MÉTHODES DE SONDAGES UTILISÉES DANS LES PROGRAMMES D'ÉVALUATIONS DES ÉLÈVES

Émilie Garcia, Marion Le Cam et Thierry Rocher
MENESR-DEPP, bureau de l'évaluation des élèves

Cet article porte sur les méthodes de sondages utilisées à la DEPP dans le cadre des dispositifs d'évaluations standardisées des acquis des élèves. Chaque année, plusieurs échantillons d'élèves sont tirés au sort pour passer ces évaluations. Des problématiques classiques du domaine des sondages se posent, concernant par exemple la définition du champ, les bases de sondage, les modalités de tirage, etc. qui doivent répondre à certaines contraintes pratiques. En outre, dans la mesure où plusieurs échantillons sont tirés à partir des mêmes bases, la question de la coordination de leur tirage doit être traitée. Dans un premier temps, nous présentons les choix faits en matière de méthode de sondage, à toutes les étapes, du tirage des échantillons au redressement de la non-réponse. Dans un second temps, nous conduisons plusieurs simulations qui visent à montrer l'intérêt d'utiliser des informations auxiliaires, c'est-à-dire disponibles pour l'ensemble des élèves. Ces informations peuvent être prises en compte lors du tirage, avec les méthodes d'équilibrage, ou lors du redressement de la non-réponse, avec les méthodes de calage sur marges. Nous montrons que les stratégies prenant en compte l'information auxiliaire, employées dans les évaluations nationales menées par la DEPP, améliorent la qualité des estimateurs, en comparaison d'autres stratégies telles que celles employées dans le cadre des évaluations internationales comme PIRLS ou PISA.

En France, chaque année, la direction de l'évaluation, de la prospective et de la performance (DEPP) conduit des programmes d'évaluation des acquis des élèves [Trosseille et Rocher, dans ce numéro, p. 15]. Il peut s'agir d'évaluations nationales comme les évaluations des compétences du socle commun ou les évaluations Cedre (Cycle des évaluations disciplinaires réalisées sur échantillons), mais également d'évaluations internationales telles que PISA (*Programme for International Student Assessment*) ou PIRLS (*Progress in International Reading Literacy Study*).

Ces évaluations sont réalisées sur des échantillons composés de plusieurs milliers d'élèves, le plus souvent scolarisés soit en fin de CM2, soit en fin de troisième. Le **tableau 1** liste les échantillons concernés pour l'année 2013-2014 et montre qu'au total, près de 80 000 élèves ont été échantillonnés cette année-là. Notre article se concentre sur les programmes d'évaluations des élèves, mais notons que la DEPP est amenée à tirer des échantillons pour répondre à d'autres problématiques, par exemple pour l'enquête nationale de climat scolaire et de victimation [Hubert, 2014].

Tableau 1 **Échantillons des programmes d'évaluations des élèves 2014**

	Établissements	Élèves
Primaire		
Cedre mathématiques	290	7 952
Cedre maîtrise de la langue	175	4 532
LSE (lecture sur support électronique)	386	9 932
Socle CE1 compétences 1 et 3	628	20 160
TIMSS CM1 - expérimentation	40	1 280
Total	**1 519**	**43 856**
Secondaire		
Cedre mathématiques	323	8 026
Cedre compétences générales et langagières – expérimentation	179	4 551
LSE (lecture sur support électronique)	315	8 070
Socle sixième – expérimentation	391	10 071
PISA – expérimentation	55	2 360
TIMSS *Advanced* – expérimentation	32	1 821
Total	**1 295**	**34 899**

Lecture : en 2014, l'échantillon Cedre mathématiques au primaire comptait 7 952 élèves répartis dans 290 écoles.

L'objet de cet article est double : descriptif et analytique. Tout d'abord, il s'agit de présenter les méthodes de sondage utilisées dans le cadre de ces programmes d'évaluation. Les problématiques soulevées concernent des aspects très divers du domaine des sondages : définition des plans de sondage, modalités de tirage, gestion du recouvrement des échantillons, redressement de la non-réponse, calcul de précision, etc. Des choix ont été opérés pour chacune de ces étapes, en fonction d'un ensemble de contraintes, et font l'objet d'une description précise dans la première partie de cet article. Dans un second temps, nous analysons dans quelle mesure l'échantillonnage peut profiter des informations disponibles dans les bases de données de la DEPP, pour tous les élèves scolarisés en France. La prise en compte de cette information, appelée information auxiliaire, permet en théorie d'améliorer la qualité du tirage ainsi que du redressement de la non-réponse. L'information auxiliaire peut être utilisée en amont au moment du tirage des échantillons (échantillon équilibré) mais aussi en aval pour le redressement de la non-réponse (calage sur marges). Les méthodes de sondage appliquées dans les évaluations internationales ne font pas appel à de l'information auxiliaire du fait de la trop grande diversité des informations disponibles dans chaque pays. Nous vérifions empiriquement l'intérêt de prendre en compte l'information auxiliaire, au moyen de simulations qui portent sur chacune des étapes du sondage : tirage de l'échantillon, redressement de la non-réponse, calcul de précision.

PLANS DE SONDAGE

Pour un échantillon donné, la définition du plan de sondage découle d'un ensemble de contraintes liées aux objectifs de l'évaluation, à la précision recherchée et aux coûts induits. En outre, les informations disponibles dans les bases de sondage impliquent également certaines contraintes.

Champ et exclusions

De manière générale, il est possible de distinguer au moins trois possibilités pour définir la population-cible d'une évaluation :
– les élèves d'un niveau scolaire donné (par exemple, Cedre avec les niveaux CM2 et troisième) ;
– les élèves entrant dans un niveau (par exemple, le panel d'élèves entrant en sixième) ;
– les élèves d'un âge donné (par exemple, l'évaluation PISA avec les élèves de 15 ans quel que soit leur niveau scolaire).
Dans le premier degré, le champ des évaluations comprend les écoles publiques et privées sous contrat en France métropolitaine et DOM, à l'exception du cycle Cedre qui concerne uniquement la France métropolitaine. Sont toujours exclues du champ les écoles des COM, les écoles privées hors contrat, les écoles à l'étranger, les écoles spécialisées. Enfin, pour des raisons de coût, les écoles de moins de 6 élèves du niveau scolaire visé sont exclues de la base de sondage. En guise d'illustration, en 2014, pour l'évaluation des compétences du socle en CE1, il y avait environ 2 500 écoles qui accueillaient moins de 6 élèves de CE1. Ces écoles représentent près de 7,5 % de l'ensemble des écoles, mais elles accueillent moins de 1 % des élèves.
Dans le second degré, les établissements publics et privés sous contrat en France métropolitaine et DOM constituent le champ. Sont toujours exclus les établissements des COM, les établissements privés hors contrat, les établissements à l'étranger, les EREA (Établissements régionaux d'enseignement adapté). Comme pour le premier degré, les évaluations Cedre ne visent que la France métropolitaine. En outre, PISA n'interroge ni La Réunion ni Mayotte, en raison d'une différence de calendrier scolaire.

Les bases de sondage

La base de sondage est la base de données dans laquelle sont tirés les échantillons. Pour les échantillons relevant du second degré, la base de sondage utilisée est la base dite Scolarité construite par la DEPP. C'est une base de données individuelles et anonymes contenant de nombreuses informations sur les élèves scolarisés une année scolaire donnée (date de naissance, profession et catégorie socioprofessionnelle des parents, etc.). Nous disposons également d'informations sur les établissements scolaires (le secteur d'enseignement, par exemple). Ces informations, qualifiées de variables auxiliaires, peuvent être utilisées au moment du tirage des échantillons, pour définir les variables de stratification. Dans le premier degré, le système d'information est nettement moins développé et les plans de sondage s'appuient sur des données qui concernent les écoles, pas les élèves.

Modalités de tirage

Très majoritairement, les plans de sondage sont à deux degrés : un premier degré qui concerne le tirage d'établissements scolaires ou de classes ; un second degré qui concerne le tirage des élèves eux-mêmes. Ce type de sondage est soumis à ce qu'on appelle des effets de grappe : concrètement, les élèves d'un même établissement ou d'une même classe ont tendance à avoir des caractéristiques communes. Ainsi, la variabilité totale va dépendre de la variabilité entre établissements ou entre classes et moins de la variabilité entre élèves. Ce phénomène a une conséquence directe : à effectif égal, un sondage par grappe est moins précis qu'un sondage aléatoire simple qui vise à tirer directement les élèves dans une liste.
Pour le primaire, les écoles sont sélectionnées puis tous les élèves du niveau scolaire visé sont évalués. Il s'agit donc d'un sondage par grappe. Le nombre d'élèves d'un niveau s'élève en moyenne à 25 et ne dépasse que rarement 40 élèves, ce qui limite les effets de grappe.
Pour le secondaire, deux options sont considérées : soit un sondage par grappe en sélectionnant un échantillon de classes et où tous les élèves des classes tirées au sort participent à l'évaluation ; soit un premier degré qui concerne les établissements puis un second degré où un nombre d'élèves fixe dans chaque établissement est sélectionné[1]. Les programmes nationaux suivent la première option tandis que l'évaluation PISA suit la seconde [OCDE, 2012].

Le choix de sondages par grappe est motivé par la facilité de gestion. En effet, le fait de sélectionner tous les élèves d'un niveau scolaire donné au primaire ou tous les élèves d'une classe au collège permet d'éviter de mettre en place des procédures de tirage au sort d'élèves une fois les établissements tirés. Comme la DEPP ne dispose pas de la liste nominative des élèves, et pour être certain que le tirage au sort soit respecté dans les établissements, dans le cadre de PISA, les établissements sélectionnés envoient à la DEPP la liste des élèves de 15 ans, à charge pour la DEPP de sélectionner un nombre fixe d'élèves parmi eux.

Stratification et choix de l'allocation

De manière générale, les échantillons sont stratifiés avec une allocation proportionnelle à la répartition selon la zone de scolarisation, à savoir : public hors éducation prioritaire ; éducation prioritaire ; privé. Cela signifie que le tirage de l'échantillon est réalisé de manière à respecter la répartition des élèves selon la zone de scolarisation de l'établissement scolaire, telle qu'observée sur l'ensemble de la population-cible. Le nombre d'établissements (ou de classes) alloués à chaque zone est ainsi déterminé selon ce principe.

Pour les évaluations des compétences du socle commun, qui alimentent les indicateurs de résultats de la LOLF, un indicateur est demandé pour chacune des quatre strates suivantes : public hors éducation prioritaire ; réseau de réussite scolaire (RRS) ; écoles, collèges et lycées pour l'ambition, l'innovation et la réussite (Éclair) ;

1. Dans ce second cas, les établissements sont tirés au sort proportionnellement à leur taille (nombre d'élèves). En effet, une fois que les établissements sont échantillonnés, un nombre fixe d'élèves est alors sélectionné quel que soit l'établissement. Par conséquent, les élèves des grands établissements ont moins de chance d'être tirés au sort que les élèves des petits établissements. Le tirage proportionnel à la taille permet ainsi de rétablir l'égalité des probabilités de tirage.

privé. Pour ces évaluations, les strates de l'éducation prioritaire vont donc être « sur-représentées » afin de garantir une précision suffisante de nos estimateurs. Par exemple, les élèves des écoles Éclair peuvent représenter environ un quart des élèves de l'échantillon, alors qu'ils représentent en réalité environ 5 % de l'ensemble des élèves.
Le nombre d'élèves sélectionnés est établi en fonction du coût de l'évaluation et du degré de précision attendu. Par exemple, dans le cadre des indicateurs de la LOLF, une précision de 2 points de pourcentage est demandée concernant la proportion d'élèves maîtrisant les compétences du socle commun. L'élaboration du plan de sondage est préparée en fonction de cette attente.

TIRAGE DES ÉCHANTILLONS

De manière classique, on note U la population visée par une évaluation donnée, Y la variable d'intérêt (typiquement le score à l'évaluation, ou bien une indicatrice de difficulté), X une variable auxiliaire, c'est-à-dire connue pour l'ensemble des élèves de la population U. Un échantillon S d'élèves est sélectionné dans la population U. Chaque élève i a la probabilité π_i d'être sélectionné dans l'échantillon S (probabilité d'inclusion). Enfin, les poids de sondages, définis comme les inverses des probabilités d'inclusion π_i, sont notés d_i.

Tirage équilibré

Un échantillon équilibré est un échantillon qui est représentatif de la population au regard de certaines variables auxiliaires. Cela signifie que dans un échantillon équilibré, l'estimateur du total d'une variable auxiliaire X sera exactement égal au vrai total de la variable X dans la population.

Cette propriété s'écrit[2] :

$$\sum_{i \in S} \frac{X_i}{\pi_i} = \sum_{i \in U} X_i \qquad (1)$$

Les échantillons équilibrés ont donc comme propriété de fournir une photographie parfaite de la population, au regard des variables auxiliaires connues, ce que ne garantit pas une procédure aléatoire simple d'échantillonnage. En théorie, ils permettent également d'améliorer la précision des estimateurs s'il existe un lien entre la variable d'intérêt et les variables auxiliaires.

Le tirage équilibré est réalisé grâce au programme CUBE développé par l'Insee et mis à disposition sous forme de macros SAS. La documentation complète est disponible sur le site Internet de l'Insee [Rousseau et Tardieu, 2004]. L'algorithme CUBE permet de choisir de manière aléatoire un échantillon parmi tous les échantillons possibles

2. Le terme de gauche de l'équation (1) représente l'estimateur du total, dit estimateur de Horitz-Thompson. En outre, l'indice i peut représenter ici aussi bien les élèves que les établissements ou les classes, en considérant les sous-totaux par unités comme variables auxiliaires.

respectant les contraintes reposant sur les variables auxiliaires. Il se déroule en deux phases : une « phase de vol » et une « phase d'atterrissage ». Durant la phase de vol, toutes les contraintes sont respectées. Elle se termine si un échantillon équilibré de manière parfaite est trouvé ou s'il n'est pas possible de trouver un échantillon en respectant toutes les contraintes. Si la phase de vol n'a pas abouti à un échantillon, la phase d'atterrissage débute. Elle consiste au relâchement des contraintes et au choix optimal de l'échantillon selon le critère choisi par l'utilisateur (ordre de priorité sur les contraintes, relâchement de la contrainte avec un coût minimal sur l'équilibrage ou garantie d'un échantillon de taille fixe).

Afin de conduire un tirage équilibré, il est nécessaire de disposer d'informations auxiliaires pour l'ensemble des élèves. Concernant les évaluations réalisées à l'école primaire, comme nous l'avons signalé précédemment, les bases de sondage contiennent très peu d'informations. C'est pourquoi les échantillons sont simplement stratifiés selon la zone de scolarisation de l'école, mais ils ne sont pas équilibrés sur d'autres variables.
En revanche, dans le secondaire, de nombreuses informations sont disponibles. Nous réalisons un tirage équilibré des établissements (ou des classes) selon les variables suivantes :
– l'effectif : le nombre total d'élèves du niveau visé dans la population (typiquement, les élèves de troisième) ;
– l'indice de position sociale, qui consiste en une transformation de la PCS des parents des élèves, tel que défini par Le Donné et Rocher [2010] : la somme de cet indice par établissements (ou par classes) est retenue comme variable d'équilibrage ;
– le retard scolaire : le nombre d'élèves en retard du niveau visé ;
– le sexe : le nombre de filles du niveau visé.

Non-recouvrement entre les échantillons

Chaque année, plusieurs programmes d'évaluation étant conduits au même moment, certains établissements peuvent être échantillonnés pour participer à plusieurs évaluations. Afin d'alléger la charge de travail des établissements, et pour assurer un taux de retour maximal, une procédure de non-recouvrement des échantillons est employée. Au moment du tirage d'un échantillon, les écoles ou les collèges dont une classe a déjà été sélectionnée pour une autre évaluation la même année sont exclus de la base de sondage. Les probabilités d'inclusion doivent donc être recalculées pour tenir compte de ces exclusions tout en gardant une représentativité nationale. En outre, il convient d'adapter la procédure de tirage équilibré à cette contrainte de non-recouvrement. L'**encadré** précise la méthode retenue pour corriger les probabilités de sélection.

REDRESSEMENT DE LA NON-RÉPONSE : CALAGE SUR MARGES

Comme toute enquête réalisée par sondage, les évaluations des élèves sont exposées à la non-réponse. Cependant, les taux de participation sont généralement très satisfaisants. Par exemple, pour l'évaluation Cedre histoire-géographie au collège, 96,5 % des établissements ont participé, et au final le taux de réponse des élèves s'est élevé à 90 %.

TIRAGE ÉQUILIBRÉ APRÈS ÉLIMINATION DE LA BASE DES ÉCHANTILLONS PRÉCÉDEMMENT TIRÉS

La situation est la suivante : un échantillon d'établissements a été sélectionné pour participer à une évaluation ; un deuxième échantillon doit être tiré pour une autre évaluation. Nous souhaitons éviter que des établissements soient interrogés deux fois. Il s'agit donc de gérer le non-recouvrement entre les échantillons et d'assurer également un tirage équilibré du deuxième échantillon.

Nous nous concentrons ici sur le non-recouvrement des échantillons, mais notons qu'une approche plus générale incluant un taux de recouvrement non nul (pour permettre des analyses croisées entre enquêtes) est en cours de développement avec une application à des données issues d'évaluations standardisées [Christine et Rocher, 2012].

Formulation du problème et notations

Un échantillon S_1 a été tiré. Il est connu et les probabilités d'inclusion des établissements π_j^1 sont également connues.

On souhaite alors tirer un échantillon S_2 dans la population U avec les probabilités π_j^2, mais sans aucun recouvrement avec l'échantillon S_1. On va donc tirer l'échantillon S_2 dans la population $U(S_1)$, c'est-à-dire la population U privée des établissements de l'échantillon S_1 qui appartiennent à U. Notons d'emblée que S_1 n'a pas nécessairement été tiré dans U, mais potentiellement dans une autre population, plus large ou plus réduite ; cela n'affecte en rien la formulation envisagée ici.

Notons également que l'indice j est utilisé ici : il concerne les établissements et non les élèves, représentés par l'indice i.

Il s'agit donc de procéder à un tirage conditionnel. On note π_j^{2/S_1} les probabilités d'inclusion conditionnelles des établissements dans le second échantillon S_2, sachant que le premier échantillon est connu.

Ces probabilités conditionnelles peuvent s'écrire :

$$\pi_j^{2/S_1} = \begin{cases} \lambda_j \text{ si } j \notin S_1 \\ 0 \text{ si } j \in S_1 \end{cases}, \text{ avec } \lambda_j \in [0,1]$$

On a $\pi_j^2 = E(\pi_j^{2/S_1}) = \lambda_j(1-\pi_j^1)$, d'où : $\lambda_j = \dfrac{\pi_j^2}{1-\pi_j^1}$

Équilibrage

On souhaite maintenant que l'échantillon S_2 soit équilibré selon certaines variables (nombre d'élèves en retard, etc.). Soit X une variable d'équilibrage, la condition s'écrit :

$$\sum_{j\in S_2} \frac{X_j}{\pi_j^2} = \sum_{j\in U} X_j$$

Pour arriver à ce résultat, le principe est de tirer S_2 dans $U(S_1)$ avec les probabilités d'inclusion λ_j et avec une condition d'équilibrage sur la variable $X_j / (1-\pi_j^1)$.

Ainsi, on aura :

$$\sum_{j\in S_2} \frac{X_j}{\pi_j^2} = \sum_{j\in S_2} \frac{X_j}{\lambda_j(1-\pi_j^1)} = \sum_{j\in U(S_1)} \frac{X_j}{(1-\pi_j^1)}$$

Or, en espérance, on a :

$$E\left(\sum_{j\in U(S_1)} \frac{X_j}{(1-\pi_j^1)}\right) = E\left(\sum_{j\in U} \frac{X_j}{(1-\pi_j^1)} I_{j\notin S_1}\right) = \sum_{j\in U} X_j$$

La condition d'équilibrage initiale est donc remplie.

Condition fondamentale

Comme il s'agit d'une probabilité, la condition fondamentale est que $\lambda_j \in [0,1]$.

Comme $\lambda_j = \dfrac{\pi_j^2}{1-\pi_j^1}$, la condition est en fait que :

$$\pi_j^1 + \pi_j^2 \leq 1$$

Dans certains cas, par exemple, des strates souvent sur-représentées comme les établissements situés dans des zones spécifiques concernant peu d'élèves (ex. : Éclair), cette condition pourrait ne pas être satisfaite. Cependant, de façon concrète, la condition a toujours été respectée dans les plans de sondage réalisés.

Bien que ces taux de retour soient élevés, il est nécessaire de tenir compte de la non-réponse dans les estimations, car celle-ci n'est pas purement aléatoire (par exemple, la non-réponse est plus élevée chez les élèves « en retard »). Afin de la prendre en compte, un calage sur marges est effectué à l'aide de la macro CALMAR, également disponible sur le site Internet de l'Insee. La méthode de calage sur marges consiste à modifier les poids de sondage d_i des répondants de manière à ce que l'échantillon ainsi repondéré soit représentatif de certaines variables auxiliaires dont on connaît les totaux sur la population [Sautory, 1993]. C'est une méthode qui permet de corriger la non-réponse, mais également d'améliorer la précision des estimateurs. En outre, elle a pour avantage de rendre cohérents les résultats observés sur l'échantillon pour ce qui concerne des informations connues sur l'ensemble de la population.
Les nouveaux poids w_i, calculés sur l'échantillon des répondants S', vérifient l'équation suivante pour les K variables auxiliaires sur lesquelles porte le calage :

$$\forall k = 1...K, \sum_{i \in S'} w_i X_i^k = \sum_{i \in U} X_i^k \qquad (2)$$

Ils sont obtenus par minimisation de l'expression $\sum_{i \in S'} d_i G(\frac{w_i}{d_i})$ où G désigne une fonction de distance, sous les contraintes définies dans l'équation (2).
Pour le redressement de la non-réponse, seules les marges des variables sont nécessaires, c'est-à-dire leur somme sur l'ensemble de la population. À l'école comme au collège, les variables auxiliaires de calage utilisées sont :
– le nombre total d'élèves du niveau visé dans la population ;
– le nombre d'élèves du niveau visé en retard dans la population ;
– le nombre de garçons du niveau visé dans la population.
Au collège, le nombre d'élèves de classe sociale défavorisée est également introduit.

CALCUL DE PRÉCISION

Les résultats des évaluations sont soumis à une variabilité qui dépend notamment des erreurs d'échantillonnage. Il est possible d'estimer statistiquement ces erreurs d'échantillonnage et de produire des intervalles de confiance sur les différents estimateurs calculés.

On note Y la variable d'intérêt (typiquement le score obtenu à une évaluation) et $\hat{\bar{Y}}$ l'estimateur de la moyenne de Y, qui constitue un estimateur essentiel sur lequel nous insistons dans la suite, bien que d'autres soient également au centre des analyses, comme ceux concernant la dispersion. La méthode retenue est cependant applicable à différents types d'estimateurs.
Nous souhaitons estimer la variance de cet estimateur, c'est-à-dire $V(\hat{\bar{Y}})$. En absence de formule théorique pour calculer $V(\hat{\bar{Y}})$, il existe plusieurs procédures permettant de l'estimer, c'est-à-dire de calculer $\hat{V}(\hat{\bar{Y}})$, l'estimateur de la variance d'échantillonnage. Il peut s'agir de méthodes de linéarisation des formules (Taylor) ou bien de méthodes empiriques (méthodes de réplication, jackknife, etc.). Ces méthodes sont bien décrites dans la littérature. Le lecteur est invité à consulter Tillé [2001] ou Ardilly [2006].

Cependant, lorsqu'un calage sur marges a été effectué, il faut en tenir compte pour le calcul de la précision. Dans ce cas, la variance de $\hat{\bar{Y}}$ est asymptotiquement équivalente à la variance des résidus de la régression de la variable d'intérêt sur les variables de calage [Deville et Särndal, 1992]. En pratique, pour estimer la variance d'échantillonnage de $\hat{\bar{Y}}$, tenant compte du calage effectué, il convient alors d'appliquer la procédure suivante :
1 – On effectue la régression linéaire de la variable d'intérêt sur les variables de calage, en pondérant par les poids initiaux. Les résidus e_i de cette régression sont calculés.
2 – Les valeurs $g_i e_i$ sont calculées, où g_i représente le rapport entre les poids CALMAR (w_i) et les poids initiaux (d_i) : $g_i = \frac{w_i}{d_i}$.
3 – La variance d'échantillonnage de $\hat{\bar{Y}}$ est alors obtenue en calculant la variance d'échantillonnage de $g_i e_i$.

LES AVANTAGES DE L'UTILISATION DE L'INFORMATION AUXILIAIRE : TROIS SIMULATIONS

Comme nous l'avons vu en première partie, dans le second degré, la DEPP dispose de bases de sondage avec de l'information disponible sur l'ensemble de la population. L'information auxiliaire est utilisée à chaque étape, du tirage des échantillons au calcul de précision. Cette approche est assez spécifique aux programmes nationaux d'évaluation. En effet, les évaluations internationales, telles que PISA, empruntent une autre voie, qui n'utilise que très peu l'information auxiliaire. Les responsables de ces enquêtes avancent que la disponibilité d'informations auxiliaires est très variable selon les pays et que les variables elles-mêmes sont différentes selon les pays, ce qui nécessiterait une investigation coûteuse pour définir les méthodes, car personnalisée pour chaque pays.

Dans cette partie, nous souhaitons cependant montrer l'intérêt de recourir à l'information auxiliaire, ainsi que cela est fait dans les programmes nationaux d'évaluation. Trois simulations sont présentées : la première concerne le tirage équilibré des échantillons, la seconde le redressement de la non-réponse *via* un calage sur marges et la troisième le calcul de précision. Ces simulations sont menées de manière à évaluer les deux types de procédures : celle utilisée dans les évaluations internationales (tirage non équilibré, redressement par ajustement, calcul simple de précision) et celle utilisée dans les programmes nationaux tels que Cedre (tirage équilibré, redressement par calage sur marges, calcul de précision tenant compte du calage). Ainsi, les simulations ne sont pas envisagées de manière indépendante, mais dans l'optique d'une comparaison globale des deux approches.
Les simulations ont été réalisées sur la base exhaustive des notes obtenues à l'examen terminal du brevet en 2009 en mathématiques, français et histoire-géographie. Dans cette base, nous disposons aussi des caractéristiques des élèves

(sexe, année de naissance, PCS des parents) et des établissements. Ces informations serviront de variables auxiliaires. Les notes fournissent un *proxy* intéressant par rapport aux scores obtenus aux évaluations, car portant sur des objets proches en principe. Nous estimons ainsi qu'en matière de sondage, les deux formes d'évaluation – scores aux évaluations et notes à l'examen – devraient conduire aux mêmes types de conclusion. Pour nos simulations, les variables d'intérêt sont donc les notes à l'examen, à défaut des scores obtenus aux évaluations standardisées.

Simulation 1 : impact du tirage équilibré

Pour montrer les avantages de l'utilisation de l'information auxiliaire lors du tirage des échantillons, nous avons comparé quatre stratégies d'échantillonnage :
1 – sondage à probabilités proportionnelles à la taille sur les établissements, puis tirage aléatoire simple de 30 élèves dans chacun des établissements ;
2 – sondage à probabilités proportionnelles à la taille et équilibré des établissements, puis tirage aléatoire simple de 30 élèves dans chacun des établissements ;
3 – tirage aléatoire simple de classes, puis tous les élèves des classes tirées au sort participent à l'évaluation ;
4 – tirage équilibré de classes, puis tous les élèves des classes tirées au sort participent à l'évaluation.

Le premier plan de sondage est celui utilisé pour PISA tandis que le dernier est celui utilisé à la DEPP. Les plans 2 et 4 sont les versions équilibrées des plans 1 et 3. Tous les échantillons sont stratifiés selon le secteur d'enseignement : public hors éducation prioritaire, RAR (réseaux ambition réussite), RRS (réseaux de réussite scolaire) et privé. Dans chacune des quatre strates, 2 000 élèves sont sélectionnés. En effet, il est souvent demandé un indicateur pour chacune de ces strates. Les variables d'intérêt sont les notes au brevet obtenues en français, en mathématiques et en histoire-géographie ; les paramètres d'intérêt sont les moyennes et les pourcentages d'élèves obtenant une note inférieure à un certain seuil.
Pour les plans de sondage 2 et 4, les variables auxiliaires utilisées pour l'équilibrage sont :
– le sexe ;
– le fait d'être en retard ;
– l'indice social moyen de l'établissement (en quatre groupes selon les quartiles).

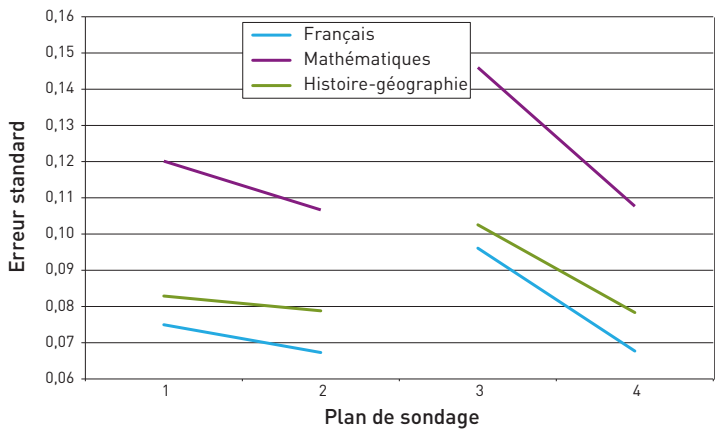
Pour chaque plan de sondage, 1 000 échantillons ont été tirés. Pour chaque échantillon s on calcule la moyenne des notes $\widehat{\bar{y}_s}$, ensuite on calcule la moyenne des $\widehat{\bar{y}_s}$ ainsi que l'écart-type des $\widehat{\bar{y}_s}$ qui correspond à l'erreur standard, c'est-à-dire à la racine carrée de la variance d'échantillonnage.
Le **tableau 2** présente les notes moyennes estimées pour chacun des plans de sondage et les compare avec la note moyenne réelle de l'ensemble de la population. Pour chaque simulation, on retrouve la note moyenne observée dans la population. Dans le cadre des plans de sondage employés, la moyenne simple sur l'échantillon est en effet un estimateur sans biais de la moyenne sur la population. On observe de légers écarts (à la deuxième décimale) pour les proportions d'élèves ayant obtenu une note inférieure à un certain seuil.
Si les biais sont très faibles, voire nuls, la précision est variable selon les plans de sondage considérés. Ainsi, comme le montre la **figure 1**, l'erreur standard est plus

Tableau 2 Simulation 1 – notes moyennes selon les plans de sondage

		Population	Plans de sondage			
			1	2	3	4
Français		11,41	11,41	11,41	11,42	11,41
Mathématiques		9,68	9,68	9,68	9,68	9,68
Histoire-géographie		10,74	10,74	10,74	10,74	10,74
Français < 8	(en %)	13,82	13,82	13,85	13,80	13,82
Mathématiques < 6	(en %)	21,49	21,52	21,52	21,51	21,48
Histoire-géographie < 8	(en %)	20,46	20,45	20,50	20,42	20,44

Lecture : la note moyenne des élèves en français est de 11,41 sur l'ensemble de la population.
La moyenne des 1 000 estimations de la note moyenne en français pour le plan de sondage 1 est aussi de 11,41.
Le pourcentage d'élèves ayant une note de français inférieure à 8 est au total de 13,82 % ; les quatre plans de sondage donnent des valeurs proches en moyenne.

Figure 1 Simulation 1 – erreurs standard selon les plans de sondage



Lecture : en mathématiques, l'erreur standard de la note est d'environ 0,12 pour le premier plan de sondage et elle est inférieure à 0,11 pour le deuxième.

faible dans le cas d'un sondage équilibré (plans 2 et 4). Dès lors que l'on utilise de l'information auxiliaire pour réaliser un tirage équilibré, la précision est nettement améliorée. Ainsi, il apparaît que l'enquête PISA gagnerait à adopter une démarche de tirage équilibré (plan 2 par rapport au plan 1 de PISA). Ce constat est encore plus marqué dans le cas où le tirage au premier degré concerne des classes et pas des établissements : le fait d'échantillonner des classes au lieu d'établissements dégrade la précision, en l'absence de tirage équilibré (plan 3 en comparaison du plan 1). Ce phénomène est à relier au fait que l'effet de grappe est plus important avec un tirage de classes qu'avec un tirage d'établissements puis d'élèves. En revanche, dès lors que l'on utilise l'information auxiliaire disponible, le mode de tirage (établissements ou classes) conduit à des résultats comparables du point de vue de la précision (plans 2 et 4). Ce point est important pour les aspects logistiques gérés par la DEPP qui sont moins coûteux dans le cadre d'un tirage de classes entières plutôt que d'un tirage d'établissements puis d'élèves.

Simulation 2 : impact du calage sur marge

Dans cette partie, au-delà du choix de tirage des échantillons, nous comparons également deux stratégies de repondération en présence de non-réponse. La première, utilisée dans le cadre des évaluations internationales, est basée sur des coefficients d'ajustement concernant les établissements et les élèves, de manière à ce que les répondants représentent les non-répondants. La seconde approche, utilisée à la DEPP, emploie une procédure de calage sur marges qui consiste à modifier les poids de sondage des élèves de manière à ce que l'échantillon soit représentatif de la population au regard de certaines variables auxiliaires choisies.

Plus précisément, concernant la première approche, nous avons repris la démarche suivie dans les évaluations internationales PIRLS [Martin, Mullis, Kennedy, 2007] ou PISA [OCDE, 2012], et également appliquée sur les premières évaluations du cycle Cedre. Il s'agit d'appliquer aux poids de sondage des coefficients d'ajustement. Pour un élève i d'un l'établissement j, nous définissons les nouveaux poids de la manière suivante[3] :

$$w_{ij}^{1} = f_{1j} f_{2ij} d_{2ij} d_{1j} \qquad (3)$$

avec :

– d_{1j} le poids de sondage initial de l'établissement j ;

– d_{2ij} le poids de sondage initial de l'élève i au sein de l'établissement j ;

– f_{1j} le coefficient d'ajustement pour la non-réponse des établissements (rapport entre le nombre d'établissements de la strate échantillonnés au départ et le nombre d'établissements répondants de la strate) ;

– f_{2ij} le coefficient d'ajustement de la non-réponse des élèves au sein des établissements répondants (rapport entre le nombre d'élèves échantillonnés de l'établissement et le nombre d'élèves répondants), distingué selon le sexe.

La deuxième approche repose sur un calage sur marges selon les totaux des distributions suivantes :

– le nombre d'élèves dans chaque strate ;

– la répartition par sexe dans la population ;

– le nombre d'élèves en retard dans la population ;

– le nombre d'élèves dans chaque quartile de la population découpée grâce à l'indice social.

Afin de comparer ces deux stratégies dans le cadre d'une nouvelle simulation, nous utilisons la même base de sondage que pour la simulation 1. Notre simulation procède selon les étapes suivantes :

1 – tirage d'un échantillon ;

2 – génération de non-réponse ;

3 – repondération des élèves répondants.

Concernant le **tirage des échantillons**, nous avons repris deux plans de sondage utilisés précédemment. Le premier, celui de PISA, est un sondage proportionnel à la taille des établissements puis une sélection aléatoire de 30 élèves dans chacun des établissements (plan de sondage 1 de la simulation précédente). Le second plan de

3. C'est une version simplifiée de ce qui est fait dans PIRLS et PISA où les coefficients d'ajustement sont plus nombreux, mais la démarche est similaire.

sondage est un échantillon équilibré sur les établissements puis sélection de 30 élèves dans chacun des établissements (plan de sondage 2 de la simulation précédente). Nous n'avons pas retenu le plan de sondage 4 utilisé à la DEPP car il est fondé sur un échantillon de classes et non d'établissements. Ainsi, les modalités de modélisation de la non-réponse n'auraient pas été comparables. En outre, nous avons pu observer que les plans de sondage 2 et 4 étaient assez proches en termes de précision. Au final, les résultats de la simulation 2 nous renseigneront sur l'impact de la procédure de repondération mais dans deux cadres de tirage différents, l'un équilibré et l'autre pas. Cette simulation nous permet ainsi de nous prononcer sur une démarche globale intégrant le tirage et la repondération.

S'agissant de la **génération de non-réponse**, nous avons au préalable modélisé la non-réponse, sachant qu'elle comporte deux types : celle des établissements scolaires échantillonnés et celle des élèves au sein des établissements répondants. Ces non-réponses ont été modélisées à l'aide de régressions logistiques appliquées à des données réelles de programmes d'évaluation.
Pour caractériser la non-réponse des établissements, nous avons utilisé l'évaluation Cedre compétences générales de 2009 en fin de troisième. Pour cette évaluation, les établissements qui n'ont pas répondu sont connus, en particulier nous disposons pour chacun d'entre eux de la note moyenne au brevet et de l'indice moyen de position socio-scolaire. Le modèle retenu prédit la non-réponse des établissements en fonction de ces deux variables.
Pour caractériser la non-réponse des élèves au sein des établissements répondants, nous avons dû utiliser une autre opération : l'évaluation Cedre histoire-géographie de 2012 en fin de troisième. Pour cette évaluation, nous disposons d'informations sur les élèves non-répondants, en particulier le sexe, le retard scolaire et l'origine sociale. La seule variable significative du modèle explicatif de la non-réponse est le retard scolaire, variable retenue pour la génération de non-réponse des élèves, qui a consisté à tirer au sort des élèves non-répondants, en distinguant les élèves, « en retard » des élèves « à l'heure ».

Pour chaque **stratégie de repondération**, 1 000 échantillons ont été tirés. La première stratégie est envisagée à partir du premier plan de sondage (PISA) et la seconde stratégie est appliquée au deuxième plan de sondage (PISA équilibré). Pour chaque échantillon de chaque stratégie, on calcule la moyenne des notes aux épreuves finales du brevet. Au final, le **tableau 3 p. 114** présente la moyenne de ces moyennes. Il ressort que la première stratégie, fondée sur les coefficients d'ajustement, apparaît légèrement biaisée, surtout pour les fractiles, c'est-à-dire les indicateurs concernant le pourcentage d'élèves en-deçà d'un certain seuil de note.

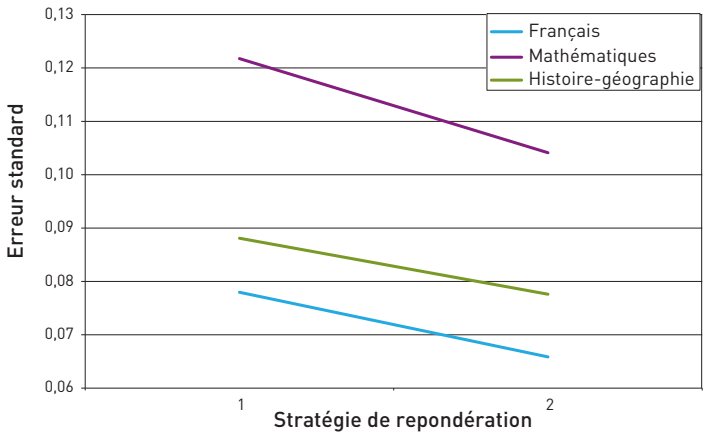
Au-delà du biais, nous avons également étudié la précision respective de ces deux stratégies. Nous avons ainsi calculé l'écart-type des 1 000 moyennes des échantillons simulés, c'est-à-dire l'erreur standard de la note moyenne. Il ressort que l'erreur standard est plus faible dans le cas d'un sondage équilibré et d'un redressement de la non-réponse avec un calage sur marges ▸ **Figure 2 p. 114**. Notons que la stratégie 2 conduit à des niveaux de précision comparables à ceux observés dans le cas de sondages équilibrés sans non-réponse (plans 2 et 4 de la figure 1). La stratégie 2 corrige donc la non-réponse de manière très satisfaisante tandis que la stratégie 1 est moins efficace.

Tableau 3 Simulation 2 – moyennes après repondération

Notes		Population	Stratégies	
			1	2
Français		11,41	11,44	11,41
Mathématiques		9,68	**9,78**	9,70
Histoire-géographie		10,74	**10,79**	10,75
Français < 8	(en %)	13,82	**13,54**	13,84
Mathématiques < 6	(en %)	21,49	**20,77**	**21,42**
Histoire-géographie < 8	(en %)	20,46	**20,02**	20,45

Lecture : la note moyenne des élèves en français est de 11,4. La moyenne des 1 000 estimations de la note moyenne en français pour la stratégie 1 est de 11,44.

Figure 2 Simulation 2 – erreurs standard selon la stratégie de repondération



Lecture : en mathématiques, l'erreur standard de la note est d'environ 0,12 pour la première stratégie de repondération et elle est supérieure à 0,10 pour la seconde.

Ces écarts peuvent s'expliquer par le fait que dans le cadre de la première stratégie, le calcul des coefficients d'ajustement repose sur l'hypothèse que les répondants et les non-répondants sont « similaires », s'agissant des élèves au sein d'un établissement, ou des établissements au sein d'une strate. Cette hypothèse a pu induire un biais dans l'estimation, dans la mesure où certaines variables individuelles, comme le retard scolaire, expliquent la non-réponse, quels que soient les établissements.

Quoi qu'il en soit, la comparaison des résultats de la simulation 1 avec ceux de la simulation 2 conduit à souligner l'importance du redressement de la non-réponse. En effet, la stratégie 2 donne des niveaux de précision du même ordre que ceux du plan 2 de la simulation 1, ce qui montre que le calage sur marges corrige parfaitement la non-réponse. En revanche, la méthode de redressement de la stratégie 1 est moins efficace et conduit à des niveaux de précision moins élevés.

Simulation 3 : impact de l'utilisation de l'information auxiliaire sur le calcul de la précision

Dans cette dernière section, nous nous intéressons aux procédures de calcul de précision. En effet, la publication des résultats des évaluations standardisées est accompagnée d'une estimation des erreurs liées à l'échantillonnage. Plus précisément, nous calculons l'erreur standard – soit la racine carrée de la variance d'échantillonnage – des différents paramètres d'intérêt, tels que le score moyen. De nombreuses méthodes de calcul existent. Dans le cadre des évaluations standardisées, les méthodes employées sont de nature empirique, car il peut être délicat de déterminer la formule théorique de l'erreur standard, au vu des plans de sondage et des redressements effectués. Nous avons conduit une simulation de manière à déterminer quelle était la meilleure méthode pour effectuer les calculs de précision pour chacune des deux stratégies de repondération présentées précédemment.

Dans le cadre de la simulation 2, pour chacun des 1 000 échantillons de chaque stratégie de repondération, nous avons utilisé la procédure *Surveymeans* de SAS pour calculer la précision des estimateurs[4]. Comme nous l'avons vu dans la section précédente, lorsqu'un calage sur marges a été effectué, il faut en tenir compte dans le calcul de la précision. Nous avons alors, pour chacune des deux méthodes, calculé la moyenne des 1 000 valeurs obtenues pour estimer l'espérance de l'erreur-standard en fonction de la méthode retenue. Par ailleurs, nous considérons que, pour une variable donnée, l'erreur standard « vraie » peut être approchée par l'écart-type de la distribution des 1 000 estimateurs obtenus par simulations.

Le **tableau 4** montre que, dans le cas de la première stratégie, le calcul de précision conduit à surestimer l'erreur standard, en particulier pour les fractiles. En revanche, dans le cadre de la deuxième stratégie procédant par calage sur marges, l'emploi de la procédure reposant sur les $g_i e_i$ présentés dans la partie précédente, permet d'aboutir à des estimations correctes de la précision, avec cependant là encore de légers biais subsistant en ce qui concerne les fractiles ▸ **Tableau 5**. Des investigations supplémentaires mériteraient d'être conduites afin d'identifier les raisons des biais d'estimation observés selon la stratégie 1 (tableau 4). Une explication pourrait

▸ **Tableau 4** **Stratégie 1 : précision pour un sondage à allocation proportionnelle à la taille et un redressement de la non-réponse avec des coefficients d'ajustement**

	Erreur standard	Estimation
Français	0,08	0,09
Mathématiques	0,12	0,14
Histoire-géographie	0,09	0,10
Français < 8	0,73	0,83
Mathématiques < 6	0,94	1,07
Histoire-géographie < 8	0,87	0,96

Lecture : l'erreur standard de la note moyenne en français est de 0,08 (*i.e.* l'écart-type de la distribution des 1 000 notes moyennes obtenues). La moyenne des 1 000 estimations de la précision de la note moyenne en français est de 0,09.

4. Nous avons utilisé deux méthodes, l'une empirique (Jackknife), l'autre par linéarisation des formules de variance (Taylor). Elles donnent des résultats quasi-identiques, nous reproduisons ici la méthode empirique (Jackknife).

Tableau 5 Stratégie 2 : précision pour un tirage équilibré et un redressement de la non-réponse avec un calage sur marges

	Erreur standard	Estimation
Français	0,07	0,07
Mathématiques	0,10	0,11
Histoire-géographie	0,08	0,08
Français < 8	0,70	0,72
Mathématiques < 6	0,89	0,93
Histoire-géographie < 8	0,83	0,85

Lecture : l'erreur standard de la note moyenne en français est de 0,07. La moyenne des 1 000 estimations de la précision de la note moyenne en français est de 0,07.

être avancée en observant que le redressement effectué avec les coefficients d'ajustement consiste en une forme de correction comparable à un « calage » (sur le sexe et la variable de stratification), correction qui n'est pas prise en compte explicitement dans le calcul de précision.

CONCLUSION

Plusieurs types de questions se posent au moment du tirage d'un échantillon et les choix sont toujours contraints. Concernant les évaluations, une des premières contraintes porte sur les bases de sondage disponibles. Pour le second degré, en France, elles sont relativement riches et elles permettent d'utiliser l'information auxiliaire pour le tirage de l'échantillon et pour le redressement de la non-réponse. Ce n'est malheureusement pas le cas pour les évaluations réalisées dans le premier degré, car les bases de données sont très limitées.

Or, les simulations montrent que l'utilisation de l'information auxiliaire, quand elle est pertinente, permet d'améliorer la précision des estimateurs. Lorsque des contraintes pratiques dégradent la précision, comme le fait de sélectionner des classes plutôt que des établissements pour les évaluations nationales, les simulations montrent que l'utilisation de l'information auxiliaire dans ce cas permet de ne pas perdre en précision.

Ces résultats doivent interroger les pratiques des évaluations internationales qui mobilisent très peu l'information auxiliaire, avec l'argument de la standardisation, c'est-à-dire de la même procédure appliquée dans tous les pays, même si certains pays disposent de nombreuses informations susceptibles d'améliorer la qualité des échantillons. Or, des procédures différentes selon les pays pourraient être étudiées sans que cela nuise à la comparabilité des résultats, mais au contraire pour que chaque pays optimise son plan d'échantillonnage.

BIBLIOGRAPHIE

Ardilly P., 2006, *Les techniques de sondage*, Paris, Technip.

Christine M., Rocher T., 2012, « Construction d'échantillons astreints à des conditions de recouvrement par rapport à un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes : aspects théoriques et mise en œuvre dans le cadre du renouvellement des échantillons des enquêtes d'évaluation des élèves », *Journées de Méthodologie Statistique*, Paris, janvier 2012.

Deville J. C., Särndal C. E., 1992, "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, vol. 87, No. 418, p. 376-382.

Hubert T., 2014, « Un collégien sur cinq concerné par la "cyber-violence" », *Note d'information*, n ° 39, MENESR-DEPP, Paris.

Le Donné N., Rocher T., 2010, « Une meilleure mesure du contexte socio-éducatif des élèves et des écoles – Construction d'un indice de position sociale à partir des professions des parents », *Éducation et formations*, n° 79, MENJVA-DEPP, p. 103-115.

Martin M., Mullis I., Kennedy A., 2007, *PIRLS 2006 Technical Report*, Chestnut Hill, TIMSS & PIRLS International Study Center, Boston College.

OCDE, 2012, *PISA 2009 – Technical Report*, Paris, OCDE.

Rousseau S., Tardieu F., 2004, *La macro SAS CUBE d'échantillonnage équilibré – Documentation de l'utilisateur*, Paris, Insee.

Sautory O., 1993, « La macro CALMAR – Redressement d'un échantillon par calage sur marges », *Série des documents de travail*, Document n° F9310, Paris, Insee.

Tillé Y., 2001, *Théorie des sondages – Échantillonnage et estimation en populations finies – Cours et exercices avec solutions*, Paris, Dunod.