Direction des Statistiques Démographiques et Sociales
Département de l'Emploi et des Revenus d'Activités
Division Emploi

NOTE

Dossier suivi par :
Fabrice Murat
Tél. : 01 41 17 54 52
Fax : 01 41 17 61 63
Messagerie : fabrice.murat@insee.fr

Paris, le
N° /F230

Objet : Calcul d'une pondération globale en 2005 pour le panel d'élèves entrés en 6e en 1995

La DEPP (Direction de l'évaluation, de la prospective et de la performance) a commencé à suivre en 1995 un échantillon de 17 830 élèves entrant en 6e cette année-là, dans l'ensemble des établissements secondaires publics ou privés : il s'agissait des élèves nés le 17 d'un mois (sauf mars, juillet et octobre, pour obtenir un taux de 1/40e environ). Depuis cette date, outre une attrition « naturelle » due aux décès et aux départs à l'étranger, l'échantillon a régulièrement diminué, du fait des refus de répondre et des pertes de contact avec les enquêtés (notamment suite à un déménagement). D'autre part, jusqu'en 2005, l'enquête était entièrement gérée par la DEPP, qui ne suivait que les jeunes encore scolarisées. Depuis cette date, un dispositif a été mis en place à l'Insee pour retrouver les jeunes sortis de l'échantillon géré par la DEPP, soit parce qu'ils avaient quitté le système éducatif, soit parce qu'ils avaient été perdus. L'échantillon DEPP et l'échantillon Insee font l'objet de deux interrogations séparées, un certain nombre de questions étant toutefois communes aux deux opérations. Dans l'échantillon DEPP, il faut aussi distinguer la partie « Enseignement secondaire » et la partie « Enseignement supérieur », qui font l'objet d'une collecte et de questionnaires différents.

La DEPP procède chaque année au calcul d'une pondération pour corriger la non-réponse, faible mais non nulle, à l'enquête dans l'enseignement supérieur (le taux de non-réponse est inférieur à 10 %). L'objectif de cette note est d'effectuer ce traitement sur l'ensemble des individus, qu'ils soient dans l'échantillon Insee, dans celui des jeunes scolarisés dans l'enseignement supérieur, dans celui des jeunes encore scolarisés dans l'enseignement secondaire[1].

Constitution du fichier des jeunes interrogés en 2005

La première étape consiste à attribuer aux 17 830 personnes de l'échantillon initial un statut de réponse en 2005, distinguant : hors-champ (les décès notamment), les répondants, les

[1] Certains jeunes interrogés par l'Insee se sont révélés encore en formation initiale en 2005. Ils ont été « rendus » à la DEPP pour qu'elle les suive jusqu'à la fin de leurs études. Pour le calcul des pondérations, on considèrera qu'ils appartiennent à l'échantillon Insee en 2005.

non-répondants. Pour ce faire, il faut apparier les différentes sources de données disponibles :

- le fichier « historique » du panel concernant les élèves dans l'échantillon initial (17 830 individus) : il comporte des informations sur ces élèves jusqu'à l'année scolaire 2002-2003, y compris les données de l'enquête Jeune 2002 ;
- le fichier des jeunes à retrouver, transmis à l'Insee par la DEPP (6 893 individus) ;
- le fichier des répondants à l'enquête de l'Insee (3 257 individus) ;
- le fichier des adresses jugées inexploitables en 2005 (2 282 individus)[2] ;
- le fichier des enquêtés dans l'enseignement secondaire (1 654 individus dont 955 répondants) ;
- le fichier des enquêtés dans l'enseignement supérieur (8 249 individus dont 7 509 répondants ; en fait, ce fichier comporte, en plus de ces 8 249 personnes, 1 202 individus que la DEPP devait normalement transmettre à l'Insee en 2005 et que l'on auraient dû donc retrouver tous dans le fichier des jeunes à retrouver, ainsi que 327 individus qui n'ont pas été interrogés par la DEP en 2005 mais n'ont pas été transmis à l'Insee en 2005 ; ils l'ont été en 2006).

Les 4 derniers fichiers (décrivant la situation des jeunes en 2005) ont d'abord été appariés avec la variable IDENTSCO (identifiant long construit avec le nom, le prénom, la date de naissance, etc.). Cet appariement a fait apparaître différents problèmes, soumis aux responsables de la collecte à la DEPP et à l'Insee, qui ont apporté les précisions permettant de trouver une solution :

c1. 1 individu se trouve dans le fichier des répondants à l'enquête Insee mais pas dans l'échantillon des jeunes transmis par la DEPP. Il a en effet été transmis avec un peu de retard. Il fait donc clairement parti de l'échantillon Insee.

c2. 327 individus ont été repérés comme transmis à l'Insee seulement pour 2006 (voir plus haut) : ce sont des bacheliers 2002, qui ont répondu à la première enquête dans l'enseignement supérieur, mais pas à la seconde. Ils ont été oubliés lors de la transmission entre la DEPP et l'Insee. Ils sont considérés comme non-répondants à l'enquête SUP en 2005, car leurs caractéristiques sont a priori proches de cette population.

c3. 22 individus sont indiqués dans le fichier de l'enseignement supérieur comme transmis à l'Insee (la variable ECH05 vaut 3) mais ne se retrouvent pas dans le fichier correspondant (l'un est dans le fichier de l'enseignement secondaire, mais il ne répond pas à l'enquête) ; leur situation a en effet été stabilisée tardivement et ils n'ont pu être envoyés à l'Insee à temps pour 2005 ; l'un a été transmis pour 2006 ; tous les autres le seront pour 2007 ; en 2005, on les considère dans l'échantillon Insee comme « non-répondants » ; il en sera de même en 2006, à l'exception de l'individu transmis (signalons que certains ont été quand même interrogés par la DEPP en 2006 et que des données peuvent être récupérées).

c4. 10 individus sont à la fois dans le fichier des jeunes transmis à l'Insee et dans celui de l'enseignement supérieur (en excluant les cas où ECH05=3). Ils ont été transmis par erreur à l'Insee, on les laisse dans le fichier DEPP pour l'enseignement supérieur (il faudra aussi les retirer du fichier Insee 2006 et ne pas les faire interroger par l'Insee en 2007, à moins qu'ils aient fini leurs études entre temps bien sûr). Sur ces 10 personnes, 8 répondent à l'enquête SUP (dont 5 répondent aussi à l'enquête Insee). Les 2 autres ne répondent à aucune des deux enquêtes.

c5. et c6. 10 individus sont à la fois dans le fichier des jeunes transmis à l'Insee et dans celui de l'enseignement secondaire. Deux d'entre eux (cas c5) n'ont effectivement pas le bac et doivent être laissés dans le fichier DEPP pour l'enseignement secondaire : un seul a effectivement répondu à l'enquête

[2] Ce fichier repère non seulement les adresses jugées inexploitables lors de la phase de recherche d'adresse début 2005 (cette information se trouve déjà dans le deuxième fichier décrit ici), mais aussi les adresses jugées inexploitables à l'issue de la collecte courant 2005, car le questionnaire avait été retourné en NPAI. Dans les premiers travaux sur les pondérations, on n'avait pas repéré cette catégorie.

DEPP, tandis que les deux répondent à l'enquête Insee. On peut donc envisager un rapatriement des données pour celui qui ne répond pas à l'enquête DEPP et le considérer comme répondant. Les 8 autres cas (cas c6) concernent des personnes ayant refait des études secondaires après leur bac : ils relèvent de l'enquête Insee. Deux d'entre eux ont effectivement répondu à l'enquête Insee (dont 1 aussi à l'enquête DEPP dans le secondaire) ; 3 sur les 6 autres ont répondu à l'enquête DEPP dans le secondaire. Pour ces 3 personnes, on rappatriera par la suite les réponses dans le fichier Insee : on peut donc les considérer comme répondants à l'enquête Insee.

c7. 19 individus sont à la fois dans le fichier de l'enseignement secondaire et dans celui de l'enseignement supérieur (en excluant les cas où ECH05=3). Aucun n'a répondu à l'enquête dans l'enseignement supérieur ; 12 ont répondu à l'enquête dans l'enseignement secondaire. L'information sur leur réussite au bac est arrivée trop tardivement pour qu'ils soient interrogés par l'enquête SUP, mais c'est bien à ce champ qu'ils appartiennent (ils ont été interrogés en 2006). Les 12 répondants à l'enquête DEPP seront considérés comme répondants à l'enquête SUP en 2005 (plus un cas apparu dans un fichier transmis par la suite) ; les autres comme non-répondants à l'enquête SUP.

Appariement avec l'échantillon initial

Le fichier résultant de cet appariement comporte 17 106 individus. Lors de l'appariement avec le fichier historique, ils sont tous retrouvés, mais 724 individus sont dans ce dernier fichier sans être dans aucun des autres. Dans 369 cas, il s'agit d'attrition naturelle entre 1995 et 2005 : 35 décès, 118 problèmes de santé graves, 216 départs définitifs à l'étranger. Ces personnes seront considérées comme hors champ et auront une pondération à blanc.

Restent 355 cas au départ inexpliqués. Les services de la DEPP ont distingué différentes situations :

c8. 20 cas ont été transmis à l'Insee par la DEPP, bien qu'on ne les retrouve pas dans le fichier initial. On les considèrera comme non-répondants à l'enquête Insee. Il faudra s'assurer qu'ils ont été interrogés par l'Insee en 2006 et qu'ils le seront en 2007.

c9. 2 élèves ont été repérés trop tardivement comme encore dans l'enseignement secondaire pour pouvoir être interrogés : on les considèrera comme non-répondants à l'enquête DEPP de l'enseignement secondaire (en 2006, ils ont été transmis à l'Insee).

c10. 121 jeunes poursuivent des études dans l'enseignement supérieur, sans être suivis par la DEPP. En effet, 4 ont eu leur bac avec un an d'avance et ont ainsi échappé au dispositif de suivi ; les autres ne sont pas bacheliers (30 ont cependant un brevet de technicien). On propose de les considérer comme sortants du dispositif DEPP, donc entrants dans le dispositif Insee. En 2005 et en 2006, on les considèrera comme non-répondants à cette enquête. Il faudra que l'Insee les interroge en 2007.

c11. 211 jeunes ont été perdus par la DEPP (presque tous entre 2001 et 2003) sans avoir été transmis à l'Insee. On les considérera comme non-répondants à l'enquête Insee en 2005 et 2006. Il faudra que l'Insee les interroge en 2007.

c12. 1 élève est encore scolarisé dans le secondaire : il n'a pas été interrogé en 2005, mais l'a été en 2006. On le considèrera comme non-répondant à l'enquête DEPP de l'enseignement secondaire en 2005.

En définitive, en excluant les personnes hors-champ, il reste 17 461 jeunes dans le champ de l'enquête en 2005, se répartissant de la façon suivante :

- **7 256 dans le champ Insee, dont 3 253 répondants (45 %)**
- **1 629 dans le champ DEPP-secondaire dont 940 répondants (58 %)**
- **8 576 dans le champ DEPP-supérieur dont 7 522 répondants (88 %)**

Différentes pistes pour la pondération

Les trois enquêtes (Insee, DEPP-secondaire, DEPP-supérieur) se distinguent tant par les caractéristiques des populations concernées que par les questionnaires et les procédures de collecte et de relance, si bien qu'il paraît préférable d'étudier la non-réponse sur chaque enquête indépendamment.

De plus, pour la partie Insee, il faut bien prendre conscience que la non-réponse va renvoyer à trois phénomènes assez différents :

- *problèmes de transmission DEPP-Insee* : ce sont les cas c3, c6, c8, c10 et c11 évoqués au-dessus. On n'a même pas tenté de les retrouver en 2005.
- *contacts impossibles en 2005* : on n'a pas réussi à retrouver leur adresse en 2005.
- *refus en 2005* : la personne a refusé de répondre à l'enquête.

Ces trois cas relèvent sans doute de mécanismes assez différents : erreur aléatoire pour le premier, forte mobilité géographique pour le second, problème de temps ou de motivation pour le troisième… Il faudra envisager de faire trois modèles distincts. Notons que la distinction entre ces différents cas n'est pas toujours facile. Lors de la phase de recherche d'adresses, des jeunes ont renvoyé le coupon de confirmation, mais par la suite, ils n'ont pas répondu à l'enquête. Ce cas est probablement un refus de l'enquête, même si un déménagement n'est pas à exclure. Par ailleurs, on a aussi envoyé le questionnaire, alors que le coupon n'avait pas été renvoyé par le jeune, mais que le courrier n'a pas été retourné en NPAI. Un cinquième de ces questionnaires ont été remplis ; il est difficile de séparer dans les cas restants (4/5) les problèmes de contact et les refus de répondre.

Une autre distinction peut être faite selon le type d'informations disponibles. Toutes les caractéristiques des élèves sont connues en 1995, mais certains sont ensuite perdus. L'enquête Famille et l'enquête Jeune peuvent fournir certaines informations, mais elles n'ont pas été passées par tous. Des modèles spécifiques auraient pu être élaborés pour les personnes ayant répondu à ces enquêtes.

Enfin, on a aussi envisagé une procédure de calage sur marges externes. Il paraît difficile de trouver des informations en 2005 (dans les enquêtes Insee, on ne peut isoler la population des personnes entrées en 6^{e} en 1995). En revanche, les données de l'éducation nationale fournissent des statistiques plus précises sur cette population en 1995 (proportions de filles, d'enfants d'ouvriers, etc.) afin de vérifier la représentativité de l'échantillon des 17 830 élèves au départ et celui des 11 715 répondants en 2005 (après correction de la non-réponse) auxquels il faut ajouter les personnes « disparus » (décédés, partis à l'étranger, malades).

Compte tenu du fait que la pondération devra être recalculée chaque année (en intégrant l'attrition à venir et le « retour » de certains non-répondants), la procédure finalement retenue vise la meilleure qualité possible des redressements, en limitant la complexité des traitements.

Modélisation de la non-réponse

On a procédé à modélisation de la non-réponse dans les trois enquêtes, en tenant compte de différentes caractéristiques disponibles :

- les caractéristiques socio-démographiques de l'élève en 1995 (pays de naissance, profession des parents, type de ménage en 1995, etc.) ;

- les caractéristiques scolaires de l'élève à la rentrée 1995 (retard scolaire, résultats aux évaluations 6e, etc.) ;

- les informations relatives à l'année scolaire 1995-1996 (secteur de l'établissement, appartenance à une ZEP, etc.)

- quelques informations sur le parcours scolaire (orientation en 3e, optention du bac, mention éventuelle)

- le « comportement de réponse » (raison et année de la sortie de l'échantillon DEPP, réponse à l'enquête famille ou à l'enquête Jeune, résultats de la recherche d'adresse ; ces variables ont surtout une importance pour l'échantillon Insee)[3].

On procède alors à quatre modélisations sur cet échantillon.

- la non réponse dans l'échantillon « secondaire », donnant la probabilité P_{sec} de ne pas répondre à cette enquête ;

- la non réponse dans l'échantillon « supérieur », donnant la probabilité P_{sup} de ne pas répondre à cette enquête ;

- sur l'ensemble de l'échantillon Insee, on modélise la probabilité P_{eva1} de ne pas avoir avoir d'adresse[4] ;

- sur l'ensemble des adresses retrouvées de l'Insee[5], on modélise la probabilité P_{eva2} de ne pas répondre à l'enquête.

Dans ce cas, la pondération POIDS est égale à :

- l'inverse de (1- P_{sec}) pour l'échantillon « secondaire »

- l'inverse de (1- P_{sup}) pour l'échantillon « supérieur »

- l'inverse de (1- P_{eva1})* (1- P_{eva2}) pour l'échantillon Insee

Les résultats de ces modélisations sont donnés en annexes. On a généralement retenu les variables significatives au seuil de 5 %[6].

La pondération ainsi obtenu a les caractéristiques suivantes :

[3] En revanche, on a laissé de côté les données de l'enquête Jeune (non disponibles) et celles de l'enquête Famille, car cela aurait obligé à construire une double série de modèles (pour chaque échantillon, l'un sur les répondants à l'enquête Famille ; l'autre sur les non-répondants, utilisant moins de variables), sans que le gain de qualité soit très important.

[4] Cela inclut : les problèmes de transmission entre la DEPP et l'Insee et les cas où l'adresse du jeune est inconnue, que ce manque soit apparu lors de la recherche d'adresse ou lors de la collecte.

[5] On regroupe à la fois les adresses « sûres » (le coupon-réponse lors de la recherche d'adresse a été renvoyées ou le jeune a été retrouvé dans l'annuaire) et les adresses « douteuses » (pas de coupon-réponse, pas de succès avec l'annuaire, mais le courrier n'a pas été renvoyé en NPAI). On tient compte de la différence importante de taux de réponse entre ces deux cas, en ajoutant cette distinction dans la modèle.

[6] On a accordé le « bénéfice du doute » à quelques variables sont la « significativité » était comprise entre 5 % et 10 %.

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
1,49	1,18	1,03	1,05	1,07	1,09	1,16	1,38	1,97	7,02	39,27

Les différences selon l'échantillon sont bien sûr importantes :

	Secondaire	Supérieur	Insee
Pondération	1,74	1,14	2,21

On compte 33 personnes, toutes dans l'échantillon Insee, qui se voient affecter une pondération supérieure à 10. Ces personnes cumulent un certain nombre de caractéristiques associées positivement avec la non-réponse (mais ont elles-mêmes répondu). En particulier, elles ont souvent changé d'établissement (76 % contre 32 % en général), ont été perdu par la DEPP au cours de leur scolarité dans le secondaire (52 % contre 18 % en général) et n'ont pas répondu à l'enquête famille (64 % des cas contre 21 %) et à l'enquête Jeune (97 % des cas contre 46 % en général). Lors des premières études, il faudra étudier l'influence des ces personnes sur les résultats.

Calage sur des données externes

La DEPP dispose d'un certain nombre de données sur les élèves entrées en 6e en 1995, qu'il peut être intéressant de confronter avec l'échantillon des répondants[7]. On n'a pu retenir la profession du responsable de l'élève en 1995, parmi les critères de calage, car les concepts de l'enquête différait trop fortement des sources disponibles (en particulier, les inactifs ayant déjà travaillé sont dans le panel recodés avec la CS de leur dernier emploi, ce qui n'est pas le cas dans les fichiers administratifs). On a donc calé uniquement sur le croisement du secteur, du sexe et de l'âge d'entrée en sixième (le croisement par sexe et âge n'étant disponible que pour le secteur public). Voici les répartitions, avant et après calage (par règle de trois sur les poids) :

	Répartition dans l'échantillon		Réparition dans la population	
Secteur Privé	3442	19,4%	146602	19,4%
Secteur Public				
Garçons en avance	193	1,1%	8218	1,1%
Garçons à l'heure	5178	29,1%	215848	28,6%
Garçons en retard	1920	10,8%	84927	11,3%
Filles en avance	217	1,2%	10423	1,4%
Filles à l'heure	5365	30,2%	224779	29,8%
Filles en retard	1459	8,2%	63128	8,4%

L'échantillon avant calage diffère peu des marges disponibles. Le recalage semble cependant utile, en particulier pour les garçons âgés, sans doute parce qu'ils ont un taux de réponse plus faible (ils sont en effet sous-représentés dans l'échantillon du supérieur, ayant le plus fort taux de réponse ; le modèle de non-réponse ne permet pas de corriger ce léger décalage).

La DEPP a aussi fournit des résultats sur les bacheliers 2002, qu'il peut être intéressant de confronter avec les données. Voici un extrait de ces résultats :

[7] Pour avoir une population comparable, on a rajouté les élèves hors champ (décédés ou partis à l'étranger), en leur attribuant un poids de 1.

	garçon	fille	total
17	6523	9916	16439
18	77828	118924	196752
19	59082	64491	123573
20	44810	40991	85801

On peut établir un lien entre :

- les bacheliers 2002 de 17 ans (ayant eu leur bac avec un an d'avance) et les élèves entrées en 1995 en sixième à 10 ans et ayant eu leur bac en 2002, c'est-à-dire sans redoubler (15 717)[8].

- les bacheliers 2002 de 18 ans (ayant eu leur bac à l'âge normal) et les élèves entrées en 1995 en sixième à 11 ans et ayant eu leur bac en 2002, c'est-à-dire sans redoubler (187 997). Parmi les cas qui posent problème, il y a les élèves entrés en avance en sixième en **1994** et qui ont redoublé une fois et eu leur bac en 2002 : ils se trouvent dans la case 18 du tableaux ci-dessus. On peut avoir une idée de leur nombre en regardant dans nos données le nombre d'élèves entrés en avance en sixième en 1995 et qui ont redoublé une fois pour avoir leur bac en **2003** (4 088)

- les bacheliers 2002 de 19 ans (ayant eu leur bac avec un an de retard) et les élèves entrées en 1995 en sixième à 12 ans et ayant eu leur bac en 2002, c'est-à-dire sans redoubler (6 678). On y ajoutera les élèves entrés à 11 ans ayant redoublé une fois pour avoir leur bac en **2003** (115 369) ; les élèves entrés à 10 ans ayant redoublé deux fois pour avoir leur bac en **2004** (2 125).

- les bacheliers 2002 de 20 ans (ayant eu leur bac avec deux ans de retard) et les élèves entrées en 1995 en sixième à 13 ans et ayant eu leur bac en 2002, c'est-à-dire sans redoubler (549). On y ajoutera les élèves entrés à 12 ans ayant redoublé une fois pour avoir leur bac en **2003** (12 602) ; les élèves entrés à 11 ans ayant redoublé deux fois pour avoir leur bac en **2004** (69 526).

On obtient alors le tableau suivant :

	garçon	fille	total
17	5999	9718	15717
18	76541	115544	192085
19	60619	63553	124172
20	47556	35121	82677

Les écarts entre les deux tableaux sont relativement faibles.

[8] Les cas posant problèmes, comme les élèves à l'heure en 1995 et ayant sauté une classe au collège, doivent être rares

Annexe 1 : Liste des variables utilisées

GARS	Garçons
NBENF1	Pas de frère et sœur
NBENF23	1 ou 2 frères et sœurs
NBENF4	3 frères et sœurs ou plus
PAYMERE200	Mère née à l'étranger
PAYMERE100	Mère née en France
PAYMERE999	Pays de naissance de la mère inconnue
ENTOUR95	Vivait avec son père et sa mère en 1995
PCSP1	Père agriculteur
PCSP34	Père cadre ou de profession intermédiaire
ACTIP2	Père chômeur
ENTREL89	Entrée à l'école en 1989
ENTREL90	Entrée à l'école en 1990
ENTREL99	Année d'entrée à l'école inconnue
ZEP	En ZEP en 1995
DEMIP	Demi-pensionnaire en 1995
PETITU	Collège rural ou de petites villes en 1995
SSECTEUR	Est passé par le privé durant sa scolarité
BOURSIER	Boursier
REPFAM1	A répondu (par écrit) à l'enquête Famille
REPFAM2	A répondu (par téléphone) à l'enquête Famille
REPFAM3	N'a pas répondu à l'enquête Famille
REPJ1	A répondu (par écrit) à l'enquête Jeune
REPJ2	A répondu (par téléphone) à l'enquête Jeune
REPJ3	N'a pas répondu à l'enquête Jeune
REPJ4	N'a pas été interrogé par l'enquête Jeune
R_FRANEL0	Premier tertile de score en Français (6e)
R_FRANEL9	Score en français inconnu
R_FROR2	Troisième tertile d'appréciation en français oral
R_MATHEL2	Troisième tertile de score en math (6e)
R_MATHEL9	Score en math inconnu
CHANGET	A changé d'établissement au collège
BAC2002	A obtenu le bac en 2002
BAC2003	A obtenu le bac en 2003
BAC2004	A obtenu le bac en 2004
MENT	A eu une mention au bac
LASTC1	Dernière classe connue : avant la terminale
LASTC2	Dernière classe connue : terminale générale
LASTC3	Dernière classe connue : terminale technologique
LASTC4	Dernière classe connue : terminale professionnelle
RSORTIE001	Raison de la sortie du secondaire : vie active
RSORTIE002	Raison de la sortie du secondaire : chômage
RSORTIE003	Raison de la sortie du secondaire : non scolarisé
RSORTIE020	Raison de la sortie du secondaire : perdus
RSORTIE950	Raison de la sortie du secondaire : supérieur
SORTANTS1	Sortis de l'échantillon DEP
SORTANTS2	Perdus
SORTANTS3	Bacheliers non-répondants à l'enquête SUP
ANNEE1	Sortis avant 1999
ANNEE2	Sortis en 1999 ou 2000
ANNEE3	Sortis en 2000 ou 2001
ADRESURE	Adresse sûre

Annexe 2 : non-réponse dans le « secondaire »

Ordered Value	nrsec	Total Frequency
1	1	689
2	0	940

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	227.1586	11	<.0001
Score	218.1863	11	<.0001
Wald	193.6902	11	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.4027	0.2765	2.1208	0.1453
gars	1	0.2008	0.1122	3.2024	0.0735
nbenf4	1	0.2770	0.1458	3.6106	0.0574
R_FROR2	1	-0.2984	0.1458	4.1857	0.0408
ssecteur	1	0.5340	0.2124	6.3217	0.0119
bac2004	1	-0.4632	0.1306	12.5854	0.0004
boursier	1	0.2891	0.1360	4.5188	0.0335
petitu	1	-0.3554	0.1132	9.8587	0.0017
repj1	1	-1.2201	0.1366	79.8231	<.0001
repj2	1	-0.3206	0.1765	3.2975	0.0694
repfam1	1	-0.5288	0.1500	12.4292	0.0004
repfam3	1	0.2039	0.2097	0.9453	0.3309

Association of Predicted Probabilities and Observed Responses

Percent Concordant	70.6	Somers' D	0.425
Percent Discordant	28.1	Gamma	0.431
Percent Tied	1.3	Tau-a	0.208
Pairs	647660	c	0.713

Annexe 3 : non-réponse dans le « supérieur »

Ordered Value	nrsup	Total Frequency
1	1	1054
2	0	7522

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	393.0718	17	<.0001
Score	455.3514	17	<.0001
Wald	398.2423	17	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.2996	0.1715	3.0520	0.0806
gars	1	0.2683	0.0699	14.7332	0.0001
paymere200	1	0.2640	0.0972	7.3798	0.0066
paymere999	1	-0.3661	0.2355	2.4182	0.1199
nbenf1	1	0.2296	0.1034	4.9304	0.0264
nbenf4	1	0.2760	0.0957	8.3080	0.0039
entour95	1	-0.5104	0.0903	31.9633	<.0001
R_mathel2	1	-0.2929	0.0773	14.3502	0.0002
R_mathel9	1	0.1115	0.1650	0.4567	0.4992
changet	1	0.2807	0.0970	8.3710	0.0038
bac2003	1	-0.6311	0.0900	49.1539	<.0001
bac2004	1	-0.7603	0.1171	42.1306	<.0001
ment	1	-0.4310	0.1020	17.8399	<.0001
lastc4	1	0.2822	0.1111	6.4493	0.0111
repfam1	1	-0.1052	0.1121	0.8812	0.3479
repfam3	1	0.5765	0.1403	16.8796	<.0001
repj1	1	-1.1668	0.0975	143.1791	<.0001
repj2	1	-0.7355	0.1332	30.4876	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	66.7	Somers' D	0.349
Percent Discordant	31.8	Gamma	0.354
Percent Tied	1.5	Tau-a	0.075
Pairs	7928188	c	0.674

Annexe 4 : absence d'adresse pour l'Insee

Ordered Value	nreva1	Total Frequency
1	1	2652
2	0	4604

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	943.5246	15	<.0001
Score	915.6785	15	<.0001
Wald	810.6338	15	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.0641	0.1684	0.1450	0.7034
entour95	1	-0.3765	0.0584	41.6139	<.0001
pcsp1	1	-0.6754	0.2035	11.0139	0.0009
R_franel0	1	-0.2076	0.0577	12.9267	0.0003
R_franel9	1	-0.00637	0.1164	0.0030	0.9564
changet	1	0.2229	0.0571	15.2630	<.0001
rsortie001	1	-0.4893	0.0899	29.6256	<.0001
rsortie002	1	-0.5754	0.1225	22.0685	<.0001
rsortie003	1	-0.3404	0.1087	9.7987	0.0017
rsortie020	1	1.0004	0.1222	66.9735	<.0001
sortants3	1	-0.4447	0.1093	16.5569	<.0001
annee1	1	-0.1126	0.1022	1.2142	0.2705
annee3	1	-0.8710	0.0728	143.1771	<.0001
repj1	1	0.0621	0.1463	0.1800	0.6713
repj2	1	0.00612	0.1595	0.0015	0.9694
repj3	1	0.5851	0.1433	16.6736	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	70.4	Somers' D	0.419
Percent Discordant	28.5	Gamma	0.423
Percent Tied	1.0	Tau-a	0.194
Pairs	12209808	c	0.709

Annexe 5 : non-réponse pour l'Insee

Ordered Value	nreva2	Total Frequency
1	1	1351
2	0	3253

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2526.2721	11	<.0001
Score	2237.0770	11	<.0001
Wald	1201.2152	11	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.3251	0.3242	1.0054	0.3160
paypere200	1	0.3922	0.1208	10.5333	0.0012
paypere999	1	0.0870	0.1476	0.3476	0.5555
demip	1	0.1820	0.0946	3.7035	0.0543
changet	1	0.2370	0.1008	5.5305	0.0187
rsortie020	1	0.5768	0.2498	5.3310	0.0209
adresure	1	-3.9661	0.1236	1028.9741	<.0001
repfam1	1	-0.5307	0.1345	15.5668	<.0001
repfam3	1	0.0292	0.1537	0.0361	0.8494
repj1	1	-0.0669	0.3053	0.0480	0.8265
repj2	1	0.4844	0.3272	2.1921	0.1387
repj3	1	0.7737	0.3013	6.5938	0.0102

Association of Predicted Probabilities and Observed Responses

Percent Concordant	90.1	Somers' D	0.811
Percent Discordant	9.0	Gamma	0.818
Percent Tied	0.8	Tau-a	0.336
Pairs	4394803	c	0.905

Direction des Statistiques Démographiques et Sociales
Département de l'Emploi et des Revenus d'Activités
Division Emploi

NOTE

Dossier suivi par :
Fabrice Murat
Tél. : 01 41 17 54 52
Fax : 01 41 17 61 63
Mél : Fabrice Murat

Paris, le
N° / DG75-F230 /

Objet : Calcul d'une pondération globale en 2006 pour le panel d'élèves entrés en 6e en 1995

Cette note présente une première proposition de calcul d'une pondération en 2006, s'inspirant de ce qui a été fait pour les données de 2005, en détaillant une procédure qui puisse être réutilisée assez simplement pour les années à venir.

Fusion des fichiers

Dans un premier temps, comme en 2005, il faut apparier les fichiers provenant des trois équipes : DEPP-Secondaire, DEPP-Supérieur et Insee. Outre le fichier de 2005 (incluant les données historiques et celles recueillies en 2004-2005), on dispose en fait des fichiers indiquant :

- les bacheliers interrogés par l'Insee en 2005, répondants, envoyés à la DEPP et éventuellement retournés à l'Insee quand il n'entrait pas dans le champ du suivi dans l'enseignement supérieur[1]
- les jeunes interrogés en 2006 par l'Insee (*EVA06.txt*), avec un statut de réponse indiquant s'ils ont répondu ou non (6 385 individus)
- les jeunes interrogés en 2006 par l'équipe DEPP Supérieur (*Insee_base06*), avec un statut de réponse (8 058 individus)
- les répondants à l'enquête DEPP-Secondaire 2006 (*Troncom95_06*)[2].

On a alors apparié ces fichiers entre eux et avec celui qui a été construit pour l'enquête de 2005, en prenant les 17 830 jeunes de l'échantillon initial du panel.

Un individu apparaît en 2006 que l'on ne retrouve pas parmi les 17 830 jeunes du départ. Son identifiant est `MITCHAN830601`. En fait, il s'agit de `MICHTAN830601`. L'identifiant a été corrigé.

En comparant les interrogations par la DEPP et par l'Insee, 98 individus semblent interrogés à la fois par l'Insee et par la DEPP. Certains de ces doublons avaient été déjà été remarqués lors du travail sur les données 2005. D'autres sont nouveaux et proviennent de problèmes de transmission entre les différentes équipes pour 2006 :

[1] Il existe deux fichiers : le premier (*Bacheliers*) qui avait été envoyé comporte 535 individus ; le second (*Envdep.txt*) en comporte 483 dont 90 retournés par la DEPP.
[2] Dans ce cas, on n'a pas l'échantillon initial, mais seulement la liste de ceux qui ont répondu.

- par exemple, la situation de certains jeunes encore scolarisés dans le secondaire en 2005 s'est parfois stabilisée tardivement : ils ont d'abord semblé sortis sans le bac, d'où une transmission par l'équipe DEPP-Secondaire à l'Insee ; puis, on s'est aperçu qu'ils avaient eu le bac, d'où une transmission par l'équipe DEPP-Secondaire à l'équipe DEPP-Supérieur
- l'Insee a retrouvé en 2005 des jeunes qui étaient encore en études initiales ; il les a renvoyés à la DEPP pour réintégration éventuelle dans le suivi scolaire ; malheureusement, certains jeunes ont été ajoutés à l'échantillon DEPP-Supérieur sans être enlevés de l'échantillon Insee.

On propose les traitements suivants :

- Pour les jeunes ayant répondu à la fois à l'équipe DEPP Secondaire et à l'Insee, on privilégie la réponse à l'Insee (car elle donne plus d'informations) : cela regroupe 4 cas, dont 3 bacheliers 2004 transmis par l'équipe DEPP-Supérieur et 1 autre qui n'a pas le bac…
- Pour les jeunes ayant répondu à la fois à l'équipe DEPP Supérieur et à l'Insee, on privilégie la réponse à l'Insee (car elle donne pour l'instant plus d'informations ; quand les réponses aux questions spécifiques au questionnaire dans l'enseignement supérieur seront connues, on les réintègrera dans les données) : cela regroupe 47 cas (pour moitié, il s'agit de « renvoyés » de l'Insee vers la DEPP)
- Pour les individus interrogés par l'Insee et l'équipe DEPP-Supérieur et n'ayant répondu qu'à l'un des deux, on attribue le jeune à l'équipe qui a eu la réponse : dans 20 cas à l'Insee, dans 10 cas à la DEPP
- Quand le jeune n'a répondu à personne, on attribue le jeune à l'équipe qui l'a interrogé en 2005 : dans 13 cas sur 17, c'est l'Insee car il s'agit de « renvoyés » de l'Insee vers la DEPP.

Compte tenu de ces doublons, on obtient ainsi un total de 14 654 jeunes interrogés.

Ainsi, 3 176 jeunes de l'échantillon initial ne semblent pas avoir été interrogés en 2006. Certains cas peuvent s'expliquer grâce au travail effectué pour le fichier de 2005 :

- 369 morts, gravement malades, partis à l'étranger
- 370 problèmes de transmission déjà repérés en 2005. Il s'agit des cas 03, 08, 10 et 11 (les choses n'étaient pas claires pour le cas 08, mais à l'évidence ils n'ont pas non plus été interrogés en 2006). Notons que 4 jeunes relevant des cas ci-dessus ont été tout de même interrogés en 2006.
- 2 278 jeunes que l'Insee a éliminés définitivement de l'échantillon, car il s'agissait de NPAI ou de réponses inexploitables, lors de la recherche d'adresse ou lors de la collecte. Notons que 25 cas relevant de ces catégories ont cependant été interrogés (parfois l'équipe DEPP-Secondaire) ; certains ont même répondu.

Il reste donc 159 personnes non interrogées sans explication à partir des données de 2005, qui se répartissent de la façon suivante :

- 8 sont décédés d'après un fichier envoyé par la DEPP pour préparer l'enquête de 2007
- 149 ont été interrogés par la DEPP dans l'enseignement secondaire en 2005
- 2 ont été interrogés par la DEPP dans l'enseignement supérieur en 2005.

On va supposer que les 149 individus sans statut de réponse ayant été interrogés dans l'enseignement secondaire en 2005, sont des non-répondants à l'enquête dans l'enseignement secondaire en 2006 (on ne dispose pas de l'échantillon initial en 2006 permettant de valider cette hypothèse). Les autres cas de non-interrogation sont considérés comme normaux et exclus de l'analyse de la non-réponse.

Au total, l'échantillon initial se décompose en :

- 377 jeunes morts, gravement malades, partis à l'étranger
- 2 280 définitivement perdus
- 370 problèmes de transmission (problème normalement résolu en 2007)
- 455 jeunes interrogés par l'équipe DEPP secondaire, dont 305 répondants (66,7 %)

- 7 977 jeunes interrogés par l'équipe DEPP supérieur, dont 7 239 répondants (90,7 %)
- 6 371 jeunes interrogés par l'Insee, dont 3 946 répondants (61,9 %).

La confrontation avec le statut de réponse en 2005 est aussi instructive :

		Statut de réponse en 2006	
		Non-répondant	Répondant
Statut de réponse en 2005	Non répondant	4868	1247
	Répondant	1472	10243

Dans l'ensemble, il y a une assez bonne concordance entre les deux statuts, mais 15 % de la population a changé de statut, de façon à peu près équilibrée dans les deux sens :

- 1472 jeunes ont répondu en 2005, mais ne répondent plus en 2006 : dans la plupart des cas, ils ont été interrogés par la même équipe aux deux dates. Il semble donc s'agir d'attrition « naturelle » liée à la démotivation des participants. Ces jeunes ont été pour la plupart relancés en 2007 (en particulier, tous les non-répondants aux enquêtes DEPP ont été normalement transmis à l'Insee)
- 1247 n'ont pas répondu en 2005, mais répondent en 2006 : dans quelques cas, il s'agit de jeunes que l'on avait oublié d'interroger en 2005 et qu'on a récupérés en 2006 ; ils peuvent aussi avoir été relancés par l'Insee après un refus de répondre en 2005 ; ce sont aussi très souvent des non-répondants aux enquêtes DEPP transmis à l'Insee.

Pondération

Comme en 2005, on va procéder en deux temps (voir l'annexe 1 décrivant en détail la procédure utilisée). D'abord, on corrige la non-réponse par une série de régressions logistiques :

- un modèle pour l'échantillon « Secondaire »
- un modèle pour l'échantillon « Supérieur »
- deux modèles pour l'échantillon « Insee » : l'un modélisant la perte d'adresse[3] ; l'autre, le refus de répondre quand l'adresse a été retrouvée.

Ensuite, on procède à un calage sur données externes selon le sexe et l'âge d'entrée en sixième.

On a procédé à modélisation de la non-réponse dans les trois enquêtes, en tenant compte de différentes caractéristiques disponibles :

- les caractéristiques socio-démographiques de l'élève en 1995 (pays de naissance, profession des parents, type de ménage en 1995, etc.) ;

- les caractéristiques scolaires de l'élève à la rentrée 1995 (retard scolaire, résultats aux évaluations 6e, etc.) ;

- les informations relatives à l'année scolaire 1995-1996 (secteur de l'établissement, appartenance à une ZEP, etc.)

- quelques informations sur le parcours scolaire (orientation en 3e, optention du bac, mention éventuelle)

- le « comportement de réponse » (raison et année de la sortie de l'échantillon DEPP, réponse à l'enquête famille ou à l'enquête Jeune, résultats de la recherche

[3] Les pertes d'adresses sont seulement celles lors de la recherche d'adresse et de la collecte par l'Insee en 2005. On supposera que toutes les non-réponses à l'Insee pour des jeunes venant de la DEPP en 2006 sont des refus de répondre (hypothèse forte car leur non-réponse à la DEPP l'année précédent peut être le signe d'un déménagement).

d'adresse ; l'équipe interrogatrice en 2005 ; ces variables ont surtout une importance pour l'échantillon Insee).

On procède alors à quatre modélisations sur cet échantillon.

- la non réponse dans l'échantillon « secondaire », donnant la probabilité P_{sec} de ne pas répondre à cette enquête ;

- la non réponse dans l'échantillon « supérieur », donnant la probabilité P_{sup} de ne pas répondre à cette enquête ;

- sur l'échantillon des jeunes interrogés par l'Insee qui étaient parmi les jeunes transmis en 2005, on modélise la probabilité P_{eva1} de ne pas avoir avoir d'adresse[4] ;

- sur l'ensemble des adresses retrouvées de l'Insee[5], on modélise la probabilité P_{eva2} de ne pas répondre à l'enquête.

Dans ce cas, la pondération POIDS est égale à :

- l'inverse de (1- P_{sec}) pour l'échantillon « secondaire »

- l'inverse de (1- P_{sup}) pour l'échantillon « supérieur »

- l'inverse de (1- P_{eva1})* (1- P_{eva2}) pour l'échantillon Insee[6]

Les résultats de ces modélisations sont donnés en annexes. On a généralement retenu les variables significatives au seuil de 5 %[7].

La pondération ainsi obtenu a les caractéristiques suivantes :

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
1,52	1,12	1,02	1,04	1,04	1,06	1,11	1,52	2,23	6,09	27,62

Les différences selon l'échantillon sont bien sûr importantes :

	Secondaire	Supérieur	Insee
Pondération	1,49	1,10	2,28

Calage sur des données externes

La DEPP dispose d'un certain nombre de données sur les élèves entrées en 6e en 1995, qu'il peut être intéressant de confronter avec l'échantillon des répondants[8]. On n'a pu retenir la profession du responsable de l'élève en 1995, parmi les critères de calage, car les concepts de l'enquête différait trop fortement des sources disponibles (en particulier, les inactifs ayant

[4] Cela inclut : les problèmes de transmission entre la DEPP et l'Insee et les cas où l'adresse du jeune est inconnue, que ce manque soit apparu lors de la recherche d'adresse ou lors de la collecte.
[5] Donc à la fois les jeunes retrouvés par l'Insee en 2005 et ceux transmis par la DEPP en 2006.
[6] Pour les jeunes transmis à l'Insee par la DEPP en 2006, P_{eva1}=0.
[7] On a accordé le « bénéfice du doute » à quelques variables sont la « significativité » était comprise entre 5 % et 10 %.
[8] Pour avoir une population comparable, on a rajouté les élèves hors champ (décédés ou partis à l'étranger), en leur attribuant un poids de 1.

déjà travaillé sont dans le panel recodés avec la CS de leur dernier emploi, ce qui n'est pas le cas dans les fichiers administratifs). On a donc calé uniquement sur le croisement du secteur, du sexe et de l'âge d'entrée en sixième (le croisement par sexe et âge n'étant disponible que pour le secteur public). Voici les répartitions, avant et après calage (par règle de trois sur les poids) :

	Répartition dans l'échantillon		Réparition dans la population	
Secteur Privé	3415	19,2	146602	19,4%
Secteur Public				
Garçons en avance	220	1,2	8218	1,1%
Garçons à l'heure	5267	29,6	215848	28,6%
Garçons en retard	1850	10,4	84927	11,3%
Filles en avance	213	1,2	10423	1,4%
Filles à l'heure	5372	30,2	224779	29,8%
Filles en retard	1473	8,3	63128	8,4%

L'échantillon avant calage diffère peu des marges disponibles. Le recalage semble cependant utile, en particulier pour les garçons âgés, sans doute parce qu'ils ont un taux de réponse plus faible (ils sont en effet sous-représentés dans l'échantillon du supérieur, ayant le plus fort taux de réponse ; le modèle de non-réponse ne permet pas de corriger ce léger décalage).

On stocke le poids dans la table *pond.sas7bdat*.

Annexe 1 : procédure de calcul de la pondération

On va décrire ici en détail la procédure de calcul de la pondération, une fois que les fichiers ont été appariés et que chaque individu s'est vu attribué un statut de réponse (les valeurs à blanc doivent correspondre aux cas où le jeune n'avait pas à être interrogé : morts ou interrogés par une autre équipe).

Dans un premier temps, on considère les variables ci-dessous, issues du fichier historique après quelques regroupements de modalité.

Variables utilisés pour l'échantillon secondaire

SEXE	**Sexe**
1	Garçons
2	Filles
LIEUNAI	**Lieu de naissance de l'enfant**
1	France
3	Etranger
NATELEVE	**Nationalité de l'enfant**
100	Française
200	Autre
NATPERE	**Nationalité du père**
0	Père inconnu ou nationalité inconnue
100	Française
200	Autre
PAYPERE	**Pays de naissance du père**
0	Père inconnu ou pays de naissance inconnu
100	France
200	Etranger
NATMERE	**Nationalité de la mère**
0	Mère inconnue ou nationalité inconnue
100	Française
200	Autre
PAYMERE	**Pays de naissance de la mère**
0	Mère inconnue ou pays de naissance inconnu
100	France
200	Etranger
NBENF	**Nombre d'enfants dans la famille**
1	1
2	2
3	3
4	4 ou plus
RANG	**Rang dans la fratrie**
1	1er
2	2ème
3	3ème
4	4ème ou plus
Permer	**Vivait avec son père et sa mère**
0	Non

1	Oui
ACTIPERE	**Activité du père**
0	Père inconnu
1	Actif
2	Inactif
pcsp	**CS du père**
1	Agriculteur
2	Artisan commerçant
3	Cadre supérieur
4	Profession intermédiaire
5	Employé
6	Ouvrier
8	Inactif
9	Père inconnu ou profession inconnue
ACTIMERE	**Activité de la mère**
0	Mère inconnue
1	Active
2	Inactive
pcsm	**CS de la mère**
1	Agriculteur
2	Artisan commerçant
3	Cadre supérieur
4	Profession intermédiaire
5	Employé
6	Ouvrier
8	Inactif
9	Mère inconnue ou profession inconnue
quartot	**quartile de réussite aux évaluations 6e**
0	Premier quartile
1	Deuxième quartile
2	Troisième quartile
3	Quatrième quartile
ENTREL	**Age d'entrée à l'école élementaire**
0	Age inconnu
89	En 1989 ou avant
90	En 1990 ou après
AGE6E	**Age d'entrée en sixième**
10	10 ou moins
11	11 ans
12	12 ans ou plus
SECTECO	**Secteur de l'école primaire**
1	Public
2	Privé
ZEP1995	**Présence en Zep en 1995**
1	Oui
2	Non
HEBERG1995	**Hébergement en 1995**
1	Externe
2	Demi-pensionnaire, interne, autre
TUETAB1995	**Tranche d'unité urbaine de l'établissement de 1995+B105**
0	Commune rurale

1	Commune urbaine de moins de 5 000 habitants
2	Commune urbaine de 5 000 a moins de 10 000 habitants
3	Commune urbaine de 10 000 a moins de 20 000 habitants
4	Commune urbaine de 20 000 a moins de 50 0000 habitants
5	Commune urbaine de 50 000 a moins de 100 000 habitants
6	Commune urbaine de 100 000 a moins de 200 000 habitants
7	Commune urbaine de 200 000 a moins de 2 000 000 habitants
8	Agglomération parisienne
SECTEUR1995	**Secteur de l'établissement de 1995**
1	Public
2	Privé
changet	**Changement d'établissement durant la scolarité au collège**
0	Non
1	Oui
boursier	**Boursier au collège**
0	Non
1	Oui
repfam	**Réponse à l'enquête Famille**
1	Réponse postale
2	Réponse téléphonique
3	Non interrogé ou non répondant
repjeune	**Réponse à l'enquête Jeune**
1	Réponse postale
2	Réponse téléphonique
3	Non interrogé ou non répondant

Pour les jeunes dans l'enseignement supérieur, on ajoute les variables suivantes :

bachelier	**Année**
0000	N'a pas le bac
2002	En 2002
2003	En 2003
2004	En 2004
mention	**Mention au bac**
0-no	N'a pas le bac
0	Pas de mention
1-AB	Assez bien
2- B	Bien
3-TB	Très bien
Fichier2004	**Equipe d'interrogation en 2004**
EVA	Interrogé par l'Insee
SEC	Interrogé par l'équipe DEPP-Secondaire
SUP	Interrogé par l'équipe DEPP-Supérieur

Pour les jeunes enquêtés par l'Insee, on rajoute encore :

rsortie	Raison de la sortie du panel scolaire
0	Pas sorti
1	Vie active
2	Chômage
3	Non scolarisé
5	Décès
6	Abandon de scolarité pour raison de santé
20	Autre
950	Université
sortants	**Transmission à l'Insee**
1	Sortant avec le bac
2	Perdu avant le bac
3	Bachelier non répondant à l'enquête SUP*
ANNEE_SORTIE2004	**Année de sortie du panel scolaire**
0	Pas sorti
1995	1995
1996	1996
1997	1997
1998	1998
1999	1999
2000	2000
2001	2001
2002	2002
sortinsee	**Année de transmission à l'Insee**
2004	2004
2005	2005

* Il faudra sans doute plus tard distinguer les sortants du supérieur (ne faisant plus d'études) des non répondants.

Après un Proc Freq pour voir les effectifs des différentes modalités des variables sur le champ considéré, on étudie l'impact individuel de chaque variable sur la non-réponse, ce qui établit une première sélection. Ensuite, on construit un modèle incluant toutes les variables, avec l'option STEWISE qui sélectionne les plus pertinentes, au sens statistique du terme. La combinaison de ces deux sélections permet, après quelques regroupements de modalités, de déterminer les variables à conserver. On calcule alors les indicatrices correspondantes et on lance le modèle final, qui construira la table avec la probabilité de répondre associée à chaque répondant. C'est ce type de modèles pour chaque population, qui est présenté dans les annexes suivantes.

Pour 2006, les indicatrices ont été calculées de la façon suivante :

```
gars=(sexe='1');
lieunai1=(lieunai='1');
nateleve1=(nateleve='1');
natp000=(natpere='000');natp200=(natpere='200');
natm000=(natmere='000');natm200=(natmere='200');
payp000=(paypere='000');payp200=(paypere='200');
paym000=(paymere='000');paym200=(paymere='200');
nbenf12=(nbenf in ('1','2'));
actp0=(actipere='0');actp1=(actipere='1');
pcsp56=(pcsp in ('5','6'));
pcsp1=(pcsp='1');pcsp2=(pcsp='2');pcsp3=(pcsp='3');pcsp4=(pcsp='4');
pcsp5=(pcsp='5');pcsp6=(pcsp='6');pcsp9=(pcsp='9');pcsp8=(pcsp='8');
actm0=(actimere='0');actm1=(actimere='1');
pcsm1=(pcsm='1');pcsm2=(pcsm='2');pcsm3=(pcsm='3');pcsm4=(pcsm='4');
pcsm5=(pcsm='5');pcsm6=(pcsm='6');pcsm9=(pcsm='9');pcsm8=(pcsm='8');
quart1=(quartot='0');quart2=(quartot='1');quart3=(quartot='2');quart
4=(quartot='3');
repfam1=(repfam='1');repfam2=(repfam='2');repfam3=(repfam='3');
repjeune1=(repjeune='1');repjeune2=(repjeune='2');repjeune3=(repjeun
e='3');
ficeva=(fichier2004='EVA');ficsec=(fichier2004='SEC');
sortants1=(sortants='1');sortants2=(sortants='2');sortants3=(sortant
s='3');
annee0=(annee_sortie2004=0);
annee1=(1994<annee_sortie2004<1999);
annee2=(1998<annee_sortie2004<2001);
annee3=(2000<annee_sortie2004<2003);
rsortie001=(rsortie='001');rsortie002=(rsortie='002');rsortie003=(rs
ortie='003');
rsortie020=(rsortie='020');rsortie950=(rsortie='950');
prive=(secteur1995='2');
petitu=(tuetab1995 in ('0','1','2','3'));
heberg=(heberg1995='1');
```

Annexe 2 : non-réponse dans le « secondaire »

Ordered Value	nrsec	Total Frequency
1	1	150
2	0	305

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.2179	4	<.0001
Score	32.9738	4	<.0001
Wald	30.6238	4	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.5066	0.3484	2.1145	0.1459
pcsp56	1	-0.3304	0.2119	2.4307	0.1190
Permer	1	-0.5738	0.2489	5.3162	0.0211
repjeune1	1	0.1850	0.3114	0.3528	0.5525
repjeune3	1	1.2342	0.3503	12.4157	0.0004

Association of Predicted Probabilities and Observed Responses

Percent Concordant	57.9	Somers' D	0.301
Percent Discordant	27.8	Gamma	0.352
Percent Tied	14.4	Tau-a	0.133
Pairs	45750	c	0.651

Annexe 3 : non-réponse dans le « supérieur »

Ordered Value	nrsup	Total Frequency
1	1	742
2	0	7239

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	439.6531	14	<.0001
Score	603.1028	14	<.0001
Wald	458.4537	14	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.2944	0.2802	21.3378	<.0001
Permer	1	-0.4812	0.1050	21.0043	<.0001
actm0	1	0.2179	0.3123	0.4870	0.4853
actm1	1	-0.1992	0.0858	5.3952	0.0202
quart1	1	0.4627	0.1405	10.8423	0.0010
quart2	1	0.2652	0.1121	5.5960	0.0180
quart3	1	0.2577	0.1004	6.5820	0.0103
changet	1	0.5867	0.1050	31.2074	<.0001
repfam1	1	-0.4584	0.0947	23.4455	<.0001
repjeune1	1	-0.2688	0.1292	4.3289	0.0375
repjeune3	1	0.4726	0.1589	8.8415	0.0029
ficeva	1	1.8515	0.1176	247.9235	<.0001
ficsec	1	0.0637	0.1801	0.1250	0.7237
lieunai1	1	-0.4026	0.2219	3.2929	0.0696
pcsp8	1	-1.0385	0.5483	3.5867	0.0582

Association of Predicted Probabilities and Observed Responses

Percent Concordant	67.0	Somers' D	0.397
Percent Discordant	27.2	Gamma	0.422
Percent Tied	5.8	Tau-a	0.067
Pairs	5371338	c	0.699

Annexe 4 : absence d'adresse pour l'Insee

Ordered Value	nreva1	Total Frequency
1	1	2646
2	0	4212

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1324.7897	19	<.0001
Score	1248.8238	19	<.0001
Wald	1056.7310	19	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	0.7429	0.1319	31.7138	<.0001
Permer	1	-0.4180	0.0611	46.8242	<.0001
pcsp1	1	-0.4389	0.2489	3.1103	0.0778
pcsm1	1	-0.7549	0.3944	3.6622	0.0557
pcsm2	1	0.2209	0.1582	1.9495	0.1626
quart1	1	-0.1660	0.1049	2.5026	0.1137
quart2	1	-0.0366	0.1053	0.1210	0.7280
quart3	1	-0.0267	0.1097	0.0594	0.8075
prive	1	-0.1399	0.0775	3.2603	0.0710
changet	1	0.2893	0.0598	23.4176	<.0001
repjeune1	1	-0.1051	0.0866	1.4725	0.2250
repjeune3	1	0.4712	0.0825	32.6590	<.0001
sortants1	1	-2.2686	0.1382	269.6551	<.0001
sortants2	1	-1.3920	0.1354	105.6976	<.0001
rsortie001	1	0.7105	0.1151	38.1285	<.0001
rsortie002	1	0.7206	0.1415	25.9351	<.0001
rsortie003	1	0.9306	0.1304	50.9018	<.0001
rsortie020	1	1.2593	0.1209	108.5004	<.0001
annee1	1	0.2266	0.0968	5.4822	0.0192
annee3	1	-0.1058	0.0760	1.9371	0.1640

Association of Predicted Probabilities and Observed Responses

Percent Concordant	74.5	Somers' D	0.495
Percent Discordant	25.0	Gamma	0.498
Percent Tied	0.5	Tau-a	0.235
Pairs	11144952	c	0.748

Annexe 5 : non-réponse pour l'Insee

Ordered Value	nreva2	Total Frequency
1	1	2425
2	0	3946

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	737.8152	18	<.0001
Score	715.1897	18	<.0001
Wald	639.2524	18	<.0001

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.0417	0.2147	0.0377	0.8460
gars	1	0.1137	0.0567	4.0208	0.0449
Permer	1	-0.4194	0.0775	29.2692	<.0001
actp1	1	-0.2963	0.0952	9.6892	0.0019
actp0	1	-0.2018	0.1399	2.0813	0.1491
pcsm1	1	-1.0203	0.3127	10.6452	0.0011
quart1	1	0.3112	0.0608	26.1894	<.0001
heberg	1	0.1999	0.0577	11.9964	0.0005
petitu	1	-0.2536	0.0592	18.3707	<.0001
changet	1	0.1460	0.0635	5.2911	0.0214
repfam1	1	-0.2697	0.0782	11.9050	0.0006
repfam3	1	0.3752	0.0954	15.4694	<.0001
repjeune1	1	-0.1035	0.1605	0.4155	0.5192
repjeune2	1	-0.2821	0.1699	2.7566	0.0969
repjeune3	1	0.7916	0.1593	24.6794	<.0001
sortants1	1	-0.8446	0.1126	56.2498	<.0001
sortants2	1	-0.4412	0.1455	9.1904	0.0024
ficeva	1	0.4689	0.1182	15.7419	<.0001
ficsec	1	1.1115	0.1459	58.0456	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	69.4	Somers' D	0.393
Percent Discordant	30.1	Gamma	0.395
Percent Tied	0.5	Tau-a	0.185
Pairs	9569050	c	0.696

Calcul d'une pondération globale en 2006/2007 pour le panel d'élèves DEPP95 et le module SANTE
V3 - 17/02/2009

Rédacteur : *A. Degorre*

Introduction

Le panel DEPP95 fait l'objet d'un suivi annuel d'une cohorte d'élèves, sélectionnés lors de leur entrée en 6ème en 1995. La constitution du fichier statistique est effectuée en deux phases :

- De 1995-2002, le suivi des élèves est conduit à travers leurs études secondaires. Il s'est appuyé sur une collecte annuelle de données administratives et sur des enquêtes ponctuelles, dont l'enquête Famille de 1998 et l'enquête Jeunes de 2002. L'ingénierie statistique a été assurée par l'équipe DEPP du secondaire.
- A partir de 2002, le suivi des jeunes dépend de leur orientation : ceux qui poursuivent des études sont suivis par l'équipe DEPP du supérieur (enquête dite « SUP »), ceux qui achèvent un cursus dans le secondaire sont suivis par l'équipe DEPP du secondaire (enquête dite « SEC »), ceux qui ont terminé leurs études sont suivis par l'Insee (enquête dite « EVA » pour Entrée dans la Vie Adulte).

La collecte de données a été exposée, depuis 2002, à une perte d'informations sur une partie des jeunes de l'échantillon, pour diverses raisons :

- La « sortie » des jeunes du système éducatif ne permet pas de retrouver tous les jeunes, du fait des déménagements et des difficultés à obtenir les nouvelles coordonnées.
- Ce problème a été particulièrement prononcé pour les premiers sortants du suivi assuré par l'enquête « SEC ». Les jeunes concernés sont soit des « sortants précoces » du secondaire, c'est-à-dire des jeunes qui ont quitté leur cursus scolaire sans valider le diplôme qu'ils préparaient, soit de jeunes diplômés qui ne pousuivent pas d'études. L'incertitude sur leur parcours (poursuivent-ils des études ? se portent-ils sur le marché du travail ?) a retardé la transmission de leurs coordonnées auprès de l'Insee. De la sorte, une grande partie n'a pu être recontactée par l'Insee au sein de l'enquête EVA. En 2006, sur un échantillon initial de 17830 jeunes, 2280 jeunes ont de la sorte été « perdus » du suivi statistique.
- Une partie des jeunes sont sortis du champ du fait d'un départ à l'étranger, d'une maladie grave ou d'un décès (377 jeunes en 2006)
- La non-réponse d'une partie des jeunes, quand bien même ces derniers ont pu être contactés, conduit à une attrition régulière. Sont exclus du suivi les non-répondants récurrents, c'est-à-dire les jeunes ne répondant pas deux fois de suite au questionnaire EVA.

La perte d'information et l'attrition conduisent à une déformation de l'échantillon par rapport à l'échantillon initial, telle que présentée en **partie 1**. Pour limiter les biais induits, une pondération est calculée chaque année sur l'ensemble des jeunes qui ont répondu à l'un des questionnaires de suivi, SEC, SUP ou EVA. La **partie 2** présente une proposition de calcul d'une pondération en 2006/2007, s'inspirant de ce qui a été fait pour les données de 2004/2005 et les données de 2005/2006, en détaillant une procédure qui puisse être réutilisée assez simplement pour les années à venir.

La collecte 2006/2007 a été enrichie d'un module complémentaire, distinct du questionnaire usuel, portant spécifiquement sur la santé des personnes enquêtées. Ce module, élaboré en partenariat avec la Drees, a été transmis à l'ensemble des personnes suivies. Quoique très proches de ceux obtenus sur le questionnaire principal, les taux de non-réponse nécessitent de calculer une pondération ad hoc pour le module santé, selon une procédure décrite en **partie 3**.

PARTIE 1 - ATTRITION ET NON -REPONSE

Fusion des fichiers

Dans un premier temps, il faut apparier les fichiers provenant des diverses équipes. A la différence des années précédentes où une partie résiduelle du suivi des jeunes était conduit par la DEPP-Secondaire, à compter de 2006-2007 l'intégralité du suivi est réalisé par la DEPP-Supérieur et l'Insee. En effet, il ne restait déjà plus que 455 jeunes suivis par la DEPP-secondaire en 2005/2006, contre 1629 en 2004/2005 : lors de la collecte 2006/2007, l'intégralité de la population entrée en 6ème en 1995 est sortie du système éducatif du secondaire

Le fichier complet de 2006/2007 comprend donc :

Un échantillon de 7104 jeunes interrogés par l'Insee dans le cadre du questionnaire EVA :

- les jeunes interrogés par l'Insee en 2005 ou 2006, répondants au moins une fois à l'enquête, soit 4920 jeunes
- les jeunes interrogés par la DEPP-SUP en 2005 ou 2006 et qui sont supposés avoir fini depuis leur formation initiale, soit 1407 jeunes
- les jeunes interrogés par la DEPP-SEC en 2005 ou 2006 et qui sont supposés avoir fini depuis leur formation initiale, soit 396 jeunes
- les jeunes dont le suivi a fait l'objet d'un problème de transmission en 2006 et qui ont été affecté au suivi EVA, soit 367 jeunes
- quelques jeunes dont on a pu penser qu'ils étaient sortis du champ ou dont les adresses étaient erronées, mais qui ont pu être réintégrés au suivi EVA, soit 14 jeunes

Un échantillon de 6642 jeunes interrogés par la Depp dans le cadre du questionnaire SUP :

- les jeunes interrogés en 2006 par l'équipe DEPP Supérieur et qui sont supposés en cours de leur formation initiale, soit 6558 jeunes
- les jeunes interrogés par la DEPP-SEC en 2005 ou 2006 et qui sont supposés suivre une formation supérieure, soit 55 jeunes
- les jeunes interrogés par l'Insee en 2005 ou 2006, pour lesquels il apparaît toutefois qu'ils n'ont pas achevé leur formation initiale, soit 29 jeunes.

Un échantillon de 4084 jeunes sortis du suivi en 2006/2007, pour les motifs suivants :

- jeunes sortis du champ (maladie, décès, départ à l'étranger) : 370 jeunes
- les jeunes « perdus » (adresse inexploitable) lors de la sortie du suivi DEPP-SEC : 2268 jeunes
- les jeunes non-répondants de façon récurrente, soit 1446 jeunes

On a alors apparié ces fichiers entre eux, en reprenant les informations des 17 830 jeunes de l'échantillon initial du panel.

En comparant les interrogations par la DEPP et par l'Insee, 2 individus semblent interrogés à la fois par l'Insee et par la DEPP. Certains de ces doublons avaient été déjà été remarqués lors du travail sur les données 2005 et 2006 : ils sont désormais bien moins nombreux (98 cas en 2005/2006). Pour les jeunes ayant répondu à la fois à l'équipe DEPP Supérieur et à l'Insee, on privilégie la réponse à l'Insee (car elle donne pour l'instant plus d'informations ; quand les réponses aux questions spécifiques au questionnaire dans l'enseignement supérieur seront connues, on les réintègrera dans les données) : cela regroupe les 2 cas cités.

Analyse des taux de réponse

Quel que le soit le type de questionnaire analysé, les taux de réponse se sont nettement améliorés entre 2004/2005 et 2006/2007. Les taux de réponse ont été particulièrement bas lors de la première collecte EVA du fait d'un grand nombre de jeunes pour lesquels les coordonnées postales n'ont pas été exploitables. Par ailleurs, les non-répondants récurrents ont été exclus du champ de la collecte 2006/2007. Enfin, les nouveaux interrogés sont issus du suivi SUP, au cours duquel une relative « habitude » de réponse aux enquêtes postales a pu prendre place.

	2006/2007			2005/2006			2004/2005		
	Enquêtés	Répondants	Taux de réponse	Enquêtés	Répondants	Taux de réponse	Enquêtés	Répondants	Taux de réponse
Questionnaire EVA	7104	5090	72%	6384	3949	62%	7256	3253	45%
Questionnaire SUP	6642	6239	94%	7965	7238	91%	8576	7522	88%
Questionnaire SEC				459	302	66%	1629	940	58%

La confrontation avec le statut de réponse antérieur au sein de la collecte EVA est aussi instructif. On examine ici, parmi les personnes enquêtées l'année N dans le cadre d'EVA, le statut de réponse à l'enquête EVA l'année N+1

Collecte EVA 2005/2006 (4214 interrogés issus d'EVA 2004/2005)		Statut de réponse en 2005/2006	
		Non-répondant	Répondant
Statut de réponse en 2004/2005	Non répondant	1052 (25%)	302 (7%)
	Répondant	535 (13%)	2325 (55%)

Collecte EVA 2006/2007 (4920 interrogés issus d'EVA)		Statut de réponse en 2006/2007	
		Non-répondant	Répondant
Statut de réponse en 2005/2006	Non répondant	592 (12%)	403(8%)
	Répondant	724 (15%)	3201 (65%)

Rem : L'enquête EVA inclut également des jeunes transmis par l'équipe DEPP-SUP, qui sont ensuite « labellisés » l'année suivante comme jeunes suivis au sein de la collecte EVA.

Dans l'ensemble, on observe une « fidélisation » progressive de l'échantillon au questionnaire EVA : la part des répondants récurrents s'accroît (de 55% à 65%). Mécaniquement exclu lors de la collecte suivante, les non-répondants sont en effet de moins en moins nombreux (de 25% à 12%). Par contre, la proportion de non-répondants ponctuels reste stable : 7 à 8% pour les enquêtés ne répondant pas l'année N et répondant l'année N+1 ; 13 à 15% pour les enquêtés répondant l'année N mais pas en N+1.

PARTIE 2 - CALCUL DES PONDERATIONS EVA 2006/2007

Methodologie

Comme pour les précédentes collectes, on va procéder en deux temps (voir l'annexe 1 décrivant en détail la procédure utilisée). D'abord, on corrige la non-réponse par une série de régressions logistiques :

- un modèle pour l'échantillon « Insee »
- un modèle pour l'échantillon « Supérieur »

Ensuite, on procède à un calage sur données externes selon le sexe et l'âge d'entrée en sixième.

Les régressions logistiques mobilisées dans ces modèles prennent en compte différentes caractéristiques disponibles :

- les caractéristiques socio-démographiques de l'élève en 1995 (pays de naissance, profession des parents, type de ménage en 1995, etc.) ;

- les caractéristiques scolaires de l'élève à la rentrée 1995 (retard scolaire, résultats aux évaluations 6^{e}, etc.) ;

- les informations relatives à l'année scolaire 1995-1996 (secteur de l'établissement, appartenance à une ZEP, etc.)

- quelques informations sur le parcours scolaire (orientation en 3^{e}, obtention du bac, mention éventuelle)

- le « comportement de réponse » (raison et année de la sortie de l'échantillon DEPP, réponse à l'enquête famille ou à l'enquête Jeune).

Pour l'échantillon « Insee », on procède à deux modélisations :

- on cherche tout d'abord à modéliser la probabilité d'être inclus dans la collecte 2006-2007 par rapport au fait d'être « sorti » du suivi EVA. Il s'agit en particulier de prendre en compte le profil spécifique des jeunes qui ont été « perdus » lors de la première collecte EVA, du fait d'adresses postales inexploitables, ainsi que des jeunes qui n'ont jamais répondu au questionnaire (non-répondants récurrents). La probabilité d'être « sorti » de l'échantillon EVA est noté P_{HCH} ;
- Pour les jeunes « inclus » dans le suivi EVA, on modélise la non réponse, donnant la probabilité P_{eva} de ne pas répondre à cette enquête ;

Pour l'échantillon « Supérieur », on procède à une seule modélisation, celle de la non réponse au sein de l'échantillon, donnant la probabilité P_{sup} de ne pas répondre à cette enquête ;

Dans ce cas, la pondération POIDS est égale à :

- l'inverse de $(1 - P_{HCH})*(1- P_{eva})$ pour l'échantillon « Insee »

- l'inverse de $(1- P_{sup})$ pour l'échantillon « supérieur »

Les résultats de ces modélisations sont donnés en annexes. On a généralement retenu les variables significatives au seuil de 5 %.

La pondération ainsi obtenu a les caractéristiques suivantes :

Pondération EVA :

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
2,39	1,56	1,250	1,29	1,44	1,59	1,83	2,27	3,30	7,57	17,13

Pondération SUP :

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
1,06	0,03	1,02	1,02	1,05	1,05	1,05	1,07	1,10	1,18	1,37

La très faible non-réponse dans la collecte SUP (6%) conduit à un réajustement des poids marginal. Dans la collecte EVA, la prise en compte des « sortis » de la collecte (36% sur le champ potentiel de collecte de 11 188 jeunes en 2006/2007) puis une non-réponse plus prononcée (28% sur le champ effectif des 7104 jeunes enquêtés en 2006/2007) amèneà une déformation des poids plus notable. Il n'y a toutefois pas de concentration de pondération sur des cas isolés : le poids maximal est de « seulement » 17,13, avec un 9ème décile à 3,3.

Le détail des variables explicatives est donné en annexe. On notera l'importance jouée par des variables familiales dans le fait d'être inclus dans l'enquête ou d'y répondre : le fait que le jeune vivait, en 6ème, avec son père et sa mère réduit significativement la probabilité de sortie du panel ou de non-réponse à l'enquête EVA ou SUP de 2006/2007. Le comportement de réponse aux enquêtes conduites pendant le secondaire est également un bon prédicteur : les jeunes n'ayant pas répondu à l'enquête Famille ou à l'enquête Jeune sont moins nombreux à être inclus dans le suivi ou à répondre à la collecte en cours. Le niveau de réussite scolaire est également un facteur clé : les jeunes ayant les moins bons résultats ont eu une moindre propension à répondre (ceux du 1er quartile de l'évaluation en 6ème pour l'enquête EVA, ceux du 2nd quartile pour l'enquête SUP). Enfin, des variables sociales jouent sur la réponse : les jeunes dont les parents sont agriculteurs ou vivent dans de petites unités urbaines ont un taux de participation à l'enquête plus élevé.

Calage sur des données externes

La DEPP dispose d'un certain nombre de données sur les élèves entrés en 6e en 1995, qu'il peut être intéressant de confronter avec l'échantillon des répondants . On a donc calé sur le croisement du secteur, du sexe et de l'âge d'entrée en sixième (le croisement par sexe et âge n'étant disponible que pour le secteur public). Voici les répartitions, avant et après calage (par règle de trois sur les poids) :

	Répartition dans l'échantillon		Réparition dans la population	
Secteur Privé	3420	19,1%	146602	19,4%
Secteur Public				
Garçons en avance	204	1,1%	8218	1,1%
Garçons à l'heure	5293	29,6%	215848	28,6%
Garçons en retard	1804	10,1%	84927	11,3%
Filles en avance	219	1,2%	10423	1,4%
Filles à l'heure	5405	30,2%	224779	29,8%
Filles en retard	1547	8,6%	63128	8,4%

L'échantillon avant calage diffère peu des marges disponibles. Le recalage semble utile, en particulier pour mieux représenter le retard des garçons et filles en 6ème.

Comparaison des poids calés de 2004/2005 à 2006/2007

On vérifie dans cette partie la cohérence des pondérations d'une année sur l'autre. Il faut tout d'abord noter la réduction de la dispersion des pondérations au fil des ans, en particulier la moindre concentration de poids « extrêmes » sur quelques sujets : la valeur maximale passe de 1646 en 2004/2005, à 1132 en 2005/2006 puis 713 en 2006/2007. Ce résultat tient en partie à l'élargissement progressif de l'échantillon EVA, suite à la transmission du suivi par les équipes SUP et SEC. En effet, l'échantillon théorique EVA comportant de plus en plus de jeunes, le calcul des pondérations pour prendre en compte les « exclus » du suivi est peu à peu équilibré sur l'ensemble des répondants, là où il se concentrait sur les quelques jeunes présentant les caractéristiques appropriées dans la collecte 2004-2005.

	Nombre d'observations pondérées	Poids moyen	Min	D1	Q1	Médiane	Q3	D9	Max
2004/2005	11 715	63	43	45	46	49	59	84	1646
2005/2006	11 488	64	39	44	44	47	64	95	1132
2006/2007	11 329	67	41	43	44	46	75	102	713

Seules 9312 observations ont répondu systématiquement sur les 3 années de collecte. Pour ces dernières, on examine la stabilité de la pondération d'une année sur l'autre :

- Pour 75% des observations, l'écart-type de la pondération sur les 3 années est inférieur à 0,15 fois le poids moyen de l'observation. On peut considérer que la structure des poids est alors assez stable ;
- Pour 20% des observations, l'écart-type de la pondération sur les 3 années est compris entre 0,15 et 0,33 fois le poids moyen de l'observation. On peut considérer que la structure des poids est perturbé mais ne biaise pas nécessairement une étude de panel ;
- Pour 5% des observations, l'écart-type de la pondération sur les 3 années est supérieur à 0,33 fois le poids moyen de l'observation. Pour ces sujets l'exploitation en panel est sérieusement mise en question par la variabilité des poids

Ce dernier résultat invite à concevoir, lors de la stabilité du fichier de diffusion EVA, une méthode de calcul des poids qui soit directement conçue pour un usage en panel, et non plus en coupe. Des travaux seront ultérieurement conduits en ce sens.

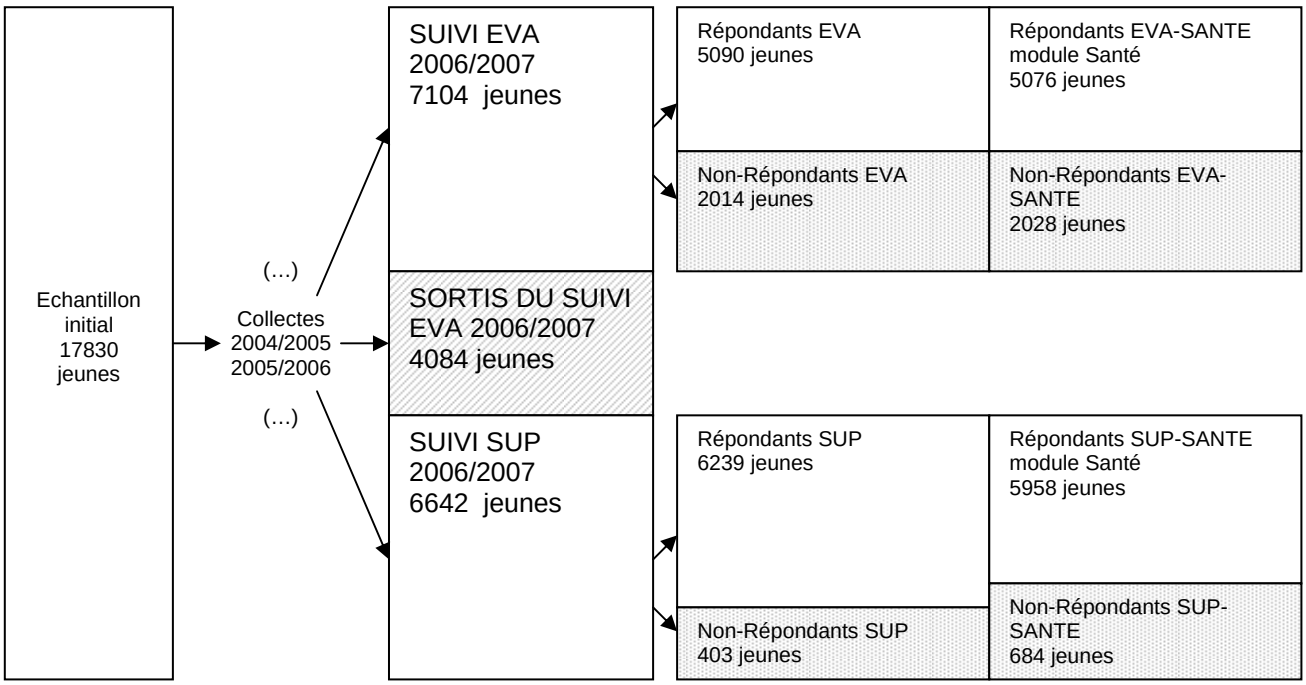
PARTIE 3 - CALCUL DES PONDERATIONS SANTE 2006/2007

Methodologie

Le comportement de non-réponse au module SANTE est très proche de celui de la non-réponse au questionnaire principal. Comme précédemment, le calcul des pondérations s'appuie sur un modèle pour l'échantillon « Insee » et un modèle pour l'échantillon « Supérieur ».

Pour l'échantillon « Insee », nous reprenons à l'identique le calcul de la probabilité ($1-P_{HCH}$) d'être inclus dans la collecte 2006-2007 par rapport au fait d'être « sorti » du suivi EVA. Pour les jeunes « inclus » dans le suivi EVA, nous modélisons ensuite la non réponse au module SANTE, donnant la probabilité P_{eva_sante} de ne pas répondre à ces questions. A noter que l'écart de réponse entre le questionnaire principal EVA (5090 répondants) et le module SANTE (5076 répondants) ne porte que sur 14 cas : les probabilités de non-réponse P_{eva} et P_{eva_sante} sont de ce fait très proches.

Pour l'échantillon « Supérieur », on procède à une seule modélisation, celle de la non réponse au module santé au sein de l'échantillon, donnant la probabilité P_{sup_sante} de ne pas répondre à ces questions. A noter que l'écart de réponse entre le questionnaire principale SUP (6239 répondants) et le module SANTE (5958 répondants) est non négligeable : les probabilités de non-réponse P_{sup} et P_{sup_sante} présentent de ce fait une distribution légèrement différente. La différence dans la non-réponse provient de ce que, en cas d'absence longue durée du jeune, l'équipe SUP de la DEPP interroge son père ou sa mère sur sa situation en matière de formation pour continuer à le suivre, ce qui n'est pas possible pour le questionnaire santé.



La pondération POIDS_SANTE est égale à :

- l'inverse de $(1 - P_{HCH})*(1- P_{eva\_sante})$ pour l'échantillon « Insee »
- l'inverse de $(1- P_{sup\_sante})$ pour l'échantillon « supérieur »

Les régressions logistiques mobilisées dans ces modèles prennent en compte les différentes caractéristiques disponibles présentées dans la partie 2. Globalement, les mêmes variables explicatives sont identifiées, avec des effets comparables (cf. annexes 5 et 6).

La pondération ainsi obtenu a les caractéristiques suivantes :

Pondération EVA-SANTE :

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
2.22	1.29	1.24	1.29	1.46	1.58	1.82	2.31	3.26	7.47	17.82

Pondération SUP :

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
1.11	0.05	1.08	1.08	1.08	1.08	1.12	1.12	1.18	1.30	1.65

Un calage est ensuite effectué sur la base des marges disponibles sur les élèves entrés en 6e en 1995, en croisant secteur, sexe et âge d'entrée en sixième. Voici les répartitions, avant et après calage (par règle de trois sur les poids) :

	Répartition dans l'échantillon		Réparition dans la population	
Secteur Privé	3420	19,1%	146602	19,4%
Secteur Public				
Garçons en avance	206	1,2%	8218	1,1%
Garçons à l'heure	5316	29,7%	215848	28,6%
Garçons en retard	1791	10,0%	84927	11,3%
Filles en avance	219	1,2%	10423	1,4%
Filles à l'heure	5398	30,2%	224779	29,8%
Filles en retard	1538	8,6%	63128	8,4%

Comparaison des poids EVA 2006/2007 et SANTE 2006/2007

On vérifie dans cette partie la cohérence des pondérations obtenues pour le questionnaire principal de l'enquête EVA d'une part et le module complèmentaire SANTE d'autre part (cf. graphique en annexe 7). Globalement, la pondération est très proche sur l'ensemble de l'échantillon des répondants. Par construction, la stabilité est plus marquée dans le sous-échantillon des jeunes suivis par l'Insee, où le taux de non-réponse est presque identique entre la collecte principale et la collecte complémentaire santé. Pour l'échantillon des jeunes suivis par la Depp, les poids sont légèrement plus importants du fait d'une non-réponse un peu plus importante sur le module santé par rapport à celle observée sur le questionnaire principal.

Annexe 1 : procédure de calcul de la pondération

On va décrire ici en détail la procédure de calcul de la pondération, une fois que les fichiers ont été appariés et que chaque individu s'est vu attribué un statut de réponse.

Dans un premier temps, on considère les variables ci-dessous, issues du fichier historique après quelques regroupements de modalité.

Variables utilisés pour l'échantillon secondaire

SEXE	**Sexe**
1	Garçons
2	Filles
LIEUNAI	**Lieu de naissance de l'enfant**
1	France
3	Etranger
NATELEVE	**Nationalité de l'enfant**
10	Française
200	Autre
NATPERE	**Nationalité du père**
0	Père inconnu ou nationalité inconnue
10	Française
20	Autre
PAYPERE	**Pays de naissance du père**
0	Père inconnu ou pays de naissance inconnu
10	France
20	Etranger
NATMERE	**Nationalité de la mère**
0	Mère inconnue ou nationalité inconnue
10	Française
20	Autre
PAYMERE	**Pays de naissance de la mère**
0	Mère inconnue ou pays de naissance inconnu
10	France
20	Etranger
NBENF	**Nombre d'enfants dans la famille**
1	1
2	2
3	3
4	4 ou plus
RANG	**Rang dans la fratrie**
1	1er
2	2ème
3	3ème
4	4ème ou plus
Permer	**Vivait avec son père et sa mère**
0	Non
1	Oui
ACTIPERE	**Activité du père**

0	Père inconnu
1	Actif
2	Inactif
pcsp	**CS du père**
1	Agriculteur
2	Artisan commerçant
3	Cadre supérieur
4	Profession intermédiaire
5	Employé
6	Ouvrier
8	Inactif
9	Père inconnu ou profession inconnue
ACTIMERE	**Activité de la mère**
0	Mère inconnue
1	Active
2	Inactive
pcsm	**CS de la mère**
1	Agriculteur
2	Artisan commerçant
3	Cadre supérieur
4	Profession intermédiaire
5	Employé
6	Ouvrier
8	Inactif
9	Mère inconnue ou profession inconnue
quartot	**quartile de réussite aux évaluations 6e**
0	Premier quartile
1	Deuxième quartile
2	Troisième quartile
3	Quatrième quartile
AGE6E	**Age d'entrée en sixième**
10	10 ou moins
11	11 ans
12	12 ans ou plus
SECTECO	**Secteur de l'école primaire**
1	Public
2	Privé
ZEP1995	**Présence en Zep en 1995**
1	Oui
2	Non
HEBERG1995	**Hébergement en 1995**
1	Externe
2	Demi-pensionnaire, interne, autre
TUETAB1995	**Tranche d'unité urbaine de l'établissement de 1995+B105**
0	Commune rurale
1	Commune urbaine de moins de 5 000 habitants
2	Commune urbaine de 5 000 a moins de 10 000 habitants
3	Commune urbaine de 10 000 a moins de 20 000 habitants
4	Commune urbaine de 20 000 a moins de 50 0000 habitants

5	Commune urbaine de 50 000 a moins de 100 000 habitants
6	Commune urbaine de 100 000 a moins de 200 000 habitants
7	Commune urbaine de 200 000 a moins de 2 000 000 habitants
8	Agglomération parisienne
SECTEUR1995	**Secteur de l'établissement de 1995**
1	Public
2	Privé
changet	**Changement d'établissement durant la scolarité au collège**
0	Non
1	Oui
boursier	**Boursier au collège**
0	Non
1	Oui
repfam	**Réponse à l'enquête Famille**
1	Réponse postale
2	Réponse téléphonique
3	Non interrogé ou non répondant
repjeune	**Réponse à l'enquête Jeune**
1	Réponse postale
2	Réponse téléphonique
3	Non interrogé ou non répondant
bachelier	**Année**
0000	N'a pas le bac
2002	En 2002
2003	En 2003
2004	En 2004
mention	**Mention au bac**
0-no	N'a pas le bac
0	Pas de mention
1-AB	Assez bien
2- B	Bien
3-TB	Très bien
rsortie	**Raison de la sortie du panel scolaire**
0	Pas sorti
1	Vie active
2	Chômage
3	Non scolarisé
5	Décès
6	Abandon de scolarité pour raison de santé
20	Autre
950	Université

ANNEE_SORTIE	Année de sortie du panel scolaire
Année_sortie_1	1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001
Année_sortie_2	2002, 2003
Année_sortie_3	2004

Après un Proc Freq pour voir les effectifs des différentes modalités des variables sur le champ considéré, on étudie l'impact individuel de chaque variable sur la non-réponse, ce qui établit une première sélection. Ensuite, on construit un modèle incluant toutes les variables, avec l'option STEPWISE qui sélectionne les plus pertinentes, au sens statistique du terme. La combinaison de ces deux sélections permet, après quelques regroupements de modalités, de déterminer les variables à conserver. On calcule alors les indicatrices correspondantes et on lance le modèle final, qui construira la table avec la probabilité de répondre associée à chaque répondant. C'est ce type de modèles pour chaque population, qui est présenté dans les annexes suivantes.

Pour 2006/2007, les indicatrices ont été calculées de la façon suivante :

```
data SASTEMP.database_pond2007;set SASTEMP.database_pond2007;
gars=(sexe='1');
lieunai1=(lieunai='1');
nateleve1=(natelevegr='01');
natp00=(natperegr='00');natp20=(natperegr='02');
natm00=(natmeregr='00');natm20=(natmeregr='02');
payp00=(payperegr='00');payp20=(payperegr='02');
paym00=(paymeregr='00');paym20=(paymeregr='02');
nbenf12=(nbenf in ('1','2'));
rang1=(rang='1');
rang2=(rang='2');
rang3=(rang='3');
rang4=(rang='4');
actp0=(actipere='0');actp1=(actipere='1');
pcsp12=(pcsp in ('1','2'));
pcsp56=(pcsp in ('5','6'));
pcsp1=(pcsp='1');pcsp2=(pcsp='2');pcsp3=(pcsp='3');pcsp4=(pcsp='4');
pcsp5=(pcsp='5');pcsp6=(pcsp='6');pcsp9=(pcsp='9');pcsp8=(pcsp='8');
actm0=(actimere='0');actm1=(actimere='1');
pcsm1=(pcsm='1');pcsm2=(pcsm='2');pcsm3=(pcsm='3');pcsm4=(pcsm='4');
pcsm5=(pcsm='5');pcsm6=(pcsm='6');pcsm9=(pcsm='9');pcsm8=(pcsm='8');
quart1=(quartot='0');quart2=(quartot='1');quart3=(quartot='2');quart4=(quar
tot='3');
zep1995_IND=(zep1995='1');
repfam1=(repfam='1');repfam2=(repfam='2');repfam3=(repfam='3');
repjeune1=(repjeune='1');repjeune2=(repjeune='2');repjeune3=(repjeune='3');
rsortie001=(rsortie='001');rsortie002=(rsortie='002');rsortie003=(rsortie='
003');
rsortie020=(rsortie in ('020','02'));rsortie950=(rsortie='950');
prive=(secteur1995='2');
petitu=(tuetab1995 in ('0','1','2','3'));
heberg=(heberg1995='1');
age6e_12=(age6e in ('12','13'));
bachelier_IND=(BAC_LAUREAT ne 'XXXX');
annee_sortie_1=(annee_sortie in
('0000','1994','1995','1996','1998','1999','2000','2001'));
annee_sortie_2=(annee_sortie in ('2002','2003'));
annee_sortie_3=(annee_sortie in ('2004'));
run;
```

Annexe 2 : exclusion de l'échantillon « eva »

Response Profile

Ordered Value	HCH_EVA	Total Frequency
1	1	4084
2	0	7104

Class Level Information

Class	Value	Design Variables
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1
age6e_12	0	0
	1	1
petitu	0	0
	1	1
bachelier_IND	0	0
	1	1

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	14686.466	13228.156
SC	14693.789	13279.414
-2 Log L	14684.466	13214.156

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1470.3102	6	<.0001
Score	1415.5193	6	<.0001
Wald	1251.3832	6	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Permer	1	121.0850	<.0001
repfam3	1	216.4365	<.0001
repjeune3	1	260.5648	<.0001
age6e_12	1	21.8804	<.0001
petitu	1	48.8526	<.0001
bachelier_IND	1	298.6360	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.2556	0.0536	22.7512	<.0001
Permer 1	1	-0.5253	0.0477	121.0850	<.0001
repfam3	1	0.7696	0.0523	216.4365	<.0001
repjeune3	1	0.7394	0.0458	260.5648	<.0001
age6e_12	1	0.2150	0.0460	21.8804	<.0001
petitu	1	-0.3084	0.0441	48.8526	<.0001
bachelier_IND	1	-0.8787	0.0508	298.6360	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
Permer	1 vs 0	0.591	0.539	0.649
repfam3	1 vs 0	2.159	1.948	2.392
repjeune3	1 vs 0	2.095	1.915	2.291
age6e_12	1 vs 0	1.240	1.133	1.357
petitu	1 vs 0	0.735	0.674	0.801
bachelier_IND	1 vs 0	0.415	0.376	0.459

Association of Predicted Probabilities and Observed Responses

Percent Concordant	69.0	Somers' D	0.424
Percent Discordant	26.6	Gamma	0.444
Percent Tied	4.4	Tau-a	0.197
Pairs	29012736	c	0.712

Annexe 3 : non-réponse dans l'échantillon « eva»

Response Profile

Ordered Value	NR_eva	Total Frequency
1	1	2014
2	0	5090

Class Level Information

Class	Value	Design Variables
gars	0	0
	1	1
natp20	0	0
	1	1
pcsp1	0	0
	1	1
pcsp8	0	0
	1	1
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1
quart1	0	0
	1	1
petitu	0	0
	1	1
annee_sortie_3	0	0
	1	1

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	8473.246	8098.359
SC	8480.114	8173.912
-2 Log L	8471.246	8076.359

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	394.8867	10	<.0001
Score	401.9145	10	<.0001
Wald	373.1143	10	<.0001

Type III Analysis of Effects

	Wald

Effect	DF	Chi-Square	Pr > ChiSq
gars	1	12.9788	0.0003
natp20	1	11.5139	0.0007
pcsp1	1	11.8090	0.0006
pcsp8	1	6.8014	0.0091
Permer	1	32.4886	<.0001
repfam3	1	34.0738	<.0001
repjeune3	1	121.2357	<.0001
quart1	1	6.0938	0.0136

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
petitu	1	28.2839	<.0001
annee_sortie_3	1	20.6606	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-0.9946	0.0742	179.4511	<.0001
gars 1	1	0.1986	0.0551	12.9788	0.0003
natp20	1	0.2800	0.0825	11.5139	0.0007
pcsp1	1	-0.6585	0.1916	11.8090	0.0006
pcsp8	1	0.4602	0.1765	6.8014	0.0091
Permer	1	-0.3776	0.0662	32.4886	<.0001
repfam3	1	0.4449	0.0762	34.0738	<.0001
repjeune3	1	0.6986	0.0634	121.2357	<.0001
quart1	1	0.1440	0.0583	6.0938	0.0136
petitu	1	-0.3071	0.0577	28.2839	<.0001
annee_sortie_3 1	1	0.2786	0.0613	20.6606	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
gars	1 vs 0	1.220	1.095	1.359
natp20	1 vs 0	1.323	1.126	1.555
pcsp1	1 vs 0	0.518	0.356	0.754
pcsp8	1 vs 0	1.584	1.121	2.239
Permer	1 vs 0	0.686	0.602	0.781
repfam3	1 vs 0	1.560	1.344	1.812
repjeune3	1 vs 0	2.011	1.776	2.277
quart1	1 vs 0	1.155	1.030	1.295
petitu	1 vs 0	0.736	0.657	0.824
annee_sortie_3	1 vs 0	1.321	1.172	1.490

Association of Predicted Probabilities and Observed Responses

Percent Concordant	63.1	Somers' D	0.287
Percent Discordant	34.4	Gamma	0.295
Percent Tied	2.5	Tau-a	0.117
Pairs	10251260	c	0.644

Annexe 4 : non-réponse dans le « supérieur »

Response Profile

Ordered Value	NR_sup	Total Frequency
1	1	403
2	0	6239

Class Level Information

Class	Value	Design Variables
natm20	0	0
	1	1
pcsp1	0	0
	1	1
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1
quart2	0	0
	1	1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	3041.637	3004.222
SC	3048.438	3051.830
-2 Log L	3039.637	2990.222

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	49.4152	6	<.0001
Score	54.5204	6	<.0001
Wald	51.8385	6	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
natm20	1	14.5010	0.0001
pcsp1	1	4.7599	0.0291
Permer	1	4.5128	0.0336
repfam3	1	5.8230	0.0158
repjeune3	1	3.6695	0.0554

quart2	1	6.6052	0.0102

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	-2.6461	0.1394	360.5113	<.0001
natm20	1	1	0.5931	0.1557	14.5010	0.0001
pcsp1	1	1	-0.9947	0.4559	4.7599	0.0291
Permer	1	1	-0.3037	0.1430	4.5128	0.0336
repfam3	1	1	0.4182	0.1733	5.8230	0.0158
repjeune3	1	1	0.3370	0.1759	3.6695	0.0554
quart2	1	1	0.3106	0.1208	6.6052	0.0102

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
natm20	1 vs 0	1.810	1.334	2.455
pcsp1	1 vs 0	0.370	0.151	0.904
Permer	1 vs 0	0.738	0.558	0.977
repfam3	1 vs 0	1.519	1.082	2.134
repjeune3	1 vs 0	1.401	0.992	1.977
quart2	1 vs 0	1.364	1.077	1.729

Association of Predicted Probabilities and Observed Responses

Percent Concordant	43.4	Somers' D	0.186
Percent Discordant	24.8	Gamma	0.272
Percent Tied	31.7	Tau-a	0.021
Pairs	2514317	c	0.593

Annexe 5 : non-réponse SANTE dans l’échantillon « eva»

Response Profile

Ordered Value	NR_sante_EVA	Total Frequency
1	1	2028
2	0	5076

Class Level Information

Class	Value	Design Variables
gars	0	0
	1	1
natp20	0	0
	1	1
pcsp1	0	0
	1	1
pcsp3	0	0
	1	1
pcsp8	0	0
	1	1
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1
quart1	0	0
	1	1
petitu	0	0
	1	1
annee_sortie_3	0	0
	1	1

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	8499.071	8116.026
SC	8505.939	8198.447
-2 Log L	8497.071	8092.026

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	405.0451	11	<.0001
Score	410.8054	11	<.0001
Wald	380.7389	11	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
gars	1	14.7599	0.0001
natp20	1	13.5704	0.0002
pcsp1	1	12.5915	0.0004
pcsp3	1	6.6516	0.0099
pcsp8	1	9.0257	0.0027
Permer	1	34.4239	<.0001
repfam3	1	33.3899	<.0001
repjeune3	1	117.7918	<.0001
quart1	1	7.7952	0.0052
petitu	1	23.5535	<.0001
annee_sortie_3	1	22.1088	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		1	-1.0257	0.0752	186.1805	<.0001
gars	1	1	0.2117	0.0551	14.7599	0.0001
natp20	1	1	0.3046	0.0827	13.5704	0.0002
pcsp1	1	1	-0.6887	0.1941	12.5915	0.0004
pcsp3	1	1	0.2415	0.0936	6.6516	0.0099
pcsp8	1	1	0.5284	0.1759	9.0257	0.0027
Permer	1	1	-0.3890	0.0663	34.4239	<.0001
repfam3	1	1	0.4404	0.0762	33.3899	<.0001
repjeune3	1	1	0.6887	0.0635	117.7918	<.0001
quart1	1	1	0.1645	0.0589	7.7952	0.0052
petitu	1	1	-0.2816	0.0580	23.5535	<.0001
annee_sortie_3	1	1	0.2880	0.0613	22.1088	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
gars	1 vs 0	1.236	1.109	1.377
natp20	1 vs 0	1.356	1.153	1.595
pcsp1	1 vs 0	0.502	0.343	0.735
pcsp3	1 vs 0	1.273	1.060	1.530
pcsp8	1 vs 0	1.696	1.202	2.394
Permer	1 vs 0	0.678	0.595	0.772
repfam3	1 vs 0	1.553	1.338	1.804

repjeune3	1 vs 0	1.991	1.758	2.255
quart1	1 vs 0	1.179	1.050	1.323
petitu	1 vs 0	0.755	0.673	0.845
annee_sortie_3	1 vs 0	1.334	1.183	1.504

Association of Predicted Probabilities and Observed Responses

Percent Concordant	63.5	Somers' D	0.291
Percent Discordant	34.4	Gamma	0.297
Percent Tied	2.0	Tau-a	0.119
Pairs	10294128	c	0.646

Annexe 6 : non-réponse SANTE dans l'échantillon « supérieur »

Response Profile

Ordered Value	NR_sante_SUP	Total Frequency
1	1	684
2	0	5958

Class Level Information

Class	Value	Design Variables
gars	0	0
	1	1
natp20	0	0
	1	1
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	4406.762	4333.985
SC	4413.563	4374.792
-2 Log L	4404.762	4321.985

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	82.7771	5	<.0001
Score	89.6100	5	<.0001
Wald	86.4071	5	<.0001

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
gars	1	33.3079	<.0001
natp20	1	7.3319	0.0068
Permer	1	15.0547	0.0001
repfam3	1	8.2779	0.0040
repjeune3	1	11.5794	0.0007

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-2.1258	0.1125	356.7788	<.0001
gars 1	1	0.4714	0.0817	33.3079	<.0001
natp20 1	1	0.3604	0.1331	7.3319	0.0068
Permer 1	1	-0.4339	0.1118	15.0547	0.0001
repfam3 1	1	0.4050	0.1408	8.2779	0.0040
repjeune3 1	1	0.4656	0.1368	11.5794	0.0007

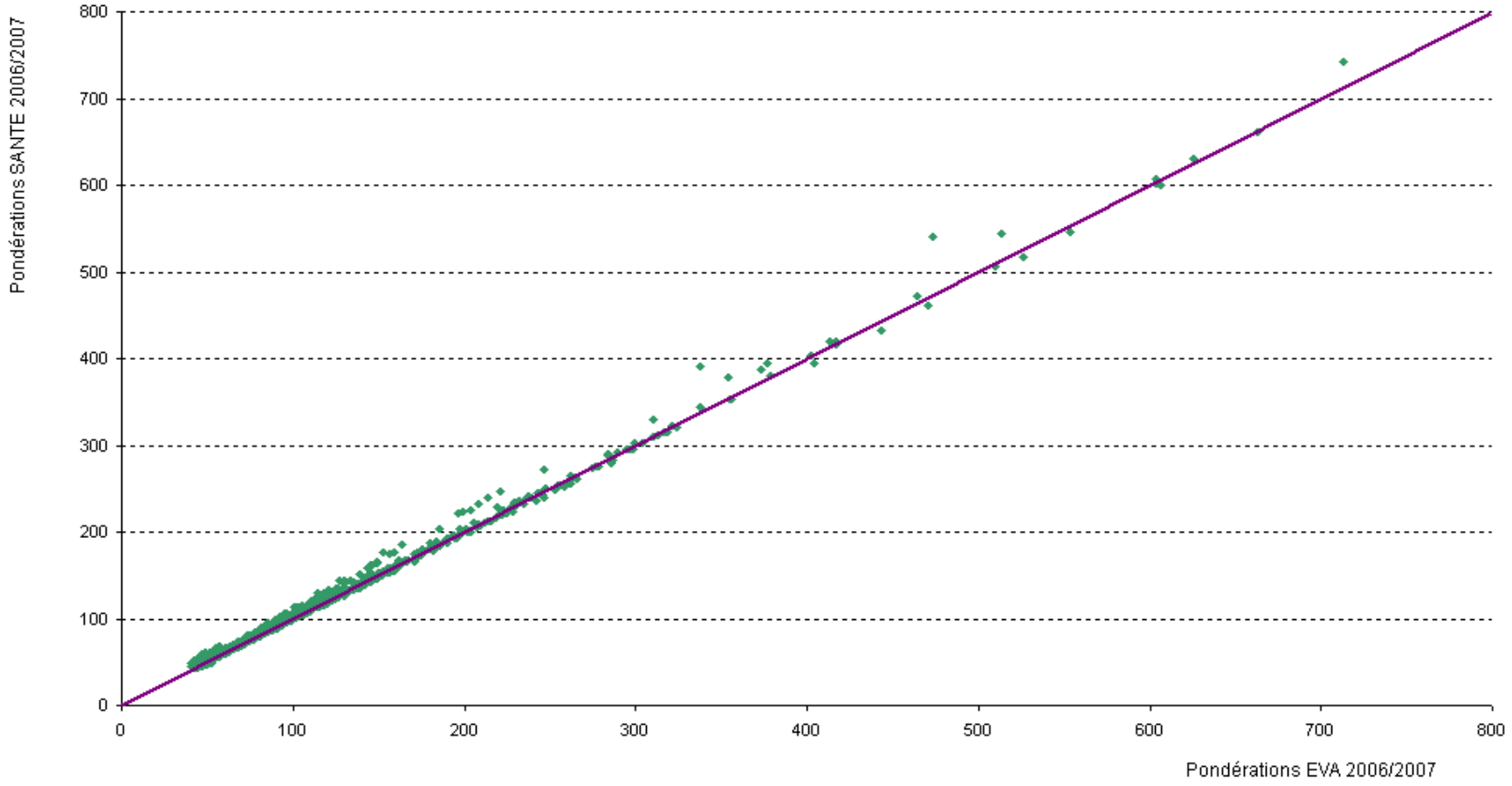
Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
gars	1 vs 0	1.602	1.365	1.880
natp20	1 vs 0	1.434	1.105	1.861
Permer	1 vs 0	0.648	0.520	0.807
repfam3	1 vs 0	1.499	1.138	1.976
repjeune3	1 vs 0	1.593	1.218	2.083

Association of Predicted Probabilities and Observed Responses

Percent Concordant	47.6	Somers' D	0.184
Percent Discordant	29.2	Gamma	0.239
Percent Tied	23.2	Tau-a	0.034
Pairs	4075272	c	0.592

Annexe 7 : comparaison des pondérations EVA et SANTE 2006/2007



Direction des Statistiques Démographiques et Sociales
Département de l'Emploi et des Revenus d'Activité
Division Emploi

NOTE

Dossier suivi par :
Sébastien Gossiaux
Tél. : 01 41 17 54 52
Fax : 01 41 17 61 63
Mél : sebastien.gossiaux@insee.fr

Paris, le
N° / DG75-F230 /

Objet : Calcul d'une pondération globale en 2007/2008 pour le panel d'élèves entrés en 6e en 1995

Le panel de 1995 compte 17 830 jeunes. Il est constitué de l'ensemble des élèves nés le 17 d'un mois (sauf mars, juillet et octobre pour obtenir un taux de 1/40e environ) qui sont entrés en 6e en septembre 1995.

De 1995 à 2002, le suivi des élèves a été assuré au moyen d'une collecte annuelle de données administratives et grâce à des enquêtes ponctuelles, l'enquête Familles de 1998 et l'enquête Jeunes de 2002. L'ingénierie statistique a été assurée par l'équipe DEPP du secondaire.

A partir de l'année scolaire 2002/2003, ceux qui ont poursuivi des études après avoir obtenu le baccalauréat ont été suivis par l'équipe DEPP du supérieur (enquête dite « SUP ») tandis que ceux qui étaient toujours dans le secondaire continuaient à être suivis par l'équipe DEPP du secondaire (enquête dite « SEC »).

Depuis 2005 l'Insee enquête annuellement les jeunes du panel qui ont terminé leurs études sur leur entrée dans la vie adulte (enquête dite « EVA »). Parmi les thèmes abordés, on peut citer les conditions de vie, la raison principale de l'arrêt des études et le parcours professionnel depuis la sortie du système scolaire.

La « sortie » des jeunes du système éducatif n'a pas permis de tous les retrouver en 2005, en raison des déménagements et des difficultés à obtenir les nouvelles coordonnées. Ce problème a été particulièrement prononcé pour les premiers sortants du suivi assuré par l'équipe DEPP du secondaire. Les jeunes concernés sont soit des « sortants précoces » du secondaire, c'est-à-dire des jeunes qui ont quitté leur cursus scolaire sans valider le diplôme qu'ils préparaient, soit de jeunes diplômés qui n'ont pas pousuivi d'études. L'incertitude sur leur parcours (poursuivent-ils des études ? se portent-ils sur le marché du travail ?) a retardé la transmission de leurs coordonnées auprès de l'Insee. 2268 jeunes ont ainsi été « perdus » et n'ont pas pu être recontactés par l'Insee lors de la première enquête EVA en 2005.

Par ailleurs, 370 jeunes sont sortis du champ lors de leurs études secondaires du fait d'un départ à l'étranger, d'une maladie grave ou d'un décès.

Enfin, la non-réponse d'une partie des jeunes, quand bien même ces derniers ont pu être contactés, conduit à une attrition régulière. Sont exclus du suivi les non-répondants récurrents, c'est-à-dire les jeunes qui ne répondent pas deux fois de suite à l'enquête EVA. Les non-répondants à l'enquête SUP sont suivis dans le cadre de l'enquête EVA l'année suivante.

La perte d'information et l'attrition conduisent à une déformation de l'échantillon par rapport à l'échantillon initial, telle que présentée en **partie 1**. Pour limiter les biais induits, une pondération est calculée chaque année sur l'ensemble des jeunes qui ont répondu à l'un des questionnaires de suivi. La **partie 2** présente une proposition de calcul d'une pondération en 2007/2008, s'inspirant de ce qui a été fait depuis 2004/2005.

1. Attrition et non réponse

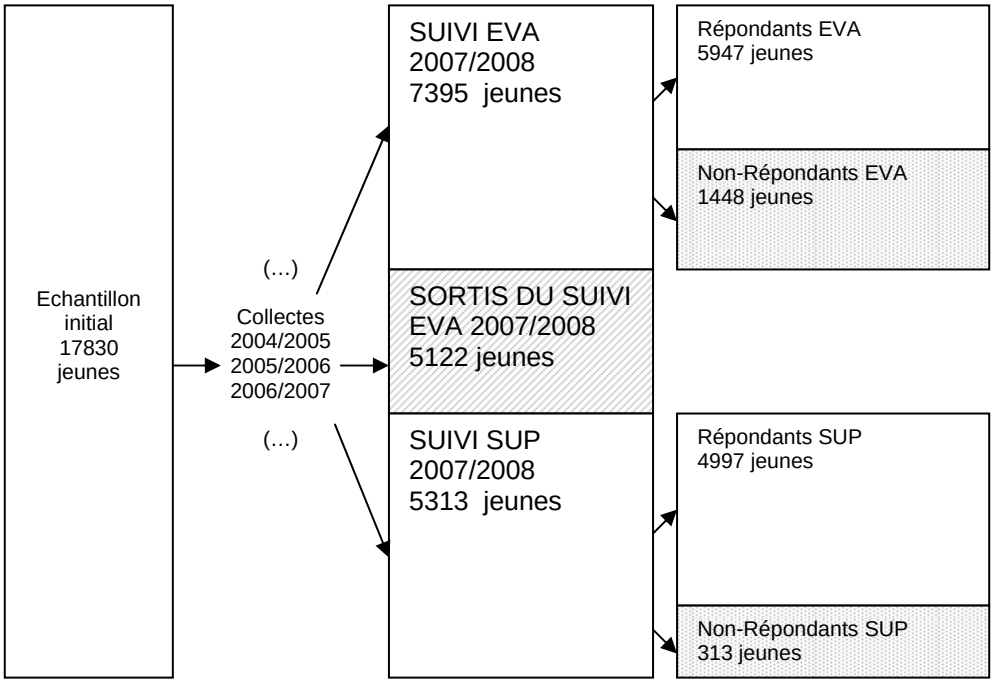
Les jeunes qui ont quitté le plus tard le secondaire y étaient pour la dernière fois en 2005/2006. De ce fait, l'intégralité du suivi est maintenant réalisée uniquement grâce aux enquêtes SUP et EVA.

Le fichier complet pour l'année 2007/2008 comprend donc

- Un échantillon de 7395 jeunes interrogés par l'Insee dans le cadre du questionnaire EVA :
  - 5102 sont en réinterrogation après avoir déjà répondu à l'enquête EVA en 2007.
  - 1328 sont de nouveaux enquêtés EVA en 2008 en sortie du SUP.
  - 532, qui avaient répondu à l'enquête EVA 2006, sont relancés en 2008 après ne pas avoir répondu en 2007.
  - 433 sont relancés en 2008 après ne pas avoir répondu à leur questionnaire de première interrogation en 2007.

- Un échantillon de 5313 jeunes interrogés par la Depp dans le cadre du questionnaire SUP.

- Un échantillon de 5122 jeunes qui n'ont pas été enquêtés pour les motifs suivants :
  - jeunes sortis du champ lors de leurs études secondaires (maladie, décès, départ à l'étranger) : 370 jeunes
  - les jeunes « perdus » (adresse inexploitable) lors de leur sortie du systèm éducatif : 2268 jeunes
  - les jeunes non-répondants de façon récurrente : 2484 jeunes.



Analyse des taux de réponse

Quel que le soit le type de questionnaire analysé, les taux de réponse se sont nettement améliorés entre 2004/2005 et 2007/2008. Les taux de réponse ont été particulièrement bas lors de la première collecte EVA du fait d'un grand nombre de jeunes pour lesquels les coordonnées postales n'ont pas été exploitables. Par ailleurs, les non-répondants récurrents ont été exclus du champ de la collecte 2007/2008. Enfin, les nouveaux interrogés sont issus du suivi SUP, au cours duquel une relative « habitude » de réponse aux enquêtes postales a pu prendre place.

	2007/2008			2006/2007		
	Enquêtés	Répondants	Taux de réponse	Enquêtés	Répondants	Taux de réponse
Questionnaire EVA	7395	5947	80%	7104	5090	72%
Questionnaire SUP	5313	4997	94%	6642	6239	94%
	2005/2006			2004/2005		
	Enquêtés	Répondants	Taux de réponse	Enquêtés	Répondants	Taux de réponse
Questionnaire EVA	6384	3949	62%	7256	3253	45%
Questionnaire SUP	7965	7238	91%	8576	7522	88%
Questionnaire SEC	459	302	66%	1629	940	58%

2. Calcul de la pondération EVA 2007/2008

Le calcul de la pondération EVA 2007/2008 repose sur la méthode utilisée au cours des collectes précédentes. A partir de régressions logistiques on commence par évaluer la probabilité de réponse à l'enquête en fonction des caractéristiques des individus.

Pour l'échantillon « Insee », on procède à deux modélisations :

- on cherche tout d'abord à modéliser la probabilité d'être inclus dans la collecte 2007-2008 par rapport au fait d'être « sorti » du suivi EVA. Il s'agit en particulier de prendre en compte le profil spécifique des jeunes qui ont été « perdus » lors de la première collecte EVA, du fait d'adresses postales inexploitables, ainsi que des jeunes qui n'ont jamais répondu au questionnaire (non-répondants récurrents). La probabilité d'être « sorti » de l'échantillon EVA est noté P_{HCH} ;
- Pour les jeunes « inclus » dans le suivi EVA, on modélise la non réponse, donnant la probabilité P_{eva} de ne pas répondre à cette enquête ;

Pour l'échantillon « Supérieur », on procède à une seule modélisation, celle de la non réponse au sein de l'échantillon, donnant la probabilité P_{sup} de ne pas répondre à cette enquête.

La richesse des informations contenues dans le panel permet de caractériser les répondants avec de nombreuses variables :

- les origines socio-démographiques en 1995 (nationalité et activité professionnelle des parents, composition de la famille, etc.) ;
- le niveau scolaire à l'entrée en 6^{e} (redoublements à l'école élémentaire, résultats aux évaluations de 6^{e}, etc.)
- le parcours scolaire dans le secondaire (orientation en fin de 3^{e}, obtention du bac, mention éventuelle, année de sortie du panel scolaire)

- le « comportement de réponse » aux précédentes enquêtes (l'enquête famille 1998 et l'enquête Jeunes 2002).

La sélection des variables utilisées pour calculer P_{HCH}, P_{eva} et P_{sup} repose sur plusieurs critères.
Tout d'abord, les variables doivent être significatives au sens statistique du terme. Les résultats doivent ensuite être robustes, ce qui implique la sélection de variables renseignées pour un grand nombre de personnes et avec des effectifs pas trop faibles dans chacune des modalités. Enfin nous ne retenons que des variables dont nous pouvons comprendre l'influence sur la présence de l'individu dans l'échantillon et/ou sur le comportement de réponse.

Le détail de la procédure de calcul de la pondération figure en annexe avec la liste des variables finalement utilisées et les programmes SAS explicitant les modèles ayant servi à calculer les probabilités de réponse à l'enquête.

Le poids de chaque individu est égal à :
- l'inverse de $(1 - P_{HCH})*(1- P_{eva})$ pour l'échantillon « Insee »
- l'inverse de $(1- P_{sup})$ pour l'échantillon « supérieur »

Avant calage, la pondération ainsi obtenue a les caractéristiques suivantes :

Pondération EVA :

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
2,10	1,47	1,39	1,39	1,39	1,39	1,59	1,98	3,37	8,99	17,76

Pondération SUP :

Moy.	σ	Min	C1	D1	Q1	Med	Q3	D90	C99	Max
1,06	0,03	1,04	1,04	1,04	1,06	1,06	1,06	1,11	1,17	1,33

La très faible non-réponse dans la collecte SUP (6 %) conduit à un réajustement des poids marginal. Dans la collecte EVA, la prise en compte des « sortis » de la collecte (41 % sur le champ potentiel de collecte de 12 517 jeunes en 2007/2008) puis une non-réponse plus prononcée (20% sur le champ effectif des 7 395 jeunes enquêtés en 2007/2008) amène à une déformation des poids plus notable. Il n'y a toutefois pas de concentration de pondération sur des cas isolés : le poids maximal est de « seulement » 17,76, avec un 9ème décile à 3,4.

On note l'importance jouée par l'environnement familial dans le fait d'être inclus dans l'enquête ou d'y répondre : le fait que le jeune vivait, en 6ème, avec son père et sa mère réduit significativement la probabilité de sortie du panel ou de non-réponse à l'enquête EVA ou SUP de 2007/2008.
Le comportement de réponse aux enquêtes conduites pendant le secondaire est également un bon prédicteur : les jeunes n'ayant pas répondu à l'enquête Familles 1998 ou à l'enquête Jeunes 2002 sont moins nombreux à être inclus dans le suivi ou à répondre à la collecte en cours.
Le niveau de réussite scolaire est également un facteur clé : les sortants précoces ont une moindre propension à répondre. A l'inverse, dans le supérieur, les jeunes ayant obtenu une mention au baccalauréat répondent mieux que les autres.
Les origines du jeune ont également de l'importance. Les enfants d'étrangers sont plus fréquemment sortis du panel et ils ont une probabilité de non-réponse plus grande. Les jeunes qui étaient dans un établissement dans l'agglomération parisienne en 6^{e} ont eux aussi une probabilité plus grande d'être sorti du panel et leur probabilité de réponse à l'enquête EVA est plus faible.

Calage sur des données externes

Pour obtenir les poids définitifs on procède à un calage sur données externes selon le sexe, l'âge d'entrée en sixième et le nature de l'établissement en 1995 (public ou privé).

	Répartition dans l'échantillon		Réparition dans la population	
Secteur Privé	3444	19,3%	146602	19,4%
Secteur Public				
Garçons en avance	222	1,2%	8218	1,1%
Garçons à l'heure	5219	29,3%	215848	28,6%
Garçons en retard	1766	9,9%	84927	11,3%
Filles en avance	202	1,1%	10423	1,4%
Filles à l'heure	5422	30,4%	224779	29,8%
Filles en retard	1555	8,7%	63128	8,4%

L'échantillon avant calage diffère peu des marges disponibles. Le recalage semble utile, en particulier pour mieux représenter le retard des garçons en 6ème.

Comparaison des poids calés de 2004/2005 à 2007/2008

On vérifie dans cette partie la cohérence des pondérations d'une année sur l'autre.
La moindre concentration de poids « extrêmes » sur quelques sujets ces deux dernières années tient en partie à l'élargissement progressif de l'échantillon EVA, suite à la transmission du suivi par les équipes SUP et SEC. En effet, l'échantillon théorique EVA comportant de plus en plus de jeunes, le calcul des pondérations pour prendre en compte les « exclus » du suivi est peu à peu équilibré sur l'ensemble des répondants, là où il se concentrait sur les quelques jeunes présentant les caractéristiques appropriées dans la collecte 2004-2005.

	Nombre d'observations pondérées	Poids moyen	Min	D1	Q1	Médiane	Q3	D9	Max
2004/2005	11 715	63	43	45	46	49	59	84	1646
2005/2006	11 488	64	39	44	44	47	64	95	1132
2006/2007	11 329	67	41	43	44	46	75	102	713
2007/2008	10 944	69	39	44	44	58	68	98	854

Seules 8646 observations ont répondu systématiquement sur les 4 années de collecte. Pour ces dernières, on examine la stabilité de la pondération d'une année sur l'autre :

- Pour 75% des observations, l'écart-type de la pondération sur les 3 années est inférieur à 0,15 fois le poids moyen de l'observation. On peut considérer que la structure des poids est alors assez stable ;
- Pour 20% des observations, l'écart-type de la pondération sur les 3 années est compris entre 0,15 et 0,34 fois le poids moyen de l'observation. On peut considérer que la structure des poids est perturbé mais ne biaise pas nécessairement une étude de panel ;
- Pour 5% des observations, l'écart-type de la pondération sur les 3 années est supérieur à 0,33 fois le poids moyen de l'observation. Pour ces sujets l'exploitation en panel est sérieusement mise en question par la variabilité des poids

Ce dernier résultat invite à concevoir, lors de la stabilité du fichier de diffusion EVA, une méthode de calcul des poids qui soit directement conçue pour un usage en panel, et non plus en coupe.

La chef de la division Emploi

Annexe : procédure de calcul de la pondération

1. Programme SAS ayant servi à sélectionner les variables et les probabilités de réponse à l'enquête.

```
%macro model(table,var1,var2,param);
%if &param=1 %then %do;
proc logistic descending data=&table(where=(&var1 ne .)) outest=i;
class &var2 /param=ref ref=FIRST
model &var1=&var2/stepwise;
run;%end;
%if &param=2 %then %do;
proc logistic descending data=&table(where=(&var1 ne .)) outest=i;
class &var2 /param=ref ref=FIRST;
model &var1=&var2;
output out=sort_&var1 p=p_&var1;
run;%end;
%mend;
```

2. Liste des variables retenues

Origines

PERMER - Vivait avec son père et sa mère en 1995
1 - Oui
0 - Non

Natm20 - Mère de nationalité étrangère
1 - Oui
0 - Non

Tuetab1995_8 - Etablissement scolaire en 1995 dans l'agglomération parisienne
1 - Oui
0 - Non

Parcours scolaire

Annee_sortie_1 - Sortie du panel scolaire avant d'avoir fait 7 année dans l'enseignement secondaire
1 - Oui
0 - Non

Bachelier_IND - Obtention du baccalauréat
1 - Oui
0 - Non

Mentionbac - Mention au baccalauréat
1 - Oui
0 - Non

Comportement de réponse aux précédentes enquêtes

Repfam3 - Non interrogé ou non répondant à l'enquête Familles 1998
1 - Oui
0 - Non

Repjeune3 - Non interrogé ou non répondant à l'enquête Jeunes 2002
1 - Oui

0 - Non

3. Calcul de P_{HCH} , la probabilité d'être « sorti » du suivi EVA

```
%model(SASTEMP.database_pond2008,HCH_eva,permer repfam3 repjeune3
tuetab1995_8 bachelier_IND natm20 annee_sortie_1,2);run;
```

Response Profile

Ordered Value	HCH_EVA	Total Frequency
1	1	5122
2	0	7395

Probability modeled is HCH_EVA=1.

Class Level Information

Class	Value	Design Variables
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1
tuetab1995_8	0	0
	1	1
bachelier_IND	0	0
	1	1
natm20	0	0
	1	1
annee_sortie_1	0	0
	1	1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	16939.187	14184.288
SC	16946.621	14243.767
-2 Log L	16937.187	14168.288

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2768.8985	7	<.0001
Score	2631.5686	7	<.0001
Wald	2210.1593	7	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Permer	1	127.2978	<.0001
repfam3	1	120.6790	<.0001
repjeune3	1	596.3712	<.0001
tuetab1995_8	1	17.3322	<.0001
bachelier_IND	1	183.6296	<.0001
natm20	1	16.8821	<.0001
annee_sortie_1	1	129.2526	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.4672	0.0531	77.5562	<.0001
Permer	1	1	-0.5465	0.0484	127.2978	<.0001
repfam3	1	1	0.6033	0.0549	120.6790	<.0001
repjeune3	1	1	1.1227	0.0460	596.3712	<.0001
tuetab1995_8	1	1	0.2392	0.0575	17.3322	<.0001
bachelier_IND	1	1	-0.6474	0.0478	183.6296	<.0001
natm20	1	1	0.2606	0.0634	16.8821	<.0001
annee_sortie_1	1	1	0.6150	0.0541	129.2526	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
Permer	1 vs 0	0.579	0.527	0.637
repfam3	1 vs 0	1.828	1.642	2.036
repjeune3	1 vs 0	3.073	2.808	3.363
tuetab1995_8	1 vs 0	1.270	1.135	1.422
bachelier_IND	1 vs 0	0.523	0.477	0.575
natm20	1 vs 0	1.298	1.146	1.469
annee_sortie_1	1 vs 0	1.850	1.664	2.056

Association of Predicted Probabilities and Observed Responses

Percent Concordant	73.3	Somers' D	0.519
Percent Discordant	21.4	Gamma	0.548
Percent Tied	5.3	Tau-a	0.251
Pairs	37877190	c	0.759

4. Calcul de P_{eva} , la probabilité de ne pas avoir répondu à l'enquête EVA 2007/2008

```
%model(SASTEMP.database_pond2008,NR_eva,permer repfam3 repjeune3
tuetab1995_8 natm20,2);run;
```

Response Profile

Ordered Value	NR_eva	Total Frequency
1	1	1448
2	0	5947

Probability modeled is NR_eva=1.

Class Level Information

Class	Value	Design Variables
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1
tuetab1995_8	0	0
	1	1
natm20	0	0
	1	1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Response Profile

Ordered Value	NR_eva	Total Frequency
1	1	1448
2	0	5947

Probability modeled is NR_eva=1.

Class Level Information

Class	Value	Design Variables
Permer	0	0
	1	1
repfam3	0	0
	1	1
repjeune3	0	0
	1	1
tuetab1995_8	0	0
	1	1
natm20	0	0
	1	1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	7316.185	7108.816
SC	7323.094	7150.268
-2 Log L	7314.185	7096.816

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	217.3689	5	<.0001
Score	235.7503	5	<.0001
Wald	222.8143	5	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Permer	1	11.1714	0.0008
repfam3	1	19.0149	<.0001
repjeune3	1	82.3445	<.0001
tuetab1995_8	1	27.8162	<.0001
natm20	1	27.1516	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.5201	0.0739	423.0494	<.0001
Permer 1	1	-0.2524	0.0755	11.1714	0.0008
repfam3 1	1	0.3787	0.0869	19.0149	<.0001
repjeune3 1	1	0.6253	0.0689	82.3445	<.0001
tuetab1995_8 1	1	0.4279	0.0811	27.8162	<.0001
natm20 1	1	0.4721	0.0906	27.1516	<.0001

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
Permer	1 vs 0	0.777	0.670	0.901
repfam3	1 vs 0	1.460	1.232	1.731
repjeune3	1 vs 0	1.869	1.633	2.139
tuetab1995_8	1 vs 0	1.534	1.309	1.798
natm20	1 vs 0	1.603	1.343	1.915

Association of Predicted Probabilities and Observed Responses

Percent Concordant	50.2	Somers' D	0.237
Percent Discordant	26.5	Gamma	0.309
Percent Tied	23.3	Tau-a	0.075
Pairs	8611256	c	0.619
Pairs	37877190	c	0.759

5. Calcul de P_{sup} , la probabilité de ne pas avoir répondu à l'enquête SUP 2007/2008

```
%model(SASTEMP.database_pond2008,NR_sup,repjeune3 natm20 mentionbac permer,2);run;
```

Response Profile

Ordered Value	NR_sup	Total Frequency
1	1	316
2	0	4997

Probability modeled is NR_sup=1.

Class Level Information

Class	Value	Design Variables
repjeune3	0	0
	1	1
natm20	0	0
	1	1
mentionbac	0	0
	1	1
Permer	0	0
	1	1

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2398.432	2362.345
SC	2405.010	2395.235
-2 Log L	2396.432	2352.345

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	44.0870	4	<.0001
Score	52.2697	4	<.0001
Wald	49.2732	4	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
repjeune3	1	17.5713	<.0001
natm20	1	14.7310	0.0001
mentionbac	1	5.3710	0.0205
Permer	1	3.9239	0.0476

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.5441	0.1548	269.9440	<.0001
repjeune3	1	1	0.7500	0.1789	17.5713	<.0001
natm20	1	1	0.6721	0.1751	14.7310	0.0001
mentionbac	1	1	-0.3556	0.1535	5.3710	0.0205
Permer	1	1	-0.3213	0.1622	3.9239	0.0476

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
repjeune3	1 vs 0	2.117	1.491	3.006
natm20	1 vs 0	1.958	1.389	2.760
mentionbac	1 vs 0	0.701	0.519	0.947
Permer	1 vs 0	0.725	0.528	0.997

Association of Predicted Probabilities and Observed Responses

Percent Concordant	41.2	Somers' D	0.188
Percent Discordant	22.4	Gamma	0.296
Percent Tied	36.4	Tau-a	0.021
Pairs	1579052	c	0.594