Enquête Génération 2010. Note sur l'attrition et la déformation de l'échantillon des enquêtes

Mady Cissé

▶ **To cite this version:**

Mady Cissé. Enquête Génération 2010. Note sur l'attrition et la déformation de l'échantillon des enquêtes. [Rapport Technique] Céreq Etudes n° 28, Céreq. 2019, 14 p. halshs-02889458

HAL Id: halshs-02889458
https://shs.hal.science/halshs-02889458

Submitted on 3 Jul 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

CÉREQ ÉTUDES
28
2019

Enquête Génération 2010

Note sur l'attrition

Mady CISSE

Équipe ingénierie et gestion d'enquête,
Département des Entrées et Évolutions dans la Vie Active (DEEVA), Céreq.

Céreq

Enquête Génération 2010

Note sur l'attrition et la déformation de l'échantillon des enquêtes

Auteur :
Mady CISSE

Membre de l'Équipe ingénierie et gestion d'enquête
Département entrées et évolutions dans la vie active du Céreq.

La présente note a pour objectif d'analyser le taux de réponse, l'attrition ainsi que la déformation de l'échantillon au cours des enquêtes auprès de la Génération 2010, effectuées en 2013, 2015 et 2017. Après un retour sur les différentes étapes de déformation de l'échantillon et de corrections des poids au cours des enquêtes, l'attention sera portée sur les taux de réponse des différentes enquêtes avant d'observer la déformation de l'échantillon selon les principaux points d'entrée adoptés dans les études mobilisant les enquêtes Génération. Par la suite, cette note se focalisera sur les évolutions et corrections de poids au cours de l'enquête, ainsi que sur leurs implications sur la précision d'indicateurs communs aux trois enquêtes.

1. Retour sur les différentes étapes de l'enquête

La base de sondage de l'enquête Génération, constituée spécifiquement par le Céreq, se compose d'une liste d'individus présumés sortant du système éducatif en 2010. Il n'existe pas de base d'élèves centralisée couvrant l'ensemble des formations initiales dispensées en France. C'est pourquoi le Céreq doit la construire à partir de différentes sources : des bases d'élèves centralisées par certains ministères ainsi qu'une collecte de liste d'élèves directement auprès de certains établissements de formation.

La base de sondage présente un défaut de sous-couverture dû aux établissements qui n'ont pas fourni de bases d'élèves lors de cette collecte. Elle présente également un défaut de sur-couverture : certains individus hors-champ de l'enquête poursuivant leur scolarité en changeant d'établissement n'ont pas pu être identifiés en amont de la base de sondage. La base comporte aussi des doublons : certains jeunes inscrits dans plusieurs établissements n'ont pas été repérés en amont de l'enquête.

Le plan de sondage de l'échantillon de l'enquête 2013 est stratifié par région de l'établissement de formation et niveau de formation (secondaire, supérieur court et supérieur long) selon 69 strates ; il est équilibré par type et spécialité de formation au sein de chacune des strates. Il s'agit d'un tirage à probabilités inégales prenant en compte pour chaque individu le taux de couverture estimé du type d'établissement dont il est issu, ainsi que d'une probabilité de répondre à l'enquête 2013, estimée à partir des enquêtes 2007 auprès de la Génération 2004 et 2010 auprès de la Génération 2007. Un sur-échantillonnage a été effectué pour certaines populations cibles d'extensions demandées par des partenaires institutionnels.

Le processus de réponse à l'enquête 2013 a été découpé en trois phases : l'obtention d'un contact téléphonique avec l'individu ou un proche, l'acceptation de la participation à l'enquête ainsi que l'achèvement du questionnaire. Il en résulte trois étapes de pertes d'individus et de corrections de poids dues à la non-réponse par modélisation logistique puis groupes de réponse homogènes, suivies d'une étape de calage sur les effectifs de sortants de formation initiale donnés par la DEPP.

Un échantillonnage a été effectué pour l'enquête 2015 afin d'exclure de la ré-interrogation les individus échantillonnés dans le cadre des extensions non reconduites. Le traitement post-collecte de l'enquête 2015 consiste en une correction de la non-réponse par modélisation logistique, puis par groupes de réponse homogènes. Un calage a été effectué sur les mêmes données, provenant de la DEPP, que pour l'enquête 2013.

L'échantillon de l'enquête 2017 est constitué de l'ensemble des répondants à l'enquête 2015, exemptés des décès et refus de répondre notifiés avant le démarrage de l'enquête. Le traitement de la non-réponse est similaire à celui effectué pour l'enquête 2015. Un calage est effectué sur les données de la DEPP et des indicateurs des enquêtes à 3 et 5 ans.

2. Taux de réponse et attrition

L'échantillon initial de l'enquête auprès de la Génération 2010 est composé de 279 484 individus présumés sortants du système éducatif en France métropolitaine et DOM en 2010. Les taux de réponse des différentes interrogations sont les suivants :

	Enquête 2013			Enquête 2015			Enquête 2017		
Niveau de sortie	Taille de l'échantillon	Taux de réponse	Taux de réponse corrigé	Taille de l'échantillon	Taux de réponse	Taux de réponse cumulé	Taille de l'échantillon	Taux de réponse	Taux de réponse cumulé
Secondaire	145 494	27,1%	31,4%	10 668	50,0%	13,5%	5 334	54,7%	7,4%
Supérieur Court	66 896	31,1%	35,8%	5 770	61,1%	19,0%	3 524	67,7%	12,9%
Supérieur Long	67 094	27,4%	33,8%	7 049	68,8%	18,8%	4 849	74,6%	14,0%
Tous	279 484	28,1%	33,0%	23 487	58,4%	16,4%	13 707	65,1%	10,7%

Tableau 1 : Taux de réponse des enquêtes auprès de la Génération 2010

Le taux de réponse de l'enquête Génération 2010 à 3 ans est de 28,1 %. Ce taux relativement faible s'explique principalement par deux facteurs. D'une part, il est difficile de récupérer des coordonnées valides pour les individus de l'échantillon : elles sont récupérées par le biais de l'établissement de sortie et datent de 3 ans avant l'enquête ; les individus ont pu changer de numéro de téléphone durant cette période.

D'autre part, l'échantillonnage tient compte d'une probabilité anticipée de répondre estimée à partir des enquêtes auprès des Générations 2004 et 2007. La probabilité de tirage dans l'échantillon de l'enquête est inversement proportionnelle à la probabilité estimée de répondre. Un taux de réponse corrigé de ce sur-échantillonnage a été évalué à 33 % en pondérant les individus par leur probabilité de réponse anticipée ; il est une estimation du taux de réponse de l'enquête si la stratégie d'échantillonnage ne tenait pas compte de la probabilité de réponse. De fait, le taux de réponse réel est volontairement dégradé pour obtenir plus de précision sur les populations répondant peu, en les échantillonnant davantage.

Les taux de réponse aux ré-interrogations en 2015 et 2017 ans sont très supérieurs à celui de la première interrogation, quel que soit le niveau de sortie d'études. L'interrogation en 2013 a permis de récupérer des coordonnées valides deux ans avant la date de l'enquête, et une mise à jour des coordonnées téléphoniques a été proposée avant les ré-interrogations pour limiter l'attrition. Cependant, le taux de réponse à ces ré-interrogations est croissant avec le niveau de sortie des études et les différences sont très élevées entre les plus bas niveaux et les plus hauts.

Les taux de réponse cumulés en 2015 et 2017 sont de 16,4 % et 10,7 %, c'est-à-dire que seuls 16,4 % des individus du champ échantillonnés en 2013 répondent en 2015, et 10,7 % de cet échantillon répondent en 2017. Il est moins élevé pour les sortants du secondaire du fait qu'à chaque enquête ils répondent en moins grande proportion que les autres.

Pour la première interrogation, le découpage selon le taux de contact, taux d'acceptation de répondre à l'enquête et taux de répondants dans le champ permet de détecter les causes de la non-réponse.

Type d'établissement	Taux de contact
Ecoles Normales Supérieures	81,0%
Ecoles administrations publiques	78,7%
Ecoles professions sociales	77,0%
Lycées agricoles	76,9%
Ecoles professions de la santé	73,5%
Ecoles formations agricoles	71,5%
Centres privés d'enseignement	70,2%
Ecoles de la DGA	69,5%
Centre de Formation des Apprentis	66,8%
Ecoles secteur industriel	65,1%
Universités	62,6%
Facultés privées	62,4%
DRJS	59,8%
DGAFP	59,6%
Ecoles ingénieurs	59,3%
Ecoles secteur service	58,3%
Lycées et collèges MEN	57,2%
Ecoles ministère de la culture	56,8%
Ecoles de commerce	56,0%
IEP	46,6%
CIFRE	38,0%
Moyenne	62,2%

Tableau 2 : Taux de contact par type d'établissement de sortie

Les difficultés de contact des individus sont très dépendantes de la qualité des coordonnées transmises par les établissements de formation, ou provenant de bases centralisées des ministères. Ainsi, selon le type d'établissement et niveaux de sortie, les taux de contact varient entre 38 % pour les conventions CIFRE et 81 % pour les Ecoles Normales Supérieures.

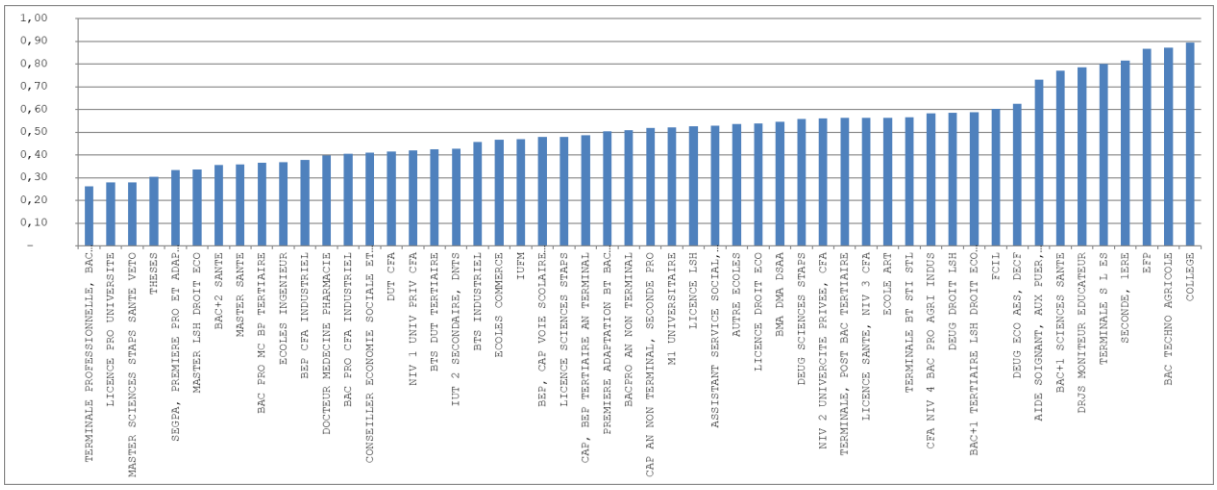

Tableau 3 : Taux de hors champ par niveau de sortie

Le taux de hors champ varie surtout en fonction du niveau de sortie des études : les années non terminales et de bas niveau ont les plus gros taux de hors champ, du fait de la poursuite d'études de certains individus en changeant de formation. Les écoles de la fonction publique ont également un fort taux de hors champ, lié au fait qu'un grand nombre de sortants de ces écoles ne sont pas en formation initiale.

3. Déformation de l'échantillon

Au fil des différentes enquêtes, les taux de réponse et l'attrition ne sont pas uniformes sur l'ensemble des individus de l'échantillon. Il en résulte que certaines populations se raréfient lors de l'avancement de l'enquête. Il s'agit ici de regarder l'évolution de la répartition du nombre d'individus répondants dans le temps selon quelques paramètres.

Plus Haut Diplôme	Répondants 2013	Echantillon 2015	Répondants 2015	Repondants 2017	Proportion de référence[1]
NON DIPLOME	11,59	13,76	10,66	7,36	16,2
CAP-BEP-MC	15,71	16,76	14,15	11,71	14,55
BACCALAUREAT	24,50	24,97	23,36	22,24	27,4
BTS-DUT, AUTRE BAC+2	8,61	9,75	10,29	10,85	11,77
BAC+2/3 SANTE/SOCIAL	8,74	3,80	4,56	5,52	3,21
BAC+3/4 HORS SANTE/SOCIAL	10,76	8,27	9,34	10,24	9,74
M2, AUTRE BAC+5, DOCTORAT	20,09	22,70	27,65	32,09	17,13

Tableau 4 : Répartition du nombre d'individus selon le niveau de diplôme

Au cours du temps, la proportion de répondants de chacun des niveaux de diplôme du secondaire diminue, alors qu'elle augmente pour chaque niveau de diplôme du supérieur. Plus le niveau est élevé, plus la proportion augmente ; la propension à répondre aux enquêtes successives augmente avec le niveau de diplôme.

Comparativement à la répartition des sortants de formation initiale en fonction du diplôme le plus élevé publiée par la DEPP[1], les enquêtes Génération comportent toujours moins de non-diplômés, et plus de sortants de niveau I. Cependant, cette comparaison ne tient pas compte des sur-échantillonnages effectués pour les partenaires d'extension. De fait, la DGESIP a financé en 2013 une extension d'échantillon de sortants du supérieur qui a été prolongée par la suite.

[1] Repères et références statistiques 2013, fiche 8.23, tableau (1)

Les sortants de formation en santé/social ont fait l'objet d'une extension en 2013 qui n'a pas été reconduite par la suite, la proportion d'individus de ce champ dans l'échantillon de 2015 est donc plus proche du repère fourni par la DEPP qu'au sein des répondants à l'enquête 2013.

Sexe	Répondants 2013	Echantillon 2015	Répondants 2015	Repondants 2017
Homme	49,50	51,81	52,32	51,36
Femme	50,50	48,19	47,68	48,64

Tableau 5 : Répartition du nombre d'individus par sexe

Les variations du rapport hommes/femmes au cours du temps est relativement faible. Les hommes répondent légèrement plus que les femmes en 2015, mais moins en 2017. Il est à noter, cependant, que le ré-échantillonnage en 2015 a ôté plus de femmes que d'hommes ; les cibles des extensions non reconduites étaient majoritairement féminines.

L'analyse selon les différentes strates d'échantillonnage permet d'avoir un premier regard sur la qualité des indicateurs de précision fournis à partir de l'enquête. En effet, pour un plan de sondage stratifié, le calcul de variance d'un indicateur simple sur une population est la somme des variances au sein des strates. Il faut donc pouvoir estimer ces variances précisément pour obtenir des indicateurs de précision fiables.

Région	Niveau de sortie	Taille de l'échantillon	Enquête 2013	Enquête 2015	Enquête 2017
Ile-de-France	Secondaire	19 545	1828	867	466
	Supérieur Court	14 106	1657	672	449
	Supérieur Long	20 310	2465	1377	1012
Champagne-Ardenne	Secondaire	2 725	314	126	66
	Supérieur Court	1 442	253	98	68
	Supérieur Long	572	107	61	42
Picardie	Secondaire	16 476	2164	238	126
	Supérieur Court	1 686	299	112	74
	Supérieur Long	649	128	69	54
Haute-Normandie	Secondaire	3 331	385	188	105
	Supérieur Court	1 637	299	108	73
	Supérieur Long	1 360	170	78	59
Centre	Secondaire	4 639	461	206	118
	Supérieur Court	2 187	350	134	89
	Supérieur Long	1 026	227	123	97
Basse-Normandie	Secondaire	2 659	297	130	75
	Supérieur Court	1 434	226	90	70
	Supérieur Long	708	138	73	55
Bourgogne	Secondaire	2 306	245	101	54
	Supérieur Court	1 452	274	85	58
	Supérieur Long	1 394	143	73	53
Nord-Pas-de-Calais	Secondaire	10 339	1174	458	221
	Supérieur Court	6 032	1000	298	195
	Supérieur Long	5 592	830	284	203
Lorraine	Secondaire	3 222	398	163	85
	Supérieur Court	2 179	458	106	74
	Supérieur Long	1 947	294	138	102
Alsace	Secondaire	2 318	316	155	83
	Supérieur Court	1 459	298	103	75
	Supérieur Long	1 774	317	159	132
Franche-Comté	Secondaire	2 447	282	109	59
	Supérieur Court	1 541	273	105	66
	Supérieur Long	786	132	65	47
Pays de la Loire	Secondaire	5 534	583	257	148
	Supérieur Court	3 713	455	158	113
	Supérieur Long	3 652	515	234	182
Bretagne	Secondaire	4 297	480	203	125
	Supérieur Court	1 875	427	128	84
	Supérieur Long	2 251	454	204	162
Poitou-Chanrentes	Secondaire	2 442	251	105	57
	Supérieur Court	1 378	191	73	49
	Supérieur Long	1 015	224	113	81
Aquitaine	Secondaire	5 119	490	231	138
	Supérieur Court	2 975	462	167	129
	Supérieur Long	2 613	461	238	189
Midi-Pyrénées	Secondaire	4 220	393	168	110
	Supérieur Court	2 861	385	126	84
	Supérieur Long	3 048	516	268	201
Limousin	Secondaire	1 953	186	81	50
	Supérieur Court	852	193	95	72
	Supérieur Long	511	116	69	63
Rhône-Alpes	Secondaire	10 220	1046	448	251
	Supérieur Court	5 343	685	252	170
	Supérieur Long	8 728	956	463	334
Auvergne	Secondaire	2 289	213	106	59
	Supérieur Court	1 420	188	71	50
	Supérieur Long	852	193	85	64
Languedoc-Roussillon	Secondaire	5 225	438	187	101
	Supérieur Court	2 552	397	138	99
	Supérieur Long	2 652	483	226	174
Provence-Alpes-Côte d'Azur	Secondaire	26 365	1839	479	247
	Supérieur Court	6 604	616	241	144
	Supérieur Long	4 574	671	349	236
Corse	Secondaire	2 267	150	76	39
	Supérieur Court	896	143	78	54
	Supérieur Long	633	99	53	38
Départements d'Outre Mer	Secondaire	5 556	494	252	133
	Supérieur Court	1 272	184	87	48
	Supérieur Long	447	95	49	36
Total	Secondaire	145 494	14427	5334	2916
	Supérieur Court	66 896	9713	3525	2387
	Supérieur Long	67 094	9734	4851	3616
	Total	279 484	33874	13710	8919

Tableau 6 : Nombre d'individus par strate d'échantillonnage

Les différentes strates d'échantillonnage ne sont pas homogènes en termes de population, et l'effet de l'attrition n'est pas homogène non plus ; les strates de sortants du secondaire ayant une plus forte attrition. De ce fait, certaines strates d'échantillonnage ont un nombre de répondants faible à très faible en 2015 et 2017. De plus, puisque le traitement de la non-réponse et les différents calages ne se font pas spécifiquement au sein de ces strates, la variance des poids au sein de ces strates peut être élevée. La précision de l'enquête sur ces strates risque donc d'être à la fois limitée et mal évaluée.

4. Corrections et distribution des poids

Le taux de réponse aux différentes enquêtes n'est pas homogène sur toute la population d'intérêt. En particulier, les individus sortants à différents niveaux d'études n'ont pas la même probabilité d'être contactés, ou de répondre à l'enquête. Les étapes de redressement de la non-réponse à chaque enquête permettent de conserver une certaine représentativité tout au long du processus d'enquêtes en modifiant les poids associés à chaque individu en fonction de sa probabilité estimée de répondre. L'objectif initial de l'échantillon de l'enquête 2013 est d'avoir un ensemble de répondants représentant l'ensemble des sortants de formation initiale en 2010. Sur les enquêtes 2015 et 2017, l'objectif est de garantir une stabilité de la composition pondérée de l'échantillon sur l'ensemble des caractéristiques observées en 2013, pour conserver une cohérence entre les enquêtes. L'enjeu principal est de minimiser les corrections de poids afin d'éviter les écarts trop importants entre les poids des individus qui tendent à diminuer la précision des indicateurs calculés à partir de l'enquête.

	Enquête 2013								Enquête 2015				Enquête 2017		
	Echantillon	Individus contactés		Individus répondant		Individus dans le champ teminant le questionnaire		Calage	Echantillon	Individus répondant		Calage	Individus répondant		Calage
Niveau de sortie		Brut	Après correction	Brut	Après correction	Brut	Après correction			Brut	Après correction		Brut	Après correction	
1-COLLEGE	9,81	8,26	9,60	10,22	9,48	1,99	1,82	2,71	2,71	1,97	2,67	2,67	1,73	2,63	2,60
2-LYCEE GENERAL TECHNOLOGIQUE	14,79	15,05	14,69	15,03	14,52	8,96	7,60	8,98	8,98	8,08	9,39	9,38	8,83	9,44	9,55
3-CAP BEP MC BAC PRO	29,41	28,52	29,49	28,11	29,92	36,01	35,89	35,42	35,41	31,00	36,23	36,19	31,50	36,59	36,22
4-BTS DUT, Autre BAC+2	18,81	20,04	18,85	18,98	18,96	20,00	18,65	19,56	19,48	19,46	19,78	19,79	20,33	19,05	19,45
5-BAC+2/3 SANTE SOCIAL	3,48	3,65	3,61	3,59	3,66	3,64	4,99	4,72	4,72	5,76	4,88	4,88	6,18	5,08	5,00
6-BAC+3/4 HORS SANTE SOCIAL	6,98	7,42	6,93	7,20	6,90	8,38	8,11	7,67	7,64	8,58	6,81	6,81	7,75	6,85	6,95
7-M1 M2 ECOLES Autre BAC+5	14,92	15,31	14,97	15,06	14,71	18,37	19,90	18,17	18,29	21,68	17,52	17,56	20,14	17,42	17,48
8-DOCTORAT	1,81	1,75	1,86	1,80	1,86	2,65	3,03	2,77	2,76	3,48	2,72	2,72	3,55	2,94	2,74

Tableau 7 : Répartition des poids en fonction du niveau de sortie au cours des enquêtes auprès de la Génération 2010

Chaque étape de correction de l'absence de contact ou de non-réponse a pour objectif de retrouver la structure de l'échantillon qui a été dégradée par le processus spécifique à chaque catégorie d'individus de l'échantillon. En analysant la structure de l'échantillon par niveau de sortie des études tout au long de l'enquête, il apparaît que chacune de ces étapes permet de redresser de façon très satisfaisante l'échantillon pour ce paramètre. Par exemple, les sortants de collège qui représentaient 9,81 % de l'échantillon tiré pour l'enquête 2013 ne représentent plus que 8,26 % du total des poids des individus contactés ; leurs poids corrigés après contact représentent 9,60 % du total.

La correction de la non-réponse de l'enquête 2015 ne parvient toutefois pas totalement à redresser parfaitement les poids des sortants de niveau BTS/DUT, bac+2 ainsi que bac+3/4 et Master bac+5, avec

des écarts dépassant 0,5 points dans la répartition des poids entre l'échantillon et les poids corrigés après réponse. Le calage sur les marges de l'enquête 2013 corrige cependant ces écarts.

L'étape de correction liée à la complétion du questionnaire au complet est spécifique du fait qu'elle ne traite pas que d'un défaut de réponse (les individus abandonnant en cours d'enquête) mais inclut aussi l'exclusion des individus ayant participé à l'interrogation jusqu'à son terme mais identifiés comme hors du champ de l'enquête. Ainsi les sortants de niveaux secondaires, qui sont en grande partie hors-champ dû à une poursuite d'études non repérée en amont de l'enquête, voient leur part diminuer drastiquement dans l'enquête pour refléter leur poids dans les sortants de formation initiale en 2010.

Cependant, l'étape de calage de l'enquête 2013 modifie assez conséquemment la répartition des poids, montrant les limites à la fois de la correction de la non-réponse, mais également de la correction des défauts de la base de sondage en amont.

Dans le but d'obtenir plus facilement un nombre suffisant de questionnaires des populations répondant peu ainsi que de limiter la dispersion des poids des individus de l'enquête, la stratégie d'échantillonnage intègre un paramètre lié à une probabilité de réponse estimée en amont de l'enquête à partir de modélisations effectuées sur les enquêtes précédentes. Plus précisément, la probabilité d'inclusion à l'échantillon de chaque individu est inversement proportionnelle à la probabilité de réponse anticipée[2]. Pendant l'étape de correction de la non-réponse, les poids initiaux inverses de la probabilité d'inclusion sont divisés par une probabilité de réponse estimée après l'enquête selon un modèle différent.

Moyenne	Ecart-type	Minimum	5ème ctl	10ème ctl	25ème ctl	50ème ctl	75ème ctl	90ème ctl	95ème ctl	Maximum
0,031	0,104	-0,421	-0,142	-0,101	-0,032	0,031	0,102	0,162	0,198	0,336

Tableau 8 : Répartition de la différence entre la probabilité de répondre estimée pour le redressement de la non-réponse et celle estimée a priori

La différence est en moyenne 0,03 ; les probabilités de réponse anticipées sont, en moyenne, plus pessimistes que celles estimées a posteriori. Pour la plupart des individus, la différence est proche de 0 entre les deux estimations, pour la moitié d'entre eux elle est comprise entre -0,10 et 0,03. Cependant les deux modélisations ne prédisent pas des valeurs proches pour l'ensemble des individus, pour 5 % elles diffèrent d'au moins 0,20, jusqu'à 0,33 ; elles peuvent différer de -0,42 en regardant les valeurs négatives. Ces différences élevées peuvent conduire à une dispersion des poids élevée, ce qui est l'objectif inverse de l'anticipation de la probabilité de répondre dans l'échantillonnage.

En effet, il est possible de construire pour chaque individu un facteur d'évolution de son poids à cause de la prise en compte de la non-réponse qui se constitue du rapport entre la probabilité de réponse anticipée et la probabilité de réponse estimée pour le redressement de la non-réponse[3]. Il permet de visualiser l'impact de l'anticipation des probabilités de répondre en comparaison avec un

[2] Pour certains individus, la multiplication de la probabilité d'inclusion socle par l'inverse de la probabilité de réponse anticipée dépasse 1, leur probabilité d'inclusion est alors fixée à 1.

[3] Pour les individus dont la probabilité d'inclusion est 1, il est substitué à la probabilité de réponse anticipée la probabilité d'inclusion socle. En effet, puisque la probabilité de tirage finale est égale à 1, l'effet de la prise en compte en amont de la non-réponse est de diviser la probabilité socle par elle-même, et non plus par la probabilité de répondre anticipée

échantillonnage classique où la seule modification de poids liée à la non-réponse est la multiplication par l'inverse de la probabilité de répondre

Variable	MEAN	STD	MIN	Q1	MEDIAN	Q3	MAX	MAX/MIN	IQR	CV
Facteur de correction de la non-réponse	1,17	0,41	0,16	0,90	1,10	1,37	4,81	30,68	1,51	0,35
Inverse de la probabilité de réponse	3,63	1,95	1,52	2,45	3,09	4,24	58,53	38,61	1,73	0,54

Tableau 9 : Répartition du facteur de correction de la non-réponse et de l'inverse de la probabilité de répondre

Le facteur de correction est naturellement inférieur à l'inverse de la probabilité de réponse puisque la probabilité de réponse anticipée est comprise entre 0 et 1. En moyenne, dans l'enquête 2013, les poids sont multipliés par 1,7 à cause de la non-réponse dans l'échantillonnage, alors qu'ils l'auraient été par 3,6 sans cette prise en compte[4].

« L'analyse comparative des statistiques de dispersion permet de montrer une relative efficacité de l'anticipation de la non-réponse du point de vue de la dispersion des poids. En effet, pour le rapport max/min, le rapport Q3/Q1 ou le coefficient de variation, le facteur de correction utilisé est inférieur à l'inverse de la probabilité de réponse du modèle de non-réponse ; ce qui indique une dispersion des poids moindre a posteriori que dans le cas où la non-réponse n'avait pas été anticipée et donc une situation dans laquelle les poids initiaux auraient été plus homogènes. Les poids finaux de chaque enquête sont ventilés, et évoluent, comme ceci selon les différents niveaux de plus haut diplôme obtenu :

	Enquête 2013			Enquête 2015			Enquête 2017		
Plus haut diplôme obtenu en 2010	Min	Moyenne (CV)	Max	Min	Moyenne (CV)	Max	Min	Moyenne (CV)	Max
1-NON DIPLOME	3,34	**31,75** (66,84)	170,76	15,23	**81,34** (47,17)	403,29	32,75	**183,39** (71,48)	1057,94
2-CAP-BEP-MC	2,15	**19,17** (63,15)	89,10	6,16	**52,74** (42,30)	189,29	14,91	**99,05** (64,32)	691,14
3-BACCALAUREAT	2,35	**23,98** (77,15)	143,83	7,52	**63,73** (49,06)	329,34	7,73	**103,11** (57,82)	492,62
4-BTS-DUT, AUTRE BAC+2	2,39	**27,30** (54,93)	104,75	7,00	**55,16** (39,87)	182,98	9,68	**80,57** (47,52)	330,48
5-BAC+2/3 SANTE SOCIAL	1,89	**9,74** (74,13)	70,67	5,75	**45,51** (42,72)	102,22	6,41	**57,82** (46,86)	224,01
6-BAC+3/4 HORS SANTE SOCIAL	2,32	**19,89** (71,72)	104,75	5,38	**45,15** (41,68)	143,67	7,08	**63,37** (55,02)	429,53
7-M2,AUTRE BAC+5	1,99	**23,11** (68,02)	98,88	4,18	**46,73** (39,55)	163,91	5,63	**65,12** (55,00)	541,45
8-DOCTORAT	1,58	**8,19** (141,04)	81,13	2,07	**11,01** (139,50)	109,43	2,42	**13,80** (165,66)	496,26

Tableau 10 : Répartition des poids par plus haut diplôme obtenu en 2010.

Pour l'enquête 2013, les différents niveaux de diplôme ont des poids moyens différents. Les docteurs ont en moyenne les poids les plus faibles car une extension sur les docteurs hors-santé a induit un sur-échantillonnage massif de cette population. Cependant les docteurs en santé n'ont pas été cibles de cette extension et ont des poids plus élevés, ce qui explique le coefficient de variation élevé des poids

[4] Les modèles de correction de la non-réponse étant pondérés par les poids d'échantillonnage, un changement de la stratégie d'échantillonnage ne les perturbe pas s'ils sont bien spécifiés.

des individus de ce niveau de diplôme. Le sur-échantillonnage est reconduit sur les trois enquêtes ; les poids moyens sont les plus faibles et le coefficient de variation reste le plus élevé tout au long du processus d'enquêtes.

Les coefficients de variation des poids se sont réduits pour chacun des niveaux de diplôme entre les deux premières enquêtes. Cela s'explique par la non reconduction en 2015 de la plupart des extensions d'échantillon de 2013 : les poids des individus ré-échantillonnés pour participer en 2015 ont alors été rehaussés pour devenir similaires, dans leur composition, aux individus non-ciblés par ces extensions. Par exemple, les diplômés de formations de niveau bac+2/3 en santé/social qui sont, en grande partie, cibles d'une extension non reconduite voient leur poids moyen augmenter de 9,74 à 45,15. Cependant, la hausse des poids entre les enquêtes 2013 et 2015 s'explique également par le traitement de la non-réponse.

Les changements de poids entre 2015 et 2017 sont uniquement provoqués par le traitement de la non-réponse à l'enquête 2017. L'effet de l'hétérogénéité de l'attrition sur les différents niveaux de diplôme se traduit par l'augmentation différente des poids moyens de chacun de ces niveaux ; l'effet de la diversité des individus au sein de ces strates par l'augmentation du CV. Les individus ayant les poids maximums en 2015 font partie de ceux ayant la plus faible probabilité de réponse en 2017, leurs poids sont donc grandement augmentés, pour atteindre plus de 400 fois les poids minimums (1057,94 pour le maximum, 2,42 pour le minimum), alors que c'était près de 200 fois en 2015 et 110 fois en 2013.

Les effets de ces changements de répartition des poids entre les différentes enquêtes se manifestent sur la précision de l'enquête dans l'exemple suivant : l'évolution de la précision de l'estimation de la part d'individus de la Génération 2010 ayant reçu une bourse sur critères sociaux durant leurs études.

	Enquête 2013				Enquête 2015				Enquête 2017			
Plus Haut Diplôme	Estimation	Ecart-type	Coefficient de variation	Intervalle de confiance à 95%	Estimation	Ecart-type	Coefficient de variation	Intervalle de confiance à 95%	Estimation	Ecart-type	Coefficient de variation	Intervalle de confiance à 95%
NON DIPLOME	**24,1%**	0,5%	2,0%	23,2% - 25,0%	**24,6%**	1,1%	4,5%	22,4% - 26,7%	**23,8%**	1,9%	8,2%	20,0% - 27,6%
CAP-BEP-MC	**18,7%**	0,7%	3,6%	17,4% - 20,0%	**18,1%**	1,0%	5,5%	16,1% - 20,1%	**19,3%**	1,4%	7,3%	16,5% - 22,0%
BACCALAUREAT	**31,0%**	0,5%	1,6%	30,0% - 31,9%	**30,9%**	0,8%	2,7%	29,3% - 32,6%	**30,7%**	1,1%	3,6%	28,5% - 32,8%
BTS-DUT, AUTRE BAC+2	**38,9%**	0,9%	2,2%	37,2% - 40,6%	**37,1%**	1,3%	3,4%	34,6% - 39,6%	**38,1%**	1,6%	4,1%	35,1% - 41,2%
BAC+2/3 SANTE/SOCIAL	**35,9%**	1,0%	2,8%	33,9% - 37,9%	**38,6%**	2,2%	5,6%	34,4% - 42,9%	**37,9%**	2,5%	6,6%	33,0% - 42,7%
BAC+3/4 HORS SANTE/SOCIAL	**45,5%**	0,9%	2,0%	43,7% - 47,3%	**45,1%**	1,8%	4,0%	41,5% - 48,6%	**44,6%**	2,1%	4,6%	40,5% - 48,6%
M2, AUTRE BAC+5	**37,2%**	0,8%	2,2%	35,6% - 38,8%	**36,9%**	1,1%	3,0%	34,7% - 39,1%	**38,3%**	1,4%	3,6%	35,7% - 41,0%
DOCTORAT	**28,0%**	1,4%	4,9%	25,3% - 30,7%	**27,1%**	1,7%	6,4%	23,7% - 30,5%	**28,4%**	2,8%	9,9%	22,9% - 33,9%
Total	**31,4%**	0,3%	0,8%	30,9% - 31,9%	**30,9%**	0,4%	1,4%	30,0% - 31,8%	**31,1%**	0,6%	2,0%	29,9% - 32,4%

Tableau 11 - Intervalles de confiance à 95 % de la part de boursiers par niveau de diplôme sur chacune des enquêtes.

La part d'élèves boursiers sur l'ensemble de la population est estimée entre 30,9 % et 31,9 % à l'aide de l'enquête 2013. La longueur de l'intervalle de confiance à 95 % est d'1 point de pourcentage. Pour la même variable estimée à l'aide des répondants et des poids de l'enquête 2015, la longueur de l'intervalle devient 1,5 point, puis de 2,4 avec les données de l'enquête 2017.

En comparant les différents niveaux de diplôme, les intervalles de confiance à 95 % des estimations de part d'étudiants boursiers parmi les non-diplômés et parmi les diplômés de CAP-BEP-MC sont disjoints en 2013 et en 2015, mais se recoupent en 2017. À partir des résultats de l'enquête 2017, il est donc impossible de dire que les sortants non-diplômés ont, en moyenne, reçu plus de bourses sur critère sociaux que les sortants de CAP-BEP-MC durant leur scolarité, alors que c'est possible à partir des deux autres enquêtes. La dégradation de la précision de l'enquête, montrée sur une variable présente dans la 1ère enquête, dévoile les limites des interprétations des questions présentes en 3ème interrogation.

ISSN
2497-6873

Céreq

Établissement public national sous la tutelle du ministère chargé de l'Éducation et du ministère chargé de l'Emploi.

DEPUIS 1971

Mieux connaître les liens formation - emploi - travail.
Un collectif scientifique au service de l'action publique.

• **12** centres associés sur le territoire et de nombreuses coopérations internationales

+ d'infos
et tous les travaux

À explorer
www.cereq.fr

+ de 600 publications
Accessibles librement