Composante du produit nº 12-001-X
Division des méthodes d'enquêtes auprès des entreprises

Article

Évaluation des règles de sélection dans les ménages sous un plan à plusieurs degrés

par Tom Krenzke, Lin Li et Keith Rust

Juin 2010

Statistique Canada Statistics Canada

Canada

Techniques d'enquête, juin 2010
Vol. 36, N° 1, pp. 119-128
Statistique Canada, N° 12-001-X au catalogue

Évaluation des règles de sélection dans les ménages sous un plan à plusieurs degrés

Tom Krenzke, Lin Li et Keith Rust [1]

Résumé

La National Assessment of Adult Literacy (NAAL) de 2003 et l'Enquête internationale sur la littératie et les compétences des adultes (ELCA) comportaient chacune un plan d'échantillonnage aréolaire stratifié à plusieurs degrés. Le dernier degré consistait à dresser la liste des membres du ménage, à déterminer la situation d'admissibilité de chaque individu et à appeler la procédure de sélection pour sélectionner aléatoirement une ou deux personnes admissibles dans le ménage. L'objectif du présent article est d'évaluer les règles de sélection dans les ménages sous un plan d'échantillonnage à plusieurs degrés en vue d'améliorer la procédure dans de futures enquêtes sur la littératie. L'analyse est fondée sur la distribution courante des ménages américains selon leur taille et sur les coefficients de corrélation intra-grappe en utilisant les données sur la littératie des adultes. Nous étudions plusieurs règles de sélection dans les ménages, en prenant en considération les effets de la mise en grappes, des taux d'échantillonnage différentiels, du coût par interview et du fardeau de réponse au niveau du ménage. Dans ce contexte, nous étendons une évaluation de l'échantillonnage dans les ménages sous un plan à deux degrés à un plan à quatre degrés et nous procédons à certaines généralisations aux échantillons à plusieurs degrés pour divers rapports de coûts.

Mots clés : Corrélation intra-grappe ; effets de plan ; échantillonnage à plusieurs degrés.

1. Introduction

La National Assessment of Adult Literacy (NAAL) de 2003, réalisée par le National Center for Education Statistics, a fourni aux chercheurs, aux praticiens, aux décideurs et au grand public un indicateur du progrès de la nation en ce qui concerne la maîtrise de l'anglais. Comme dans la National Adult Literacy Study (NALS) de 1992, on a évalué la compréhension de textes suivis, de textes schématiques et de textes au contenu quantitatif par les adultes présents dans les ménages. La conception des livrets de réponse était basée sur celle de la NALS de 1992 pour permettre la mesure des tendances entre 1992 et 2003.

Afin de réduire le coût des déplacements des intervieweurs pour rendre visite aux ménages, la NAAL comporte un plan d'échantillonnage en grappes stratifié à quatre degrés qui a produit 18 500 évaluations complètes administrées à des adultes de 16 ans et plus. Dans la NAAL, les comtés ont été groupés pour former des unités primaires d'échantillonnage (UPE), qui ont été stratifiées et sélectionnées au premier degré. Au deuxième degré, des unités secondaires d'échantillonnage (USE) ont été formées et sélectionnées parmi les UPE échantillonnées. Les USE correspondaient à des îlots de recensement individuels ou à des groupes d'îlots adjacents comptant au moins 60 ménages (MN) qui ont été formés à l'intérieur des limites des secteurs. Subséquemment, les ménages ont été sélectionnés parmi les USE, et une personne échantillonnée (1 PE) a été sélectionnée aléatoirement dans les ménages dont la taille était égale ou inférieure à 3 ($B \leq 3$) et deux personnes (2 PE) ont été sélectionnées dans les ménages dont la taille était supérieure à 3 ($B > 3$), où B désigne le nombre de personnes admissibles par ménage. Cette règle suit l'approche d'échantillonnage dans les ménages utilisée au premier cycle de la NAAL (NCES 2001), réalisée en 1992. Une évaluation de la règle de sélection a été effectuée en utilisant la distribution courante des ménages américains selon la taille et les coefficients de corrélation intra-classe calculés d'après les données de l'enquête de 2003. Pour cela, nous avons étendu une évaluation de l'échantillonnage dans les ménages sous un plan d'échantillonnage à deux degrés (Clark et Steel 2007) à un plan d'échantillonnage à quatre degrés, tel que celui utilisé dans l'enquête NAAL et avons fait certaines généralisations aux échantillons à plusieurs degrés pour divers ratios de coûts.

Les données utilisées pour l'évaluation comprennent les mesures de la littératie faites sur trois échelles correspondant à trois types de littératie, à savoir la compréhension de textes suivis, la compréhension de textes schématiques et la compréhension de textes au contenu quantitatif. Pour plus de renseignements au sujet des types de littératie de la NAAL, consulter http://nces.ed.gov/NAAL/fr_tasks.asp. Deux types d'estimations sont utilisés, à savoir les moyennes (par exemple, score moyen de compréhension de textes suivis) et le pourcentage d'adultes à un certain niveau de littératie (par exemple, pourcentage *sous le niveau de base* pour la compréhension de textes suivis). Pour une discussion des niveaux de littératie utilisés dans la NAAL, consulter http://nces.ed.gov/NAAL/perf_levels.asp. En plus des données de la NAAL, l'évaluation s'appuie sur les données de l'échantillon des États-Unis de l'Enquête sur la littératie et les compétences des adultes (ELCA), qui a été menée par

1. Tom Krenzke, Statistical Group, Westat, Rockville, Maryland 20850. Courriel : tomkrenzke@westat.com ; Lin Li, Statistical Group, Westat, Rockville, Maryland 20850. Courriel : linli@westat.com ; Keith Rust, Statistical Group, Westat, Rockville, Maryland 20850. Courriel : keithrust@westat.com.

Statistique Canada. En 2003, l'échantillon des États-Unis, commandité par le NCES, faisait partie de l'étude comparative destinée à mesurer les compétences des adultes dans plusieurs pays. Comme la NAAL, l'ELCA a été réalisée auprès d'un échantillon en grappes à plusieurs degrés et mesurait la compréhension de textes suivis et de textes schématiques, ainsi que la numératie (OCDE 2005). L'échantillon de la NAAL était beaucoup plus grand (18 500 évaluations achevées) que celui de l'ELCA (3 400 évaluations achevées), et la population cible de la NAAL comprenait les personnes de 16 ans et plus, tandis que celle de l'ELCA comprenait les personnes de 16 à 65 ans. Le tableau 1 résume le plan et la structure de chaque enquête.

À la section 2, nous présentons une discussion des aspects du plan de sondage pris en considération qui ont aidé à formuler l'évaluation des règles d'échantillonnage dans les ménages. À la section 3, nous discutons du calcul des corrélations intra-ménage sous les plans d'échantillonnage à plusieurs degrés et nous nous concentrons sur l'intégration de l'effet de grappe résultant des premiers degrés de sélection de l'échantillon pour décider d'une règle de sélection dans les ménages. À la section 4, nous procédons à une évaluation des règles de sélection en utilisant des données provenant d'enquêtes sur la littératie des adultes menées sur place et présentons les résultats. Enfin, à la section 5, nous résumons brièvement l'étude.

2. Considérations concernant le plan d'échantillonnage

Un certain nombre de facteurs doivent être pris en considération pour évaluer les règles de sélection dans les ménages pour des enquêtes telles que la NAAL et l'ELCA. La suite de la présente section porte sur l'effet qu'ont sur l'échantillonnage dans les ménages le fardeau de réponse, la mise en grappes des personnes dans les ménages, les taux d'échantillonnage différentiels, l'échantillonnage à plusieurs degrés, les considérations budgétaires, les systèmes informatisés, les domaines d'intérêt et la composition du ménage.

Fardeau de réponse du ménage. Dans les enquêtes sur l'alphabétisation des adultes, l'exécution de l'interview et de l'évaluation prend, en tout, environ une heure et demie. Par conséquent, l'une des préoccupations si l'on sélectionne plus d'une personne par ménage est l'accroissement du fardeau de réponse du ménage et l'effet sur les taux de réponse. Toutefois, comme le montre le tableau 2, l'écart entre les taux de refus de participer à l'ELCA et à la NAAL observés pour les ménages à 1 PE et à 2 PE n'est pas significatif (seuil de signification de 0,05).

Tableau 1
Caractéristiques de la NAAL et de l'ELCA

Enquête	Échantillon aréolaire	Évaluations complètes	Collecte des données	Évaluations	Âge	Règle d'échantillonnage dans les MN
NAAL	UPE, USE, ménages, personnes	18 500	Présélection Interview Évaluation	Textes suivis Textes schématiques Textes à contenu quantitatif	16 ans et plus	$B \leq 3$, $b = 1$ $B > 3$, $b = 2$
ELCA	UPE, USE, ménages, personnes	3 400	Présélection Interview Évaluation	Textes suivis Textes schématiques Numératie	16 à 65 ans	$B \leq 3$, $b = 1$ $B > 3$, $b = 2$

Nota : UPE = Unité primaire d'échantillonnage, USE = Unité secondaire d'échantillonnage, b = taille de l'échantillon, B = taille du ménage.

Tableau 2
Taux de refus par les ménages à 1 PE et 2 PE pour les enquêtes sur la littératie des adultes

Enquête	Sous-groupe	Taux de refus
NAAL	Ménages à 1 PE	16,3
	Ménages à 2 PE	15,7
ELCA	Ménages à 1 PE	17,6
	Ménages à 2 PE	16,2

Nota : PE = personne échantillonnée.

Mise en grappes des personnes dans les ménages. Kish (1965) discute des avantages d'un échantillon en grappes comparativement à un échantillon aléatoire simple. Habituellement, le coût par personne est plus faible pour un échantillon en grappes, mais la variance unitaire est plus élevée et pose de plus grandes difficultés dans l'analyse statistique. Kish a introduit le concept d'un effet de plan (DEFF), qui mesure l'accroissement de la variance dû aux écarts par rapport à un échantillon aléatoire simple, tels que la mise en grappes des personnes dans les ménages. Dans de nombreuses enquêtes, la sélection est limitée à une personne échantillonnée (PE) par ménage, en raison de préoccupations quant à l'accroissement de l'effet de grappe (c'est-à-dire l'accroissement de l'effet sur les estimations de la variance) associé à la sélection de plusieurs PE par ménage. L'effet de plan dû à la mise en grappes peut être exprimé sous la forme : $\text{DEFF}_{\text{grappe}} = 1 + (\bar{b} - 1)\,\text{Rhô}$, où $\bar{b} = \sum(M_B / M) b_B$, $M_B =$ nombre de ménages de taille B, $M =$ nombre de ménages, et $b_B =$ taille de l'échantillon de personnes dans les ménages de taille B (Kish 1965). Cette composante de l'effet de plan augmente quand la taille de l'échantillon dans un ménage augmente ou que la valeur de la corrélation intra-grappe (Rhô) augmente. Comme dans Cochran (1977), Rhô peut être approximé par

$$\text{Rhô} = 1 - \frac{\sigma_w^2}{\sigma^2},$$

où

$$\sigma_w^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.})^2 \,/\, (n - a)$$

et

$$\sigma^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 \,/\, (n - 1),$$

où a est le nombre de ménages échantillonnés, et b est le nombre de personnes échantillonnées par ménage. Nous examinons l'effet de plan dû à la mise en grappes de manière plus approfondie pour diverses règles d'échantillonnage dans les ménages à la section suivante.

Taux d'échantillonnage différentiels. L'effet de grappe n'est pas le seul facteur qui accroît la variance. Celle-ci peut également augmenter à cause des taux d'échantillonnage différentiels (qui donnent lieu à une pondération différentielle). Sous la stratégie de 1 PE par ménage, l'accroissement est directement associé à la variation de la taille du ménage, puisque le taux d'échantillonnage pourrait varier de 1 sur 1 à 1 sur 7 ou plus. L'effet de plan dû aux taux d'échantillonnage différentiels est exprimé sous la forme $\text{DEFF}_{\text{pond.}} = \sum(p_B / k_B)\sum(p_B k_B)$, où $p_B = N_B / N$, $N_B =$ nombre de personnes admissibles dans la population dans les ménages de taille B, $N =$ nombre de personnes admissibles dans la population, et $k_B =$ taux d'échantillonnage dans les ménages de taille B (Kish 1965). Sous certaines conditions, l'effet de plan global peut être exprimé comme étant le produit des composantes de mise en grappes et de taux d'échantillonnage différentiels, soit : $\text{DEFF} = \text{DEFF}_{\text{grappe}} \times \text{DEFF}_{\text{pond.}}$. Kalton, Brick et Lê (2005) laissent entendre que ce produit est applicable quand les poids sont aléatoires ou approximativement aléatoires.

Pour arriver à un échantillon autopondéré, les personnes pourraient être sélectionnées dans les ménages à un taux constant. Cependant, dans la plupart des enquêtes, la préférence n'est pas donnée à l'approche fondée sur les taux, parce qu'elle aboutirait à abandonner une partie des ménages ne comportant qu'une seule personne et, donc, augmenterait le coût de l'enquête. Nous limitons les diverses règles prises en considération à celles avec un minimum de 1 PE par ménage. Afin de ne pas imposer un trop lourd fardeau aux ménages, la taille maximale d'échantillon a été fixée à deux. Les règles d'échantillonnage envisagées sont les suivantes :

1. Tirer1 : 1 PE quelle que soit la taille du ménage.
2. Règle2 : 1 PE pour les tailles de ménages jusqu'à 2 ; sinon 2 PE sélectionnées.
3. NAAL3 : 1 PE pour les tailles de ménages allant jusqu'à 3 ; sinon 2 PE sélectionnées.
4. Règle4 : 1 PE pour les tailles de ménages allant jusqu'à 4 ; sinon 2 PE sélectionnées.
5. Frac5 : Tirer au moins 1 PE, mais pas plus de 2 PE et la taille d'échantillon est une fraction. Autrement dit, si la taille d'échantillon pour un ménage contenant deux personnes admissibles est 1,6, deux personnes sont sélectionnées 60 % du temps au hasard, et une personne est sélectionnée 40 % du temps.

Bien que l'approche Tirer1 n'essaye pas de réduire l'effet de plan dû aux taux d'échantillonnage différentiels, elle ne comporte pas d'effet de grappe. Cependant, les quatre autres approches susmentionnées donnent lieu à une réduction de la composante due aux taux d'échantillonnage différentiels, tout en introduisant un effet de grappe. Dans le cas de Frac5, sous l'hypothèse que les poids π sont utilisés, comme nous le supposons tout au long du présent exposé, l'approche sera celle donnant la réduction la plus importante de la composante due aux taux d'échantillonnage différentiels. L'approche des poids π est fondée sur la probabilité de sélection inconditionnelle de la personne dans le ménage. Si la taille réelle d'échantillon dans un ménage est utilisée sous la forme de poids proportionnels (ratio), le taux d'échantillonnage différentiel augmente et l'avantage est moins clair et dépend de Rhô. La figure 1 illustre les meilleures options sous un plan d'échantillonnage des ménages à deux degrés avec taille fixe de l'échantillon effectif de personnes, sans

aucune considération des coûts. Nous avons utilisé pour cet exemple la distribution nationale des ménages américains selon la taille établie d'après la Current Population Survey de 2007. Comme l'illustre la figure 1, l'approche fractionnaire est la meilleure règle pour une grande gamme de valeurs de Rhô. L'approche fractionnaire peut être programmée dans un système informatisé lorsque l'on dénombre et sélectionne les membres des ménages (une discussion plus approfondie des systèmes informatisés suit). Si aucun système informatisé n'est disponible pour la présélection, la meilleure approche pour les faibles valeurs de Rhô est celle où la mise en grappes est plus importante, c'est-à-dire la règle 2, tandis que la règle NAAL3 est la meilleure pour les valeurs de Rhô plus grandes qu'environ 0,34.

Échantillonnage à plusieurs degrés. Pour les plans d'échantillonnage aréolaire à plusieurs degrés, l'effet de la mise en grappes sur l'échantillonnage dans les ménages est affecté par la mise en grappes due aux UPE et aux USE. Comme l'a fait remarquer Kish (1965), la mise en grappes des ménages et des personnes dans les UPE et dans les USE accroît la variance d'échantillonnage (autrement dit, les unités dans les UPE et dans les USE sont plus semblables les unes aux autres). L'effet marginal de la mise en grappes dans les ménages peut être atténué par la domination des composantes de la variance liées aux UPE et aux USE (cependant, la grandeur de l'effet peut différer selon le type d'estimations et de variables). Autrement dit, un plus grand nombre de personnes peuvent être sélectionnées dans un ménage pour des enquêtes où l'importance de la mise en grappes est grande à cause des deux premiers degrés d'échantillonnage. Des précisions concernant cette distinction sont fournies à la section 3.

Considérations budgétaires. Le coût de la présélection d'un ménage dans le cas d'un plan d'échantillonnage de 1 PE par ménage comparativement au coût de l'interview/ évaluation d'une deuxième personne dans un ménage est examiné dans une analyse détaillée présentée plus loin.

Systèmes informatisés. Les systèmes informatisés, tels que l'interview sur place assistée par ordinateur (IPAO), ont la capacité de traiter des tailles d'échantillon fractionnaires. En d'autres mots, il est possible de programmer la sélection aléatoire de 1 ou 2 PE étant donné une taille d'échantillon fractionnaire préattribuée. Les systèmes informatisés ont également la capacité de trier la liste de personnes admissibles et de sélectionner 2 PE selon un échantillonnage aléatoire systématique. Un autre avantage tient au fait que le programme de sélection peut être testé et validé avant la collecte des données.

Domaines d'intérêt. Comme nous l'avons mentionné plus haut, l'échantillonnage optimal dans les ménages dépend de l'importance de l'effet de grappe associé à la variable d'intérêt. L'effet de grappe peut être plus faible quand la variable est associée à un sous-groupe de la population plutôt qu'à la population complète. Par exemple, quand un domaine de déclaration clé est le sexe dans une enquête auprès de la population adulte, la catégorie des hommes comportera vraisemblablement la sélection de 1 PE par ménage en moyenne et sera moins susceptible de comporter la sélection de 2 PE de sexe masculin qui introduirait un effet de grappe. Par conséquent, lorsqu'il existe de multiples domaines d'intérêt dans un ménage typique, il est souvent avantageux de sélectionner plus de 1 PE dans un ménage. Voir Mohadjer et Curtin (2008) pour un exemple de considérations concernant le plan d'échantillonnage pour des enquêtes axées sur de multiples sous-groupes de la population.

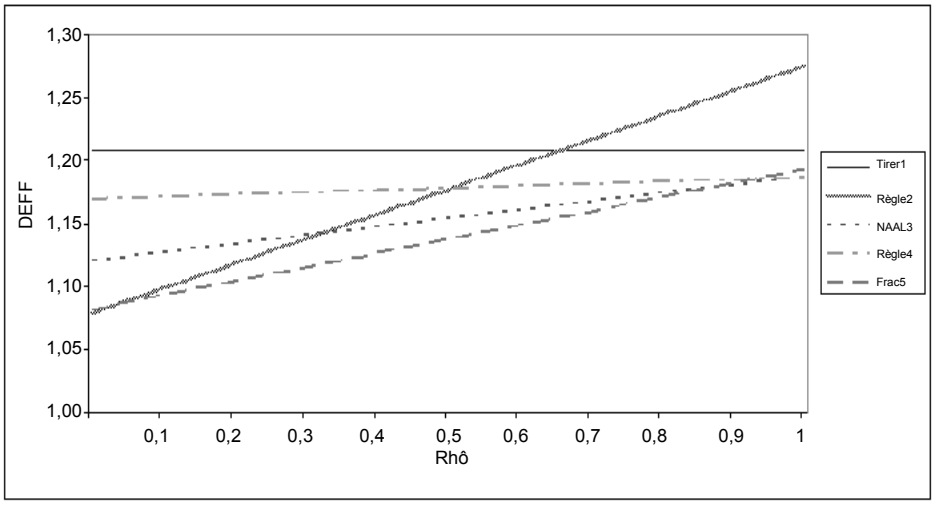

Figure 1 Analyse initiale des règles de sélection dans les ménages

Composition du ménage. Enfin, on pourrait prendre en considération la composition du ménage et les relations entre les personnes dans un ménage pour concevoir la règle de sélection. Le tableau 3 donne les valeurs de Rhô pour diverses relations entre les membres du ménage, pour un ménage avec 2 PE dans l'enquête NAAL. Rhô varie fortement selon les relations entre les membres du ménage. Ces relations ont été déterminées d'après le sexe et l'âge.

3. Estimation de Rhô *intra-ménage* et de DEFF sous échantillonnage à plusieurs degrés

Jusqu'à présent, la discussion au sujet de Rhô se rapportait à un plan d'échantillonnage à deux degrés, mais la NAAL ainsi que l'ELCA ont toutes deux un plan d'échantillonnage à quatre degrés. La variance totale peut être décomposée en quatre termes de variance de type inter attribuables aux UPE, aux USE, aux ménages et aux personnes, de la façon suivante :

$$\sigma_T^2 = \sigma_{\text{UPE}}^2 + \sigma_{\text{USE(UPE)}}^2 + \sigma_{\text{MN(USE)}}^2 + \sigma_{\text{PERS(MN)}}^2.$$

Comme nous le montrons plus bas, si nous suivons l'approche utilisée pour l'échantillonnage à deux degrés pour estimer Rhô pour un plan d'échantillonnage à quatre degrés, le numérateur contient non seulement la composante de variance inter-ménages, mais également les contributions des composantes inter-UPE et inter-USE qui accroissent les valeurs de Rhô.

$$\text{Rhô} = 1 - \frac{\sigma_{\text{PERS(MN)}}^2}{\sigma_T^2} = \frac{\sigma_{\text{UPE}}^2 + \sigma_{\text{USE(UPE)}}^2 + \sigma_{\text{MN(USE)}}^2}{\sigma_T^2}.$$

Par conséquent, pour évaluer les règles d'échantillonnage dans les ménages sous un plan à plusieurs degrés, nous supposons que le plan d'échantillonnage des UPE et des USE sera le même dans l'avenir. Nous pouvons accomplir cela en nous limitant à considérer l'échantillonnage à l'intérieur des USE. Par conséquent, le calcul de Rhô est contenu dans les USE, c'est-à-dire qu'il est effectué de manière compacte, sans effet dû aux composantes UPE et USE. Nous parlons dans ce cas du Rhô compact (c'est-à-dire dans les USE) dénoté par Rhô^*, exprimé par :

$$\text{Rhô}^* = \frac{\sigma_{\text{MN(USE)}}^2}{\sigma_{\text{MN(USE)}}^2 + \sigma_{\text{PERS(MN)}}^2}.$$

En utilisant le Rhô^* compact, nous calculons maintenant le DEFF estimé sous un plan d'échantillonnage à plusieurs degrés afin de déterminer la taille optimale d'échantillon dans les ménages. La variance d'une estimation $(\hat{\theta})$ avec b personnes par ménage peut être décomposée comme il suit :

$$\text{Var}(\hat{\theta}) = \frac{\sigma_{\text{UPE}}^2}{n_{\text{UPE}}} + \frac{\sigma_{\text{USE(UPE)}}^2}{n_{\text{USE}}} + \frac{\sigma_{\text{MN(USE)}}^2}{n_{\text{MN}}} + \frac{\sigma_{\text{PERS(MN)}}^2}{bn_{\text{MN}}}.$$

où n_{UPE}, n_{USE}, n_{MN} et bn_{MN} représentent respectivement les tailles d'échantillon des UPE, des USE, des ménages et des personnes.

Tableau 3
Rhô pour les scores d'évaluation de la NAAL selon la relation entre les membres du ménage

Estimation	Frères et sœurs	Enfant-tuteur	Mariés	Autre
Nombre de ménages avec 2 PE	111	205	180	434
Score moyen – textes suivis	0,42	0,35	0,70	0,59
Score moyen – textes schématiques	0,40	0,27	0,72	0,54
Score moyen – textes à contenu quantitatif	0,46	0,36	0,63	0,56
Pourcentage sous le niveau de base – textes suivis	0,52	0,41	0,79	0,67
Pourcentage sous le niveau de base – textes schématiques	0,54	0,40	0,78	0,60
Pourcentage sous le niveau de base – textes à contenu quantitatif	0,51	0,41	0,77	0,65

Puis, le DEFF dû à la mise en grappes, comparativement au tirage d'une personne par ménage et bn_{MN} ménages est donné par :

$$\mathrm{DEFF}_{\mathrm{grappe}}^{\mathrm{MN}}$$

$$= \frac{\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}} + \frac{\sigma^2_{MN(USE)}}{n_{MN}} + \frac{\sigma^2_{PERS(MN)}}{bn_{MN}}}{\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}} + \frac{\sigma^2_{MN(USE)}}{bn_{MN}} + \frac{\sigma^2_{PERS(MN)}}{bn_{MN}}}$$

$$= \frac{\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}} + \frac{1}{bn_{MN}}(\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)} + (b-1)\sigma^2_{MN(USE)})}{\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}} + \frac{1}{bn_{MN}}(\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)})}$$

$$= \frac{\frac{bn_{MN}\left(\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}}\right)}{\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)}} + (1 + (b-1)\,\hat{\mathrm{Rho}}^{*})}{\frac{bn_{MN}\left(\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}}\right)}{\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)}} + 1}$$

$$= \frac{k^{*} + (1 + (b-1)\,\hat{\mathrm{Rho}}^{*})}{k^{*} + 1}$$

où

$$k^{*} = \frac{bn_{MN}\left(\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}}\right)}{(\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)})}$$

$$= \frac{\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}}}{\frac{1}{bn_{MN}}(\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)})}.$$

Alternativement $\mathrm{DEFF}_{\mathrm{grappe}}^{\mathrm{MN}}$ peut être exprimé sous la forme :

$$\mathrm{DEFF}_{\mathrm{grappe}}^{\mathrm{MN}} = 1 + \frac{(b-1)\,\hat{\mathrm{Rho}}^{*}}{k^{*} + 1}$$
$$= 1 + (b-1)\,\hat{\mathrm{Rho}}^{**}$$

où

$$\hat{\mathrm{Rho}}^{**} = \frac{\hat{\mathrm{Rho}}^{*}}{k^{*} + 1}$$

$$= \frac{\left(\frac{\sigma^2_{MN(USE)}}{\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)}}\right)\frac{1}{bn_{MN}}(\sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)})}{\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}} + \frac{\sigma^2_{MN(USE)}}{bn_{MN}} + \frac{\sigma^2_{PERS(MN)}}{bn_{MN}}}$$

$$= \frac{\frac{\sigma^2_{MN(USE)}}{bn_{MN}}}{\frac{\sigma^2_{UPE}}{n_{UPE}} + \frac{\sigma^2_{USE(UPE)}}{n_{USE}} + \frac{\sigma^2_{MN(USE)}}{bn_{MN}} + \frac{\sigma^2_{PERS(MN)}}{bn_{MN}}}$$

$$= \frac{\sigma^2_{MN(USE)}}{\frac{bn_{MN}\sigma^2_{UPE}}{n_{UPE}} + \frac{bn_{MN}\sigma^2_{USE(UPE)}}{n_{USE}} + \sigma^2_{MN(USE)} + \sigma^2_{PERS(MN)}}$$

La mesure $\hat{\mathrm{Rho}}^{**}$ est une expression utile pour la corrélation intra-ménage sous un plan d'échantillonnage à plusieurs degrés, qui est égale à $\hat{\mathrm{Rho}}^{*}$ quand $\sigma^2_{UPE} = \sigma^2_{USE(UPE)} = 0$. La mesure $\hat{\mathrm{Rho}}^{*}$ compacte est utile pour évaluer les tailles optimales d'échantillon en faisant varier le ratio des variances k^{*}. Notons, toutefois, qu'en général, $\hat{\mathrm{Rho}}^{**}$ est une fonction de n_{UPE}, de n_{USE} et de la taille totale d'échantillon de personnes, tandis que $\hat{\mathrm{Rho}}^{*}$ ne dépend pas de ces paramètres.

Comme le montre le tableau 4, le ratio des variances k^{*}, qui est égal à la variance provenant des deux premiers degrés d'échantillonnage divisée par la variance provenant des deux derniers degrés pour un plan de sélection d'une personne par ménage, varie de 0,68 à 1,61 selon le type d'évaluation et d'estimation pour l'ELCA.

Le tableau 5 donne les estimations de Rhô (calculées sous une hypothèse de plan à deux degrés), du $\hat{\mathrm{Rho}}^{*}$ compact et du $\hat{\mathrm{Rho}}^{**}$ (calculé sous une hypothèse de plan à plusieurs degrés où $k^{*} = 1$) pour les scores moyens d'évaluation de la littératie de la NAAL et de l'ELCA. Quand nous incluons l'effet de grappe dû aux deux premiers degrés du plan d'échantillonnage à quatre degrés, les valeurs de $\hat{\mathrm{Rho}}^{*}$ compact et de $\hat{\mathrm{Rho}}^{**}$ sont beaucoup plus faibles que celle de Rhô. Par exemple, le Rhô à deux degrés pour le score moyen de compréhension des textes suivis de la NAAL est de 0,57, le $\hat{\mathrm{Rho}}^{*}$ compact est égal à 0,33 et $\hat{\mathrm{Rho}}^{**}$ est égal à 0,17. Le tableau montre aussi que les valeurs du $\hat{\mathrm{Rho}}^{*}$ compact pour les scores moyens sont à peu près au même niveau pour la NAAL (fourchette de 0,32 à 0,33) et pour l'ELCA (fourchette de 0,29 à 0,39). Il existe également une certaine variation selon le type d'estimation ; pour l'ELCA, les valeurs de $\hat{\mathrm{Rho}}^{*}$ pour le pourcentage calculé pour le niveau 1 ou 2 sont inférieures de 0 à 0,2 à celles pour les scores moyens. Les valeurs de $\hat{\mathrm{Rho}}^{*}$ peuvent aussi varier selon la taille du ménage, comme le montre la figure 2 à l'annexe A.

Tableau 4
Valeurs de k^* pour l'échantillon de l'ELCA

Estimation de l'ELCA	k^*
Score moyen – textes suivis	0,95
Score moyen – textes schématiques	1,56
Score moyen – textes à contenu quantitatif/numératie	1,13
Pourcentage au niveau 1 ou 2 – textes suivis	0,68
Pourcentage au niveau 1 ou 2 – textes schématiques	1,61
Pourcentage au niveau 1 ou 2 – numératie	1,10

Tableau 5
Valeurs de Rhô, Rhô* et Rhô** pour les scores d'évaluation de la littératie

	Rhô		Rhô*		Rhô**	
Estimation	NAAL	ELCA	NAAL	ELCA	NAAL	ELCA
Nombre de ménages avec 2 PE	930	162	930	162	930	162
Score moyen – textes suivis	0,57	0,60	0,33	0,38	0,17	0,19
Score moyen – textes schématiques	0,53	0,50	0,33	0,29	0,17	0,15
Score moyen – textes à contenu quantitatif/numératie	0,54	0,58	0,32	0,39	0,16	0,20
% sous le niveau de base (NAAL)/niveau 1 ou 2 (ELCA) – textes suivis	0,65	0,44	0,42	0,28	0,21	0,14
% sous le niveau de base (NAAL)/niveau 1 ou 2 (ELCA) – textes schématiques	0,61	0,37	0,39	0,28	0,20	0,14
% sous le niveau de base – textes quantitatifs (NAAL)/niveau 1 ou 2 (ELCA) numératie	0,62	0,36	0,40	0,17	0,20	0,09

Nota : Rhô** est calculé en supposant que $k^*=1$.

4. Évaluation et résultats

Nous avons comparé les règles courantes d'échantillonnage aux règles optimales d'échantillonnage en minimisant une fonction variance-coût (VC) qui est le produit des DEFF (c'est-à-dire, les accroissements de variance) dus à la formation de grappes et à la pondération ainsi que d'une fonction de coût qui est utilisée par Kish (1965) :

$$\mathrm{VC} = \mathrm{DEFF}^{\mathrm{MN}*}_{\mathrm{grappe}} \times \mathrm{DEFF}_{\mathrm{pond.}} \times n\left(c_p + \frac{c_{\mathrm{MN}}}{\bar{b}}\right),$$

où $c_p = $ coût par personne ajoutée et $c_{\mathrm{MN}} = $ coût par ménage ajouté. Notons que $n/\bar{b}$ représente le nombre de ménages échantillonnés. Afin de tenir compte des effets de grappe différentiels pour chaque taille de ménage B, nous remplaçons $\mathrm{DEFF}^{\mathrm{MN}}_{\mathrm{grappe}}$ par :

$$\mathrm{DEFF}^{\mathrm{MN}*}_{\mathrm{grappe}} = \frac{k^* + \sum_B \frac{M_B}{M}(1+(b_B-1)\,\mathrm{Rh\hat{o}}^*_B)}{k^*+1}$$

où $\mathrm{Rh\hat{o}}^*_B$ est calculé comme il est décrit à l'annexe A.

Notons que la fonction VC représente le coût additionnel de l'accroissement de la taille globale d'échantillon pour compenser l'accroissement de variance dû aux composantes du DEFF. Le tableau 6 donne les résultats pour les solutions entières optimales calculées au moyen d'un algorithme qui est décrit à l'annexe B. Le tableau montre que, quand le ratio des coûts passe de 0,5 à 1 pour $k^*=1$, nous souhaiterions tirer plus de personnes par ménage, c'est-à-dire 2 sur 2 au lieu de 1 sur 2. Quand le ratio des variances passe de 1 à 3 pour les solutions entières optimales, un changement n'est observé que pour la taille de ménage de 2 et le rapport des coûts de 0,5. Autrement dit, quand le rapport des variances est égal à 3, il est avantageux de tirer 2 personnes sur 2 au lieu de 1 personne sur 2.

Le tableau 6 donne les résultats quand des tailles fractionnaires d'échantillon sont permises. Les rapports des variances et des coûts pour la NAAL et l'ELCA ont tendance à être de l'ordre de 1, où il semble que la sélection de 1 personne sur 1, de 1,6 personne sur 2 et de 2 personnes autrement est la meilleure règle. Les effets des rapports des coûts et des variances sont plus clairs sous le scénario des tailles d'échantillon fractionnaires que sous celui des solutions entières.

Si le coût de l'exécution d'une présélection est faible comparativement à celui de l'interview, les variances peuvent être réduites en utilisant l'approche fractionnaire d'abandon de ménages. Le tableau 6 donne les tailles d'échantillon avec abandon optimales. Selon cette approche, par exemple, une taille d'échantillon de 0,9 indique que nous abandonnons 10 % des ménages, où $B=1$. Si le coût de la présélection est une part très faible du coût d'interview, le plan d'échantillonnage optimal pourrait comporter l'abandon d'un beaucoup plus grand nombre de ménages.

Sous les paramètres NAAL/ELCA probables pour les ratios des coûts $(C_{\mathrm{MN}}/C_p = 1)$ et les ratios des variance $(k^*=1)$, quand elle est comparée à l'approche Tirer1, la fonction VC peut être réduite d'environ 9 % en utilisant la règle d'échantillonnage de la NAAL/ELCA, 19 % en utilisant la solution entière optimale, 20,4 % en utilisant la solution fractionnaire optimale et 20,6 % en utilisant l'approche avec abandon de ménage optimale. En général, les gains dus à l'écart par rapport à l'approche Tirer1 augmentent à mesure que le coût par ménage supplémentaire (c'est-à-dire de la présélection) augmente. Les tailles moyennes de grappes pour chaque approche sont données au tableau 7. Pour la NAAL et la règle entière optimale, la taille moyenne de grappe indique le pourcentage de ménages avec 2 PE. Par exemple, environ 6 % des ménages compteraient 2 PE sous la stratégie NAAL3.

Tableau 6
Nombre prévu optimal de personnes par *ménage* selon le type de méthode d'échantillonnage des personnes et la taille du *ménage* (B)

		Méthode d'échantillonnage des personnes											
		Nombre entier				Nombre fractionnaire				Abandon			
k^*	C_{MN} / C_p	$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$	$B = 4$	$B = 1$	$B = 2$	$B = 3$	$B = 4$
1	0,5	1	1	2	2	1	1,4	2	2	0,6	1,3	2	2
1	1	1	2	2	2	1	1,6	2	2	0,9	1,6	2	2
1	2	1	2	2	2	1	1,9	2	2	1	1,9	2	2
3	0,5	1	2	2	2	1	1,6	2	2	0,8	1,5	2	2
3	1	1	2	2	2	1	1,8	2	2	1	1,8	2	2
3	2	1	2	2	2	1	2	2	2	1	2	2	2

Tableau 7
Réduction en pourcentage pour NAAL3 et les solutions optimales par rapport à la stratégie Tirer1 et taille moyenne des grappes

		Réduction en pourcentage par rapport à la stratégie Tirer1				Taille moyenne des grappes			
k^*	C_{MN} / C_p	NAAL3	Entier	Fractionnaire	Abandon	NAAL3	Entier	Fractionnaire	Abandon
1	0,5	8,2	13,0	15,8	18,0	1,06	1,18	1,38	1,21
1	1	9,1	19,2	20,4	20,6	1,06	1,68	1,48	1,45
1	2	9,9	26,1	26,1	26,1	1,06	1,68	1,63	1,63
3	0,5	8,6	17,3	18,7	19,0	1,06	1,68	1,48	1,37
3	1	9,5	23,7	23,9	23,9	1,06	1,68	1,58	1,58
3	2	10,4	30,2	30,2	30,2	1,06	1,68	1,68	1,68

Enfin, nous avons effectué une analyse de sensibilité en faisant varier les valeurs de Rhô*. Nous avons ajusté un modèle de régression en fonction de la réduction en pourcentage, par rapport à la stratégie Tirer1, de la fonction VC dans lequel les variables indépendantes étaient la stratégie utilisée (NAAL3, nombre entier, nombre fractionnaire, abandon de ménages), le rapport des coûts (0,1, 0,5, 1, 2, 10), le rapport des variances (1, 3, 5) et Rhô* (+/-0,1). Pour la gamme de données, Rhô* avait un effet limité (paramètre estimé de -7,4 avec une erreur-type associée de 4,5) sur la réduction en pourcentage de la fonction VC, tandis que les autres facteurs avaient un effet plus prononcé.

5. Résumé

Plusieurs aspects du plan de sondage ont été pris en compte pour évaluer la règle de sélection dans les ménages pour la NAAL et l'ELCA, y compris les effets de grappes dus aux degrés initiaux d'échantillonnage. Afin de faciliter l'évaluation, nous avons formulé un moyen d'intégrer les contributions des UPE et des USE à la variance dans le calcul de l'effet de plan dû à la mise en grappes et à la corrélation intra-ménage pour décider du nombre de personnes et du nombre de ménages qu'il convient de sélectionner dans un plan d'échantillonnage à plusieurs degrés. Pour cela, nous introduisons la mesure de Rhô* compact, qui est calculé à l'intérieur de l'USE de sorte qu'il n'est pas influencé par les composantes de la variance dues aux UPE et aux USE. Cette approche est utile si l'on veut déterminer l'effet de plan dû à la mise en grappes à l'intérieur des ménages, tout en faisant varier la contribution des degrés de sélection des UPE et des USE à la variance totale dans les plans d'échantillonnage à plusieurs degrés. La mesure Rhô** est introduite en tant qu'expression de la corrélation intra-ménage sous un plan à plusieurs degrés en tenant compte de la contribution des deux premiers degrés d'échantillonnage à la variance totale.

En outre, nous avons élaboré un algorithme pour calculer les solutions optimales concernant la taille d'échantillon en intégrant les effets de plan dus à la mise en grappes, aux taux d'échantillonnage différentiels et aux coûts.

En général, les principaux déterminants de la réduction en pourcentage de la fonction VC par rapport à l'approche Tirer1 sont le niveau de dominance des composantes de la variance dues aux UPE et aux USE dans l'échantillonnage à plusieurs degrés, le rapport des coûts et la règle de sélection utilisée. Pour la gamme de données évaluée, Rhô* a un effet limité sur la réduction de VC. Par rapport à l'approche Tirer1, en général, la règle de la NAAL produit une amélioration par rapport à la stratégie très répandue de tirage d'une personne (Tirer1). La règle d'un nombre entier optimal donne de meilleurs résultats que la règle de la NAAL. Cependant, la règle du nombre fractionnaire optimal n'offre qu'une amélioration limitée par rapport à la règle du nombre entier optimal. La règle de l'abandon optimal de ménages est meilleure que les autres règles pour les faibles rapports des coûts. Enfin, quand les deux premières composantes de la variance dominent et que le rapport des coûts est élevé, les règles du nombre entier, du nombre fractionnaire et de

l'abandon de ménages donnent essentiellement les mêmes résultats.

Remerciements

Les auteurs tiennent à souligner les contributions très utiles de Leyla Mohadjer et de Bob Fay.

Annexe A

Estimations de Rhô* selon la taille du ménage

Des estimations d'après les données d'enquête ne peuvent pas être obtenues pour Rhô* selon la taille du ménage, puisque 1 PE seulement a été sélectionnée quand la taille était égale ou inférieure à 3 et que la taille d'échantillon était trop faible pour produire des estimations pour chaque taille de ménage égale ou supérieure à 4. Par conséquent, les estimations de Rhô* selon la taille du ménage sont modélisées en utilisant des données de recensement. À la figure 2, Rhô* est représenté sur l'axe des y et la taille du ménage, sur l'axe des x. La courbe supérieure est celle obtenue pour l'échantillon du fichier de microdonnées à grande diffusion (FMGD) du Recensement des États-Unis pour le niveau de scolarité pour les personnes de 25 ans et plus. Cette courbe supérieure montre que le niveau de scolarité est plus semblable pour les ménages comptant deux adultes, qui sont peut-être plus susceptibles d'être des couples mariés. Elle révèle une diminution lorsque l'on passe de 2 à 3 adultes dans le ménage. Nous avons saisi la variation de la taille des ménages en calculant le ratio de Rhô* pour les scores de compréhension de textes suivis de la NAAL à Rhô pour le niveau de scolarité pour l'échantillon du FMGD du recensement parmi les ménages pour lesquels $B>3$ et en appliquant le ratio au Rhô de l'échantillon du FMGD pour toutes les tailles de ménages. Les valeurs résultantes sont les estimations du Rhô^*_B compact pour $B=1,\ 2,\ \ldots 11$.

Annexe B

Algorithme de calcul

Nous avons élaboré un algorithme de calcul pour arriver aux tailles d'échantillon optimales dans les ménages pour chaque taille de ménage B. Nous avons construit l'algorithme de façon à générer des solutions optimales entières ou fractionnaires qui reflètent les effets de la mise en grappes, des taux d'échantillonnage différentiels et du coût, sous les contraintes qu'on sélectionne au moins une personne par ménage et pas plus de deux. Suivent les étapes de l'algorithme (toutes les exécutions du traitement convergent en quatre itérations) :

- Initialiser en fixant $b=1$ pour toutes les valeurs de B (Tirer1).
- Calculer $\text{DEFF}^{\text{MN*}}_{\text{grappe}}$, $\text{DEFF}_{\text{pond.}}$, c_p, c_{MN} et VC(0).
- Faire $I=1$ à 5.
  - Faire $B=1$ à 11.
    - Calculer $\text{DEFF}^{\text{MN*}}_{\text{grappe}}$, $\text{DEFF}_{\text{pond.}}$, c_p, c_{MN} et VC pour tout $1\le b_B\le 2$, sachant l'ensemble de b_B, pour tout $B'\ne B$.
    - Repérer les b_B ayant la valeur la plus faible de VC.
  - Fin.
  - Si $\text{VC}(I)=\text{VC}(I-1)$, s'arrêter.
- Fin.

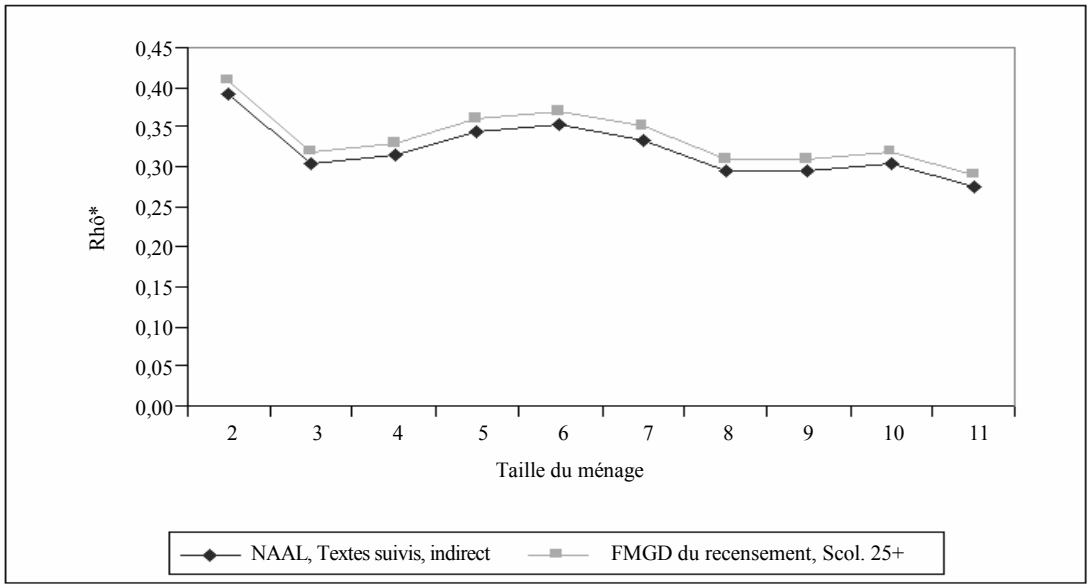

Figure 2 Estimations de Rhô* pour la NAAL selon la taille du ménage

Bibliographie

Clark, R.G., et Steel, D.G. (2007). Sampling within Households in Household Surveys. *Journal of the Royal Statistical Society*, Séries A, 170, 63-82.

Cochran, W.G. (1977). *Sampling Techniques*. 3ième Éd. New York : John Wiley & Sons, Inc.

Kalton, G., Brick, J.M. et Lê, T. (2005). Estimating Components of Design Effects for Use in Sample Design, Household Sample Surveys in Developing and Transition Countries, Chapitre VI United Nations, New York, 95-121.

Kish, L. (1965). *Survey Sampling*. New York : John Wiley & Sons, Inc.

Mohadjer, L., et Curtin, L.R. (2008). Trouver l'équilibre entre les divers objectifs du plan d'échantillonnage de la National Health and Nutrition Examination Survey. *Techniques d'enquête*, 34, 1, 131-140.

NCES (2001). Technical Report and Data File User's Manual For the 1992 National Adult Literacy Survey. U.S. Department of Education, National Center for Education Statistics, Washington, DC.

OCDE (2005). Apprentissage et réussite - Premiers résultats de l'enquête sur la littératie et les compétences des adultes. Organisation de Coopération et de Développement Économiques, Paris. Statistique Canada, Ottawa.