Loire	Insertion des jeunes ayant terminé leurs études en Pays de la Loire	Conseil régional des Pays de la Loire

Trois extensions (Bretagne, Pays de la Loire, jeunes des quartiers prioritaires de la politique de la ville (QPV)) correspondent à des extensions définies territorialement (basées sur l'adresse de l'établissement de formation pour les deux régions et sur l'adresse de la personne enquêtée pour la troisième). Elles constituent des extensions d'échantillon qui ont pour but de permettre la diffusion d'information sur ces zonages. L'extension financée par le CGET (devenu ANCT en 2019) se singularise pour deux raisons. D'une part, le CGET-ANCT est intéressé à recueillir l'adresse des personnes enquêtées au moment de leur bac (pour celles arrivées au moins en terminale). Ce type d'information permettra d'analyser, en plus des questions d'insertion (extrêmement difficiles pour ces jeunes), les parcours dans l'enseignement supérieur. D'autre part, le CGET a financé une extension portant sur les deux interrogations de la Génération 2017 (2020 et 2023).

Quatre autres extensions visent des filières de formation spécifiques (CGDD, ANFA, DGESIP, INJEP/Ministère des Sports). Ces extensions d'échantillon doivent permettre de produire de l'information statistiquement robuste sur les segments visés du système de formation. Deux de ces extensions s'accompagnent d'une demande de questionnement limité. Le CGDD souhaite savoir si la personne enquêtée estime, de son point de vue, que sa formation, le métier qu'il exerce et l'activité de l'entreprise qui l'emploie a bien une dimension environnementale. Le ministère des Sports et l'Injep souhaitent, de leur côté, identifier, en cas de multi-activité déclarée, si le deuxième emploi décrit est dans le domaine du sport ou de l'animation.

Dans le même esprit, la DARES a sollicité une extension de questionnement auprès des jeunes sortis de formations professionnelles dans le but d'identifier quel est le degré d'information dont ils disposent (qu'ils ont retenus) sur les risques physiques et chimiques dans le travail. Le questionnement sur la prévention de ces risques concerne donc à la fois la formation suivie (ont-ils bénéficié d'un module de sensibilisation ?) et l'entreprise qui les emploie (ont-ils bénéficié de consignes, de rappels, de mise en garde, concernant les risques physiques et chimiques ? A-t-on mis à leur disposition un équipement spécifique ? A quels risques sont-ils exposés ?)

La DGESCO, la DHUP, l'Injep et le SDES-CGDD sont demandeurs, de leur côté, d'un module de questionnement sur le logement, au cours de la dernière année d'études. La préoccupation qui motive cette extension concerne la distance qui peut exister entre le domicile familial et l'établissement de formation professionnelle que fréquente (ou aurait souhaité fréquenter) l'individu. Cette distance peut constituer, pour les ménages aux revenus limités, une barrière à l'accès à la formation souhaitée. Dans cette perspective, le module s'attache à saisir les conditions de logement de la personne au cours de sa dernière année d'études. Il distingue la possibilité que la personne doive se partager entre deux logements, l'un lui permettant de suivre sa formation dans l'établissement scolaire, l'autre de réaliser la période en entreprise (stage, alternance) dictée par sa formation le cas échéant. Le module cherche à identifier les potentielles difficultés induites par cet éloignement, voire les impossibilités en termes d'orientation qu'il génère.

Concernant encore le parcours scolaire, le module sur les séjours à l'étranger en cours d'études est de nouveau reconduit. Le module a été adapté pour que l'on puisse connaitre la durée totale passée à l'étranger pour chaque jeune et répondre aux deux objectifs européens concernant les mobilités en cours d'études (Benchmark). Le module vise à pouvoir mesurer l'incidence des séjours à l'étranger tant dans l'enseignement supérieur que dans l'enseignement secondaire. Ce module détaille les conditions du séjour le plus long effectué dans la période la plus récente (primaire, secondaire, supérieur). Si ce séjour est effectué dans le cadre scolaire, celui-ci est décrit plus précisément : type de validation qu'il permet, diplômé préparé, cadre dans lequel il a été financé, etc.

L'agence du service civique souhaite, pour sa part, pouvoir identifier les jeunes passés dans le dispositif qu'elle organise. D'une durée de six à douze mois, ce dispositif propose aux jeunes le souhaitant un engagement dans une activité.

Le module sur l'attractivité de la Fonction publique (module de questionnement uniquement) a pour objectif d'éclairer la DGAFP sur la perception qu'ont les jeunes de ce secteur, leurs intentions et/ou leurs motivations. Le module distingue donc ceux qui n'y ont jamais travaillé de ceux qui y travaillent ou y ont travaillé transitoirement. En complément, deux questions concerneront l'existence d'un lien dans la famille (parents) avec la Fonction publique.

- **Les points forts du dispositif Génération**

Un cadre d'analyse homogène et cohérent

Contrairement à d'autres enquêtes d'insertion qui visent des publics segmentés (apprentis, lycéens, sortants de grandes écoles ou d'université…), seul le dispositif « Génération » propose un questionnement, une méthodologie et un cadre d'analyse homogène pour tous, quels que soient le parcours scolaire, les diplômes obtenus, les domaines et voies de formation. Il est donc possible de comparer et d'évaluer l'impact de ces différentes caractéristiques sur les variations observées au cours des premières années de vie active : qui accède rapidement à un emploi ? Qui reste durablement au chômage ? À quel type d'emploi accède-t-on ? À quel niveau de rémunération ? Telles sont les questions auxquelles le dispositif permet de répondre. Plus généralement, il met en évidence les phénomènes de concurrence ou de complémentarité entre niveaux, domaines et voies de formation.

Des informations riches et diversifiées

Grâce à un questionnaire détaillé et un échantillon important, les enquêtes permettent, au-delà des caractéristiques du parcours scolaire et des diplômes obtenus, de prendre en compte d'autres critères. Le genre, l'origine sociale, l'origine nationale, le lieu de résidence, les mobilités géographiques, le statut familial, les réseaux sociaux, mais aussi la place et le rôle des dispositifs publics sont autant de dimensions que le dispositif Génération permet d'intégrer pour analyser les différences observées au cours des premières années de vie active.

Un recul temporel nécessaire

Certaines enquêtes d'insertion sont réalisées quelques mois seulement après la sortie du système scolaire. L'option retenue est alors de disposer d'indicateurs qui peuvent être mis rapidement à disposition des décideurs, des familles et des étudiants. Avec le dispositif « Génération », la première interrogation est réalisée trois ans après la sortie du système scolaire. Les résultats des premières enquêtes ont mis en évidence l'importance de ce recul temporel. En effet, il faut attendre plusieurs années pour que la stabilisation professionnelle soit établie pour le plus grand nombre. Enquêter tôt après la sortie de formation donne une photographie faussée des situations par rapport à l'emploi, qui accentue fortement les différences, alors que les enquêtes « Génération » montrent que celles-ci tendent à se réduire avec le temps.

Un suivi longitudinal

Le questionnaire permet aux jeunes débutants de décrire systématiquement, mois par mois, les différentes situations qu'ils ont connues depuis leur sortie du système éducatif. Ce mode d'interrogation permet de construire différents indicateurs comme le taux de chômage ou le taux d'emploi, et d'aborder la qualité de l'emploi (niveau de rémunération, type de contrat). Il permet aussi de construire des typologies de parcours à partir de la description des situations mois par mois. Ces typologies offrent une vision synthétique des premières années sur le marché du travail : trajectoire d'accès rapide à l'emploi, trajectoire d'accès différé à l'emploi, trajectoire de décrochage, etc. L'insertion est une réalité multidimensionnelle qui ne peut se réduire à un ou deux indicateurs.

La même conjoncture pour tous

Les « générations » sont construites en fonction de la date de sortie de formation et non de l'année de naissance. Quel que soit leur niveau de formation, les jeunes arrivent donc dans un contexte de marché du travail plus ou moins favorable, mais identique pour tous. Il est donc plus facile a priori de comparer les trajectoires d'accès à l'emploi. Mais cette conjoncture a-t-elle les mêmes effets pour tous : à qui profitent les embellies ? Qui souffre le plus des retournements ? Quels effets sur les taux de chômage, l'importance des CDD ou de l'intérim, et pour qui ? Telles sont les questions auxquelles le caractère récurrent des enquêtes « Génération » permet de répondre.

Une enquête panélisée

L'enquête « Génération », dans sa forme rénovée, comprend une ré-interrogation en 2023. La première interrogation, trois ans après la sortie du système éducatif, est principalement tournée vers la caractérisation du processus d'insertion ; la ré-interrogation, à six ans, sera plutôt centrée sur les usages analytiques (notamment sur la question des parcours et des mobilités sur moyen terme) ; cette ré-interrogation permettra aussi d'approfondir les premières analyses issues de l'exploitation de la première interrogation.

1.2 – Pilotage de l'enquête et calendrier

- Conception de l'enquête

La conception de l'enquête, en termes d'élaboration du questionnaire, est prise en charge par le DEEVA (Département Entrées et Évolutions dans la Vie Active) au Céreq.

Les réflexions sur les évolutions à apporter d'une enquête à l'autre (ajout ou suppression de modules, ajout ou suppression de questions ou de modalités) sont menées de façon collégiale au sein du département, en croisant l'approche technique de l'équipe dédiée à la production, et l'expérience des chargés d'études dans l'exploitation des précédentes enquêtes.

Pour le cas particulier des extensions thématiques, une co-construction s'opère avec les commanditaires. Les partenaires sont en effet porteurs d'une problématique qu'ils souhaitent pouvoir étudier, le Deeva les accompagne dans la formalisation de l'écriture du module dédié.

- Pilotage de l'enquête

Au sein du Deeva, l'équipe ingénierie et gestion d'enquêtes assure le pilotage de l'enquête et en particulier a coordonné le travail réalisé par le prestataire (Médiamétrie) chargé du développement du questionnaire.

- Calendrier de l'enquête

Tableau 3. Calendrier prévisionnel de l'enquête

Étape	Date
Collecte des fichiers pour la constitution de la base de sondage	Octobre 2018 – Juin 2019
Lancement de l'AO pour le développement et l'administration du questionnaire	Janvier 2019
Réunion de lancement avec le prestataire chargé du développement du questionnaire	22 mars 2019
Présentation au Cnis pour avis d'opportunité	12 avril 2019
Pilote 1 (1er test en réel du questionnaire)	16 -28 septembre 2019
Remise du dossier au Comité du label	12 Novembre 2019
Passage au comité du secret	Décembre 2019
Passage au Comité du label	18 Décembre 2019
Pilote 2	Fin janvier 2020
Pilote 3	Fin février – début mars 2020
Démarrage de la collecte	2 Avril 2020
Fin de la collecte	31 juillet 2020
Premières publications	1er semestre 2021

2. Bilan d'exécution de l'enquête et des résultats produits

- **Le protocole de l'enquête**

L'enquête Génération 2017 à 3 ans a été réalisée en multimode, à partir d'un échantillon de 303 500 individus. Cet échantillon comportait une partie « échantillon Céreq », et une partie répondant aux besoins des partenaires d'extensions.

La collecte initialement prévue au printemps 2020 a été décalée à l'automne 2020, du fait du confinement lié à la crise sanitaire du Covid. Par ailleurs, la collecte a dû être prolongée de près de 3 mois et mobiliser un échantillon de réserve, en raison des difficultés sur le plateau d'enquête (turn-over des enquêteurs et distanciation sociale en lien avec le Covid). L'enquête s'est finalement déroulée entre septembre 2020 et mars 2021, selon un protocole multimode séquentiel et concurrentiel en cinq phases :

- **Phase 1 : Priorité INTERNET** (septembre 2020) : envoi des mails-avis et courriers, appel des enquêtés sans adresse électronique.

- **Phase 2 : Choix INTERNET-TELEPHONE** (octobre 2020) : appel de l'ensemble des non-répondants, y compris ceux ayant démarré leur questionnaire par internet. Possibilité de poursuivre par internet ou par téléphone.

- **Phase 3 : Priorité TELEPHONE** (novembre à mi-décembre 2020) : appel de l'ensemble des non-répondants, y compris ceux ayant démarré leur questionnaire par internet. Incitation à terminer l'enquête par téléphone.

- **Phase 4 : Relance sur internet uniquement** (mi-décembre 2020 au 4 janvier 2021) ; arrêt momentané des appels, relance par mail pour terminer le questionnaire en ligne.

- **Phase 5 : Poursuite des relances auprès des populations à faible taux de réponse, priorité TELEPHONE** (4 janvier au 22 mars 2021) : poursuite des appels pour les strates à faible taux de réponse ; intégration d'un échantillon de réserve.

- **Le taux de collecte**

Le taux de collecte de l'enquête Génération 2017 à 3 ans s'établit à 23%, dont 9% dans le champ et 14% hors champ.

$$taux\ de\ collecte = \frac{Nombre\ d'enquêtes\ réalisées\ (Répondants\ champ + Hors\ Champ)}{Nombre\ d'individus\ dans\ l'échantillon} = \mathbf{23\ \%}$$

Ce taux de collecte est plus bas que d'habitude. Il prend en compte l'échantillon de réserve, constitué de sous-populations plus difficiles à joindre (jeunes peu ou non diplômés, originaires de QPV), et n'ayant pas bénéficié de la même durée d'exploitation que l'échantillon principal.

Par ailleurs, cette enquête a pâti d'un fort taux d'abandon en cours de remplissage (voir Tableau 5). En effet, parmi les 42 171 (25 164 + 17 007) individus s'étant qualifiés dans le champ de l'enquête (champ Céreq), seuls 25 257 (25 164 + 93) individus ont terminé leur questionnaire : 40 % des individus qualifiés dans le champ Céreq ont donc abandonné le remplissage. Ce taux d'abandon est inédit. Lors des enquêtes Génération précédentes, menées en monomode, ce taux était de 11 %, c'est-à-dire que 89% des individus ayant passé le questionnaire filtre, et appartenant au champ Céreq terminaient leur questionnaire. Avec le passage au multimode, une hausse du taux d'abandon était anticipée mais pas d'une pareille ampleur. Ce taux d'abandon nous a invité à questionner notre protocole, mais il est aussi lié en partie aux difficultés de la collecte, réalisée en pleine crise sanitaire.

Tableau 4. Classement des individus échantillonnées

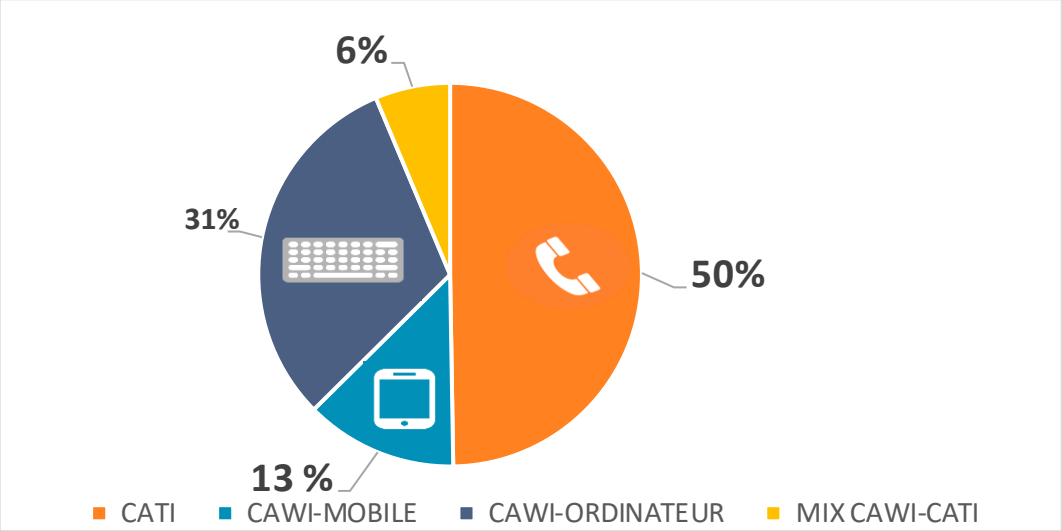
Classement des individus	Effectif	%
Répondants dans le champ (Céreq et /ou Sport)	**26 464**	**9**
Champ Céreq	25 164	9
Champ post-initiaux sport	1 207	<1
Inexploitables	93	<1
Hors champ	**43 026**	**14**
Répondants hors du champ Céreq	42 158	14
Décès / Incapacité	868	<1
Incomplets	**22 098**	**7**
Incomplets champ Céreq	17 007	5
Incomplets champ indéterminé	5 091	2
Non répondants avec contacts	**77 115**	**25**
Refus explicite de l'individu	11 076	3
Contact avec l'individu recherché CATI ou CAWI	20 243	7
Contact avec un tiers	45 796	15
Non répondants sans contact	**134 870**	**44**
Présence de coordonnées (validité indéterminée)	39 728	13
Coordonnées invalides	43 446	14
Aucune coordonnée	51 696	17
Total	**303 573**	**100**

- **Le mode de remplissage du questionnaire**

La moitié des questionnaires ont été remplis exclusivement par téléphone, 44% exclusivement par internet et 6% ont été remplis sur les deux modes (voir Figure 1). Parmi les répondants par internet, la plupart ont utilisé un ordinateur pour le remplissage du questionnaire. Une part non négligeable (environ un répondant par internet sur trois) a été complétée sur mobile : des efforts ont été faits pour adapter le questionnaire à un remplissage sur smartphone. Néanmoins, du fait de la longueur du questionnaire et de la spécificité du calendrier d'activité (dont l'ouverture a généré un fort taux d'abandons sur smartphone), la participation via ce support de collecte reste limitée.

La possibilité de remplir le questionnaire sur plusieurs modes et supports offre une certaine souplesse permettant de s'adapter aux préférences et contraintes de chacun.

Figure 1. Mode de collecte du questionnaire d'enquête des répondants dans le champ (%)



- **Résultats produits à partir de l'enquête Génération 2017 à 3 ans**

L'enquête de 2020 auprès de la Génération 2017 a permis de produire de nombreux résultats et alimenté les publications du Céreq.

Ces publications ont débuté par des études de premiers résultats pointant l'amélioration des conditions d'insertion des jeunes de cette Génération vis-à-vis de ses aînées, en premier lieu du fait du contexte économique présidant à leur arrivée sur le marché du travail, au moins jusqu'à l'émergence de la crise de la Covid. Les différentes études ont donné lieu à des exploitations centrées sur le devenir des jeunes selon leur niveau de sortie ou passés par une voie de formation par apprentissage. L'impact de la crise Covid sur les parcours des jeunes a donné lieu à des études approfondies, l'enquête permettant de documenter les situations des personnes pendant le premier confinement et les mois qui ont suivi, jusqu'en octobre 2020.

D'autres études se sont focalisées sur les parcours des jeunes. Ces études ont établi que les parcours en formation ont une importance dans le devenir ultérieur des jeunes sur le marché du travail, au-delà du seul plus haut diplôme atteint. Cette influence passe par la construction des cursus dans laquelle les initiatives, individuelles ou collectivement organisées, contribuent à davantage individualiser les parcours que par le passé. Ces initiatives passent entre autres par des mobilités en cours d'études, qu'elles soient temporaires (séjours à l'étranger) ou plus durables (mobilités d'études).

Au-delà de la formation, les transformations des parcours professionnels dans les premières années de vie active sont et ont aussi été étudiées. Elles ont mis en évidence la prégnance plus grande des retours en formation après quelques années passées sur le marché du travail, surtout le fait de jeunes sortis du système éducatif sur un échec scolaire. Cette montée des reprises d'études souligne la porosité croissante en France de la frontière entre système éducatif et système productif. Dans le même esprit, l'apprentissage, voie de formation particulière qui permet de poursuivre sa formation tout en se positionnant déjà dans le marché du travail, a continué à être étudié ; les travaux se sont particulièrement intéressés à son développement dans l'enseignement supérieur et ont souligné son impact différencié sur l'insertion selon le niveau du diplôme de sortie.

Finalement, différentes études ont pointé le rôle des inégalités sociales comme facteur d'hétérogénéité pesant sur la valorisation du diplôme au moment de l'insertion. Par exemple, plusieurs études soulignent la remarquable progression des conditions d'insertion des jeunes femmes ; cependant, une fois remis en contexte leur progression dans la scolarité plus remarquable encore, elles concluent à la persistance d'inégalités sur le marché du travail, se traduisant notamment par des différences d'accès à la catégorie cadres entre jeunes hommes et femmes diplômés de l'enseignement supérieur.

Différentes études méthodologiques ont également été menées, à partir de l'enquête Génération 2017 à 3 ans. En effet, cette enquête a vu la mise en place et la réalisation d'une collecte multimode dans le cadre du nouveau dispositif Génération. Ce passage au multimode constitue une adaptation nécessaire du dispositif, à la fois aux évolutions de comportement et de pratiques des jeunes enquêtés et au cadre plus structurel visant à un contrôle plus strict des coûts de l'enquête. Ces études ont permis de rendre compte de cette première expérience.

3. Méthodologie statistique

3.1 - Champ, unités enquêtées

- **Définition du champ de la Génération 2017**

La Génération 2017 concerne les individus « primo sortants » de formation initiale en 2016-2017 (année scolaire), qu'ils soient diplômés ou non. Dans « primo-sortants » sont inclus les individus ayant interrompu leurs études par le passé pendant une durée comprise entre 1 et 16 mois. Tous les niveaux et domaines de formations sont concernés.

De façon plus précise, les critères d'éligibilité pour être dans le champ retenu, nommé ensuite « champ Céreq », sont les suivants :

- Avoir été inscrit dans un établissement de formation en France métropolitaine ou dans un DROM durant l'année scolaire 2016-2017 ;

- Avoir quitté le système éducatif entre octobre 2016 et décembre 2017 ;

- Ne pas avoir interrompu ses études durant plus de 16 mois avant l'année scolaire 2016-2017 (sauf pour raison de santé) ;

- Ne pas avoir repris ses études au cours des 16 mois qui ont suivi l'entrée sur le marché du travail ;

- Avoir 35 ans ou moins en 2017.

Ces conditions sont cumulatives.

Quelques points particuliers concernent l'application de ces critères :

- Bien qu'il s'agisse de contrats de travail, les contrats d'apprentissage et les contrats de professionnalisation à visée diplômante et qui interviennent dans la continuité du parcours scolaire initial sont considérés comme relevant de la formation initiale. Une personne repérée comme sortant de formation en 2016-2017 qui poursuit par un contrat d'apprentissage ou un contrat de professionnalisation en 2017-2018 est donc considérée en poursuite d'études, et donc hors champ.

- Une personne sortie d'un établissement de formation en 2016-2017 qui poursuit des cours par correspondance ou des cours du soir en 2017-2018 est considérée comme en poursuite d'études, donc hors champ, si elle n'a pas d'emploi en parallèle. Si elle a un emploi en parallèle, sa situation d'emploi prime.

3.2 - Paramètres d'intérêt de l'enquête

L'enquête 2020 sur la Génération 2017 a pour objet principal de caractériser l'insertion professionnelle des jeunes, dans une perspective multicritère. Dans cette optique, elle doit permettre :

- De distinguer, une fois la personne sortie de formation initiale, **mois par mois, ses différentes positions sur le marché du travail**. Le premier paramètre d'intérêt distingue différents types de séquences : emploi, chômage, formation ou reprise d'étude, autre situation. Parmi les séquences collectées, celle correspondant au moment de l'interrogation (situation actuelle) revêt une importance particulière ;
- Parmi les séquences d'emploi, d'identifier pour les salariés **le contrat de travail** prévalant à l'embauche, à la fin du de la séquence, et le moment où s'est opéré ce changement. Ces informations permettent ainsi de distinguer les personnes employées en CDD (sous ses différentes formes) de celles employés en CDI ou bénéficiant du statut de fonctionnaire, et du moment du passage de l'un à l'autre.

Ces deux premiers points aboutissent à reconstituer des calendriers professionnels couvrant les trois premières années selon les positions d'activité successives des personnes. Ces calendriers permettent, à leur tour, de :

- Calculer les **durées passées** dans les différents états
- Calculer le **nombre de séquences** par état.

Ces deux types de paramètres permettent une première caractérisation des cheminements professionnels des jeunes (par exemple, ils peuvent permettre de distinguer les situations de chômage récurent des situations de chômage persistant). Ils s'avèrent importants pour compléter l'information sur la seule situation observée en coupe, au moment de l'enquête, trois ans après la sortie et permettent de distinguer les rythmes d'insertion, au-delà du mouvement de convergence vers l'emploi qui caractérisent une majorité des trajectoires, particulièrement parmi les diplômés de l'enseignement supérieur. Le calendrier permet également de :

- Calculer la **durée mise pour atteindre un état donné**. Typiquement, la durée d'accès au premier emploi est un paramètre utilisé (particulièrement parmi les peu diplômés) ; la durée d'accès au premier CDI est également un paramètre largement mobilisé.
- Réaliser une **analyse typologique** des calendriers professionnels, aboutissant à identifier des trajectoires-type intégrant les différentes dimensions. Classiquement, cette analyse a identifié sur les Générations précédentes une trajectoire de référence de stabilisation rapide et durable à l'emploi / en EDI.

L'enchainement des séquences, la nature des transitions constituent également des indicateurs clés. Parmi les séquences autres que la situation à la date d'enquête, le **premier emploi** représente une séquence-clé des analyses puisqu'il permet de caractériser les conditions initiales d'accès à l'emploi. Différents travaux ont pointé l'ambivalence des CDD, entre opportunité (tremplin vers le CDI) et risque (trappe à précarité).

Cependant, pour les jeunes entrant sur le marché du travail, l'accès à l'emploi, ou non, et la nature du contrat de travail ne suffisent pas à caractériser les conditions d'insertion des différentes sous-populations étudiées. En effet, ces paramètres permettent d'opposer les grands niveaux de diplômes (non diplômés / diplômés du secondaire / diplômés du supérieur) mais ils sont beaucoup moins informatifs concernant des distinctions plus fines, particulièrement à la sortie de l'enseignement supérieur. Ils nécessitent donc d'être complétés par d'autres paramètres d'intérêt :

- La **rémunération perçue**,
- Le **niveau de qualification des emplois**

Ces paramètres permettent de développer des analyses en termes de rendement des diplômes mais également des approches en termes de déclassement (normatif, statistique ou salarial). Cet aspect permet du coup de souligner la place dans l'enquête de l'information sur **l'acquisition de diplôme** *après* la sortie de formation initiale.

Au-delà de la qualité de l'emploi occupé, l'enquête s'intéresse également à la nature de l'activité exercée et son éventuel lien avec la formation suivie. La **profession occupée** (PCS 2020 pour l'enquête de 2020) constitue donc un autre paramètre d'importance, au-delà du seul niveau de qualification du poste.

Outre ces informations factuelles caractérisant les situations professionnelles des personnes, un apport important de l'enquête est de pouvoir rendre compte et **caractériser le niveau de satisfaction / insatisfaction des personnes par rapport à leur activité de travail** ; sentiment d'être utilisé en dessous ou à son niveau de compétence, sentiment d'être (plutôt) bien ou mal rémunéré, sentiment de se réaliser professionnellement, etc., de pouvoir caractériser son rapport au travail exercé. Dans le même esprit, l'enquête permet de caractériser la dynamique temporelle des parcours, des aspirations et des anticipations des personnes, par des questions sur leurs priorités professionnelles, leurs aspirations à la mobilité (personnes en emploi déclarant rechercher un autre emploi), le sentiment d'insécurité qu'ils sont susceptibles d'exprimer[1] (« être plutôt inquiet / optimiste pour son avenir professionnel »), le ressenti de discrimination à leur encontre, etc. Ces approches prennent davantage de relief encore dans la mesure où la Génération 2017 est un dispositif panélisé permettant donc de

[1] cf. Séance du 5 novembre 2019 de la Commission Emploi, qualification et revenus du travail du CNIS sur la mesure du sentiment d'insécurité sur le marché du travail)

suivre les évolutions professionnelles y compris dans leur dimension subjective à travers ce rapport au travail exprimé.

Enfin, au-delà de la seule caractérisation de la « vie professionnelle » des personnes enquêtées, l'enquête constitue une ressource pour étudier les conditions de vie des jeunes. Ainsi, le **calendrier habitat** pointe les difficultés qu'éprouvent une partie de ces jeunes à accéder à l'autonomie résidentielle, quitter sa famille d'origine pour éventuellement en fonder une nouvelle. Dans le dispositif Génération, ces problématiques sont habituellement amorcées au cours de la première interrogation et développées dans l'interrogations suivante. Plus largement, les variables sociodémographiques permettant de caractériser sociologiquement les sous-populations étudiées constituent des variables importantes de l'enquête.

3.3 - Description du sondage

Etape 1 : collecte de fichiers d'élèves et création de la base de sondage

Il n'existe pas de base nationale d'élèves nominative couvrant l'ensemble des sortants du système éducatif une année donnée. Le Céreq constitue donc cette base de sondage pour chaque Génération, à partir de différentes sources et en opérant divers retraitements.

Pour cela, deux principales opérations de collecte ont lieu pour récupérer des listes nominatives d'élèves ou d'étudiants inscrits dans les établissements français. Dans un premier temps, une collecte de fichiers de données nominatifs centralisés au niveau national sur un champ spécifique est réalisée. Les fournisseurs de ces données sont : le ministère en charge de l'Education, le ministère en charge de la Culture, le ministère de l'Agriculture, la Direction des Sports sur les diplômés jeunesse et sport, la direction de l'Animation de la recherche, des Études et des Statistiques (DARES), l'Association nationale de la recherche et de la technologie (ANRT). Dans un second temps, une collecte complémentaire auprès de tous les autres établissements de formation (universités, écoles d'ingénieurs, diplômés d'écoles de la santé et du social, etc.) est réalisée spécifiquement pour le Céreq avec l'aide d'un sous-traitant externe pour le contact des différents établissements.

Une fois cette collecte réalisée, un travail d'appariement de fichiers est effectué pour éliminer les élèves ayant poursuivi leurs études au cours de l'année suivant celle concernant la Génération interrogée. Par exemple, pour la Génération 2017, une comparaison a été effectuée entre les fichiers d'inscrits dans une formation en 2016-2017 et les fichiers d'inscrits dans une formation en 2017-2018. Une base de sondage d'environ 1 130 000 individus présumés sortants de formation initiale en 2016-2017 a ainsi été constituée.

Etape 2 : tirage de l'échantillon

L'enquête Génération est ouverte à des partenariats avec des acteurs intervenant dans le domaine de l'emploi et de la formation, pour répondre à des besoins de connaissance spécifiques. Elle offre notamment la possibilité d'extensions d'échantillons, pour disposer d'un nombre conséquent de répondants sur un champ spécifique (une zone géographique ou un type de formation).

Pour chaque Génération, l'échantillon est construit de manière à obtenir un échantillon de répondants qui soit représentatif de l'ensemble des sortants de formation initiale (pour les besoins du Céreq) mais aussi de satisfaire les demandes en nombre de questionnaires des partenaires d'extensions. Il répond également au besoin d'assurer un nombre suffisant de répondants pour la ré-interrogation à 6 ans, en anticipant la non-réponse et l'attrition.

Pour l'enquête Génération 2017 à 3 ans, l'échantillon a été constitué par tirages indépendants : l'échantillon « Céreq » tronc commun ainsi qu'un échantillon de réserve (pour les besoins du Céreq) et des échantillons d'extensions. Ces échantillons ont été tirés indépendamment par tirage stratifié à probabilité inégale, puis combinés en post-collecte par un partage des poids.

3.4 - Traitements statistiques

- **Estimation composite et partage des poids**

Chacun des échantillons tirés pour l'enquête Génération, tronc commun ou extension, répond à un objectif propre et permet d'obtenir un jeu de poids $\left\{ w_{i,j}^{ech} = \frac{1}{\pi_{i,j}} I[i \in S_j] \right\}_{i \in U}$ permettant d'estimer sans biais leur domaine d'intérêt (tronc commun si $j = 0$, extension E_j pour $j \in [\![1,7]\!]$) à partir des individus de l'échantillon. Ces différents domaines se recoupent fortement ; en particulier, le tronc commun inclut les domaines d'intérêt de toutes les extensions. Il semblerait dommageable de ne pas inclure tous les individus ayant répondu à l'enquête lors d'estimations sur l'ensemble du champ de l'enquête, et de ne garder que les individus tirés pour le tronc commun. Afin de tirer parti du maximum d'informations disponibles, une estimation composite est réalisée à partir des différents jeux de poids issus des échantillons tirés pour le tronc commun et les extensions. Ainsi, pour estimer le total d'une variable y un sous-domaine D du champ de l'enquête Génération, il s'agit d'agréger les estimations provenant des estimations de ce total de chacun des échantillons, en tenant compte du fait que ces échantillons ne représentent pas tous intégralement D. Par exemple, l'échantillon extension Bretagne ne permet pas d'estimer correctement les agrégats sur la population entière, mais peut améliorer l'estimation par ce qu'elle apporte sur le champ des sortants de la région Bretagne ; elle ne permet cependant pas d'améliorer les estimations sur les individus sortant d'autres régions.

Pour agréger les différentes estimations, il faut préalablement segmenter le champ de l'enquête selon les sous-domaines que les échantillons permettent ou non de représenter, en considérant la partition décrite par l'ensemble des intersections des différents champs des extensions.

Tableau 5 - Partition de la base de sondage de l'enquête auprès de la Génération 2010 selon les intersections d'extensions prévues pour l'interrogation 2020 de la Génération 2017

Partie	E1	E2	E3	E4	E5	E6	E7	Effectif
1	0	0	0	0	0	0	0	478730
2	0	0	0	0	0	0	1	35032
3	0	0	0	0	0	1	0	29017
4	0	0	0	0	1	0	0	10512
5	0	0	0	0	1	0	1	847
6	0	0	0	0	1	1	0	867
7	0	0	0	1	0	0	0	4631
8	0	0	0	1	0	0	1	189
9	0	0	0	1	0	1	0	322
10	0	0	1	0	0	0	0	10031
11	0	0	1	0	0	0	1	895
12	0	1	0	0	0	0	0	429351
13	0	1	0	0	0	0	1	29099
14	0	1	0	0	0	1	0	20576
15	0	1	0	0	1	0	0	13344
16	0	1	0	0	1	0	1	773
17	0	1	0	0	1	1	0	1092
18	0	1	0	1	0	0	0	515
19	0	1	0	1	0	0	1	39
20	0	1	0	1	0	1	0	20
21	1	0	0	0	0	0	0	41316
22	1	0	0	0	0	0	1	1813
23	1	0	0	0	0	1	0	1031
24	1	0	0	0	1	0	0	510
25	1	0	0	0	1	0	1	25
26	1	0	0	0	1	1	0	18
27	1	0	0	1	0	0	0	690
28	1	0	0	1	0	0	1	17
29	1	0	0	1	0	1	0	20
30	1	0	1	0	0	0	0	487
31	1	1	0	0	0	0	0	29187
32	1	1	0	0	0	0	1	1084
33	1	1	0	0	0	1	0	702
34	1	1	0	0	1	0	0	490
35	1	1	0	0	1	0	1	24
36	1	1	0	0	1	1	0	19
37	1	1	0	1	0	0	0	37
38	1	1	0	1	0	0	1	3
39	1	1	0	1	0	1	0	1

Note de lecture *La partie n°6 est composée des individus appartenant aux sous-populations E5 et E6 sans appartenir à aucune autre sous-population. Dans la base de sondage de l'enquête auprès de la Génération 2010, ils étaient 867 individus au sein de la partie 6.*

Pour chaque élément p de cette partition, il s'agit alors d'estimer sans biais le total de y sur p à partir des différents échantillons et de pondérer ces estimations à l'aide d'un jeu de coefficients $a_{p,j}$. L'estimateur composite du total de Y sur p est alors :

$$\hat{Y}_{comp,p} = \sum_{j \in [\![0,7]\!]} \left(a_{p,j} \cdot \sum_{i \in S_j} w_{i,j}^{ech} \cdot y_i \right).$$

$$Sous\ condition : \sum_{j \in [\![0,7]\!]} a_{p,j} = 1$$

L'estimateur du total de Y sur l'ensemble de la population est alors :

$$\hat{Y}_{tot} = \sum_{p \in P} \hat{Y}_{comp,p}$$

Au niveau individuel, cela se traduit par un partage des poids. Pour chaque individu i d'un sous-domaine p, son poids de l'estimateur composite est alors

$$w_i^{comp} = \sum_{j \in [\![0,7]\!]} a_{p,j} \cdot w_{i,j}^{ech}$$

Pour chaque sous-domaine p de la partition P, le choix des coefficients $a_{p,j}$ dépend des extensions constituant p. Si p contient l'extension E_3 ou E_4 (qui sont exhaustives), alors $a_{p,3} = 1$ ou $a_{p,4} = 1$ et les autres coefficients sont nuls ; E_3 ou E_4 garantit que tous les individus de p sont au sein de l'échantillon.

Sinon, $a_{p,j}$ est égal à la part d'individus de p répondants provenant de l'extension E_j, de telle sorte que l'estimation du total à partir d'un échantillon (tronc commun ou extension) fournissant un grand nombre de répondants de p ait plus de poids que celle d'un échantillon en fournissant moins. Le poids de l'estimation composite d'un individu sera alors la moyenne pondérée par le nombre de répondants des poids d'échantillonnage de chaque échantillon (avec un poids nul si l'individu n'est pas échantillonné). De cette manière, les moyennes des $a_{p,j} \cdot w_{i,j}^{ech}$ par échantillon E_j seront a priori proches et les poids moyens des individus échantillonnés une unique fois seront en moyenne proches, indépendamment de l'échantillon initial dans lequel ils étaient tirés.

- **Traitement de la non-réponse totale**

Une correction de la non-réponse totale est faite par repondération par l'inverse de la probabilité de réponse estimée $p_{red,i}$. L'estimation est réalisée par la méthode des groupes homogènes de réponse, suite à une estimation par régression logistique, de manière analogue aux enquêtes précédentes.

Lors de la précédente enquête, la modélisation était décomposée en deux étapes, afin d'affiner l'estimation en intégrant des données relatives au processus de collecte. Une première étape était d'estimer la probabilité de contacter les individus échantillonnés ou un de leur proches à l'aide des variables présentes dans la base de sondage ainsi que du processus de contact.

La deuxième étape modélisait la probabilité d'accepter de répondre à l'enquête sachant que l'individu a été contacté et est identifié dans le champ de l'enquête. Dissocier cette modélisation de la première

permet notamment d'intégrer des variables relatives à la phase de contact du questionnaire : nature du contact, prise de rendez-vous…

Pour exemple, sur la première enquête auprès de la Génération 2013, les variables utilisées dans les modèles étaient :

- Niveau et type de formation
- Âge de l'individu (en 2013)
- Variables indicatrices sur le type de coordonnées téléphoniques disponibles dans l'échantillon (numéro de téléphone fixe, portable, issu de la base de sondage ou de recherches de coordonnées téléphoniques,…)
- Mode d'envoi de la lettre avis
- Formation effectuée par apprentissage en 2013
- Caractéristiques de la commune de résidence en 2013
- Département de l'établissement de formation en 2013
- Genre

Pour l'enquête 2020 auprès de la Génération 2017, pour inclure la dimension multimode de cette enquête, des variables renseignant le type de contact utilisé, la phase de l'enquête où le premier contact a été effectué, ou les différents types de relances utilisées pour contacter l'individu ont été utilisées dans la modélisation en addition aux variables mentionnées ci-dessus.

- **Calage sur marges et pondération finale**

Les Repères et références statistiques (RERS) du ministère de l'Éducation nationale publient un tableau donnant une estimation de la répartition par sexe et niveau de diplôme des sortants du système éducatif français. Les poids des individus répondant à l'enquête 2020 auprès de la Génération 2017 seront calés sur la répartition publiée dans le RERS.

Les poids finaux des individus répondants de l'enquête Génération sont alors

$$w_i^{tot} = \frac{w_i^{comp}}{p_{red,i}} . cal_i$$

Où cal_i est le coefficient de calage.

- **Traitement de la non-réponse partielle : les salaires**

Durant le questionnement, pour chaque emploi déclaré par l'individu, le salaire au début et à la fin de la séquence d'emploi est demandé. L'information sur le salaire étant sensible, il est laissé l'opportunité aux individus ne souhaitant pas livrer le montant exact de leur salaire de ne pas répondre ou de donner un intervalle dans lequel il est compris (déclaration de tranches de salaires). Cette information est cependant un des principaux paramètres d'intérêt et objets d'études des enquêtes Génération ; les salaires manquants sont imputés afin d'avoir des données sur chacun des emplois renseignés.

Le traitement de la non-réponse des salaires s'effectue par imputation par la régression linéaire. Les salaires déclarés en début et fin de séquence emploi sont modélisés séparément. Les salaires dont les résidus studentisés de ces modèles sont supérieurs à 3 sont jugés aberrants et mis à blanc. Les informations non renseignées sont alors imputées de manière déterministe par la projection de ce modèle.

Une modélisation similaire est effectuée pour l'imputation des salaires déclarés en tranche, à l'aide des salaires déclarés au sein de l'intervalle de salaire correspondant à la tranche.

Les variables permettant la modélisation des salaires sont :

- Sexe
- Âge de l'enquêté
- Niveau de sortie : non diplômé, secondaire, bac+2, bac+3/4, bac+5
- Plus haut diplôme obtenu en 15 positions
- Spécialité de formation : général, industriel ou tertiaire
- Type de contrat de travail : indépendant, fonctionnaire, cdi, cdd, contrats aidés
- Catégorie socio-professionnelle : ouvrier, profession intermédiaire, cadre, employé, autre
- Ancienneté
- Région de l'entreprise : Île-de-France, autre région, étranger
- Taille de l'entreprise : moins de 20 salariés
- Activité de l'entreprise : NAF en 8 postes

- **Traitement du biais de mesure entre les différents modes**

La collecte multimode implique des effets de mesure dans les réponses des individus. De fait, un même individu ne répond pas forcément de la même manière selon le mode d'interrogation. Les causes de ces différences peuvent être multiples : en présence d'un enquêteur, un individu se sent plus concerné et se concentre plus pour trouver la réponse la plus adéquate à une question difficile et peut être aidé en cas d'incompréhension de la question. Lorsqu'au contraire, lors d'un mode de collecte auto-administré, l'enquêté se contente de donner une réponse qu'il juge bonne au lieu de chercher la meilleure réponse qu'il aurait pu donner ; cet effet est dénommé *satisficing*.

Entre les collectes Internet et téléphone, les informations ne sont pas reçues identiquement : dans le cadre de la collecte par téléphone, l'enquêté doit se rappeler de la liste des modalités, alors que dans la passation en auto administré, l'enquêté lit les modalités qui apparaissent toutes ensemble sur son écran. À partir de cette différence de transmission de l'information, deux effets de mesure opposés

ont été définis. Dans une collecte par téléphone, on parle de *recency effect* : l'enquêté donnera davantage une réponse parmi les modalités dont il se souvient, c'est-à-dire une des dernières modalités entendues. À l'inverse, dans le cas de la collecte par Internet, les répondants auraient tendance à choisir une des premières modalités lues. Il s'agit, alors d'un effet de primauté (*primacy effect*).

Dans le cas d'un mode de collecte où un enquêteur est présent, l'enquêté peut craindre le jugement de l'enquêteur et donner des réponses socialement plus acceptables à des questions d'opinion ou sensibles. Cet effet est appelé biais de *désirabilité sociale*.

Le biais impliqué par les effets de mesure sur les estimations est problématique : les estimations des variables sujettes à un effet de mesure réalisées sur les populations de répondants diffèrent selon le mode de réponse à structure de population égale. La vraie valeur du paramètre n'est sûrement pas égale à l'une des estimations, effectuée sur un seul mode ; l'estimation de ce paramètre sur l'ensemble de la population dépend de la répartition des individus selon le mode de réponse.

Afin de tenir compte de ce biais de mesure dans l'enquête auprès de la Génération 2017, une quantification de l'effet de mesure sur chacune des variables du questionnaire est réalisée par un matching sur score de propension. Usuellement utilisées dans le cadre de l'évaluation des politiques publiques, le but des méthodes de matching est d'évaluer l'effet d'un traitement T sur une variable d'intérêt Y. Dans le contexte de l'enquête auprès de la Génération 2017, le traitement est le fait de répondre au questionnaire par Internet, par rapport à une réponse par téléphone ; les variables d'intérêt seront des variables du questionnaire susceptibles de présenter un effet de mesure, par exemple des variables portant sur l'opinion sur l'emploi des individus, les raisons de son arrêt des études, ou sur le sentiment de discrimination au sein de l'entreprise.

La population suivant le traitement a une variable de traitement T=1 tandis que celle ne le subissant pas a pour valeur T=0. Dans ce cadre, deux variables latentes Y1 et Y0 correspondent aux valeurs de Y que l'individu aurait renseigné en répondant, respectivement, par Internet ou par téléphone. Une seule de ces variables latentes est observée, mais la relation suivante existe :

$$Y = T.Y1 + (1 - T).Y0$$

L'hypothèse de cette approche par matching est de considérer que, conditionnellement aux variables qui expliquent la sélection, les variables latentes sont indépendantes du traitement. C'est-à-dire que conditionnellement aux covariables, il n'y a plus d'effet de sélection et l'assignation du traitement peut être considérée comme aléatoire.

L'objectif est d'alors d'estimer la valeur latente non observée. Pour cela, à chaque individu qui a reçu le traitement est attribué un contrefactuel qui n'a pas reçu le traitement et qui lui ressemble selon les variables qui expliquent la sélection. Dans le contexte de l'enquête auprès de la Génération 2017, la détermination d'un contrefactuel de chaque individu se fait par la méthode d'appariement sur score de propension ; dans le cas d'indépendance des variables latentes Y1 et Y0 par rapport aux covariables expliquant la sélection, alors il y a aussi indépendance par rapport à la probabilité de recevoir le traitement, le score de propension. Ce score de propension est estimé par modélisation logistique de la probabilité de répondre par Internet plutôt que par téléphone à partir de variables non sujettes à effet de mesure.

La valeur de la variable latente inobservée Y0 est alors estimée par la variable Y du contrefactuel. Réciproquement, la valeur de la variable latente inobservée Y0 des individus qui n'ont pas suivi le traitement peut être estimée par la valeur Y de leur contrefactuel qui a suivi le traitement.

L'effet de mesure est estimé par l'effet moyen du traitement (Average Treatment effect on the Treated ou ATT) :

$$ATT = \mathbb{E}(Y1 - Y0|X, T = 1)$$

Les variables sujettes à un fort effet de mesure sont en nombre limité, et identifiées dans le dictionnaire des variables de l'enquête afin que chaque personne utilisant les données de l'enquête soit sensibilisée au problème de mesure.