Nº 11-522-X au catalogue
ISSN 1709-822X

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Facteurs ayant une incidence sur la propension à répondre aux enquêtes, avec un intérêt pour les unités échantillonnées plusieurs fois

par Noah Johnson, Cilanne Boulet et Catherine Deshaies-Moreault

Date de diffusion : le 8 septembre 2025

Statistique Canada Statistics Canada

Canada

Recueil du Symposium de 2024 de Statistique Canada
Le futur des statistiques officielles

Facteurs ayant une incidence sur la propension à répondre aux enquêtes, avec un intérêt pour les unités échantillonnées plusieurs fois

Noah Johnson, Cilanne Boulet et Catherine Deshaies-Moreault[1]

Résumé

Le besoin en données ayant augmenté au cours des dernières années, l'effet et le fardeau de l'échantillonnage répété des mêmes unités pour plusieurs enquêtes sont devenus une préoccupation croissante. On suppose en général que le fardeau de réponse contribue à la diminution des taux de réponse; toutefois, peu d'études empiriques se penchent sur cette question. Dans le cadre de cette étude, les données sur les réponses aux enquêtes sociales menées à Statistique Canada entre 2021 et 2023 ont été agrégées afin d'examiner les facteurs contribuant aux profils de réponse observés, y compris l'effet d'avoir été sélectionnées plusieurs fois. Par rapport à certaines autres caractéristiques démographiques et géographiques, on a constaté qu'une unité échantillonnée plusieurs fois ne constituait pas un facteur influent dans la prédiction de la propension à répondre.

Mots clés : Fardeau de réponse; taux de réponse; non-réponse.

1. Introduction

Depuis de nombreuses années, les taux de réponse aux enquêtes sociales à Statistique Canada enregistrent une tendance à la baisse. Parallèlement, le nombre d'enquêtes en cours a augmenté, ainsi que la taille des échantillons au fil du temps. On suppose souvent que le fardeau d'être invité à répondre à plusieurs reprises aux enquêtes contribue à la baisse des taux de réponse (Yan et Williams, 2022). Statistique Canada a mis en œuvre des mesures pour tenter de gérer ce fardeau à l'avance, habituellement en utilisant des exclusions d'unités ayant déjà été échantillonnées pour une enquête au cours d'une certaine période. Toutefois, cela peut présenter des problèmes si une enquête cible certaines petites sous-populations d'intérêt, par exemple, les ménages avec enfants dans de petites provinces ou d'autres groupes de population. Dans cette situation, si les unités échantillonnées sont par la suite exclues d'une enquête auprès de la population générale en raison de préoccupations relatives au fardeau, il existe un risque de sous-représentation de ces groupes et un biais potentiel dans l'enquête auprès de la population générale du fait de ces exclusions, si les caractéristiques distinctives des unités exclues sont liées au paramètre d'intérêt de l'enquête. Cet effet est accru lorsque de petites sous-populations sont ciblées de façon répétée pour les enquêtes. Il est donc intéressant de connaître la mesure dans laquelle l'échantillonnage répété peut contribuer à la diminution des taux de réponse, puisque cela peut aider à orienter les décisions sur la façon dont le fardeau de réponse est géré compte tenu du risque connexe de biais.

La présente étude examine les taux de réponse parmi les unités sélectionnées pour plusieurs enquêtes de Statistique Canada. Sinibaldi et Karlsson (2017) ont effectué un travail semblable en examinant les personnes sélectionnées pour au moins deux enquêtes et en se concentrant sur la relation entre la période de repos (c'est-à-dire le temps entre la sélection initiale et la sélection suivante pour la même enquête ou une enquête différente) et les taux de réponse. Leur étude était fondée sur les données de 10 605 personnes échantillonnées deux fois pour cinq enquêtes-ménages sur une période de 12 ans. En comparaison, les travaux en cours portent sur un plus grand nombre d'enquêtes sociales ayant des échantillons de plus grande taille; la grande quantité de données fournit un niveau élevé de richesse analytique, notamment la possibilité d'examiner l'effet de la sélection multiple sur la réponse et d'inclure les unités sélectionnées un plus grand nombre de fois.

2. Préparation des données

[1]Noah Johnson, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, noah.johnson@statcan.gc.ca; Cilanne Boulet, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6, cilanne.boulet@statcan.gc.ca; Catherine Deshaies-Moreault, Statistique Canada, 150, promenade Tunney's Pasture, Ottawa (Ontario), Canada, K1A 0T6 catherine.deshaies-moreault@statcan.gc.ca

Pour procéder à cette analyse, trois sources de données ont été couplées au niveau des logements : un fichier maintenu manuellement utilisé par la plupart des enquêtes sociales de Statistique Canada pour suivre et gérer le fardeau de réponse, des paradonnées des systèmes de gestion de la collecte, et un fichier contenant des caractéristiques démographiques et géographiques supplémentaires pour les logements individuels faisant partie de la base de sondage sur les logements de Statistique Canada. Ces données ont été couplées au niveau des logements, puisqu'il s'agit de l'unité utilisée pour la coordination des échantillons pour les enquêtes sociales à Statistique Canada. Le fichier obtenu contenait des données pour 3 257 270 sélections d'enquêtes uniques dans 2 980 207 logements pour 23 programmes d'enquêtes sociales différents à Statistique Canada de 2021 à 2023 inclusivement. Ces données étaient limitées aux logements situés dans les 10 provinces du Canada. Fait important, ce fichier contenait une variable indiquant si un logement avait répondu à chaque sélection. D'autres métavariables liées à une sélection donnée et à des sélections antérieures ou subséquentes ont également été ajoutées au fichier. Des exemples de ces métavariables comprennent la période de repos, la sélection que représente un enregistrement donné (c.-à-d., la première instance chronologique de sélection du logement dans le fichier, la deuxième, etc.), et si le logement a répondu à l'enquête précédente pour laquelle il a été sélectionné.

En pratique, le processus d'élaboration de ce fichier n'a pas été aussi simple qu'espéré. Pour plusieurs enquêtes, les données du fichier sur le fardeau mis à jour manuellement étaient incomplètes; ce qui a entraîné l'exclusion de quelques enquêtes du fichier final. Il était également courant de faire face à des problèmes où les formats des identificateurs uniques différaient légèrement entre les sources; ce qui a nécessité un processus d'investigation et de résolution manuel et fastidieux. Enfin, il était possible que certains logements comportent des renseignements manquants pour différentes variables. Dans les cas de couplage réussi, ce problème a été résolu en tentant d'utiliser l'information provenant de l'autre source pour combler les lacunes. Ces défis mettent en évidence les avantages de la mise en place de systèmes interopérables; cela peut faciliter le type d'analyse effectué ici.

3. Analyse

3.1 Niveaux d'échantillonnage et répartition

Comme base d'analyse, il est utile de se faire une idée de la situation relative au fardeau de réponse dans l'ensemble du Canada. Le tableau 3.1-1 présente la répartition des sélections individuelles aux enquêtes par rapport au nombre de logements pour chaque province au Canada, suivie du ratio de ces nombres (en fait une fraction d'échantillonnage moyenne). Cela donne une idée du fardeau imposé aux logements dans chaque province et de la mesure dans laquelle il varie selon la province par rapport au ratio canadien global de 0,2. Les trois territoires du nord du Canada, qui représentent environ 0,3 % des logements au Canada, ont été exclus de cette analyse. Il est en effet particulièrement difficile de coupler des renseignements au niveau du logement dans ces régions et celles-ci font face à des défis uniques liés au fardeau de réponse.

Tableau 3.1-1
Nombre de sélections aux enquêtes par rapport au nombre de logements par province

Province	Sélections aux enquêtes (2021-2023)		Logements (2021)		Rapport des nombres
	Nombre	Pourcentage	Nombre	Pourcentage	
Terre-Neuve-et-Labrador	152 591	5,1 %	269 184	1,7 %	0,57
Île-du-Prince-Édouard	115 238	3,9 %	74 934	0,5 %	1,54
Nouvelle-Écosse	199 882	6,7 %	476 007	2,9 %	0,42
Nouveau-Brunswick	199 444	6,7 %	366 146	2,3 %	0,54
Québec	580 197	19,5 %	4 050 164	24,9 %	0,14
Ontario	807 208	27,1 %	5 929 250	36,5 %	0,14
Manitoba	229 054	7,7 %	571 528	3,5 %	0,40
Saskatchewan	218 249	7,3 %	513 725	3,2 %	0,42
Alberta	353 518	11,9 %	1 772 670	10,9 %	0,20
Colombie-Britannique	401 889	13,5 %	2 211 694	13,6 %	0,18
Canada	**3 257 270**	**100 %**	**16 235 302**	**100 %**	**0,20**

Les provinces dont la population est plus petite sont suréchantillonnées comparativement aux provinces plus peuplées. C'est le cas parce que, malgré la variation substantielle de la taille des provinces canadiennes, les données d'enquête doivent être suffisamment précises au niveau provincial pour appuyer les décisions de politique concernant les portefeuilles gérés au niveau provincial. Au cours de la période de trois ans couverte par ces données, le nombre de sélections d'échantillons dans les six provinces dont le nombre de logements est inférieur à un million équivalait à environ 49 % du nombre de logements. À titre comparatif, dans les quatre autres provinces dont le nombre de logements était supérieur à un million, le nombre de sélections d'échantillons équivalait à environ 15 % du nombre total de logements. Cet enjeu est particulièrement prononcé dans la province la moins peuplée, l'Île-du-Prince-Édouard, où les 74 934 logements ont été échantillonnés en moyenne un peu plus d'une fois et demie en trois ans. De plus, il est peu probable que la répartition des sélections d'échantillons soit uniforme parmi les logements au sein des provinces elles-mêmes. Même au sein des provinces moins peuplées, certains logements n'ont jamais été échantillonnés, tandis que d'autres ont été échantillonnés plus d'une fois. Souvent, les logements échantillonnés plusieurs fois faisaient partie d'un groupe de population d'intérêt (p. ex. les ménages avec enfants); ce qui exacerbe les problèmes liés au fardeau au sein de ces groupes. Si cette différence n'est pas prise en compte, cela entraîne également un biais potentiel dans les enquêtes subséquentes sélectionnées parmi les logements restants moins échantillonnés.

3.2 Taux de réponse par instance de sélection

Lors de l'examen des taux de réponse de l'ensemble des enquêtes de Statistique Canada, il est important de tenir compte du profil des enquêtes sociales et des facteurs qui rendent différentes enquêtes uniques. Le tableau 3.2-1 présente quatre enquêtes différentes menées entre 2021 et 2023 qui représentent bien ce panorama. Les résultats présentés dans les tableaux suivants concernent ces enquêtes en particulier.

Tableau 3.2-1
Résumé de certaines enquêtes sociales de Statistique Canada menées entre 2021 et 2023

Nom de l'enquête	Durée de l'enquête	Taille de l'échantillon	Obligatoire	Effort de collecte	Taux de réponse
Enquête sur la population active	Enquête par panel avec six itérations	Grand	Oui	Élevé	77,4 %
Enquête nationale sur les voyages	Courte	Très grand	Non	Faible	31,1 %
Enquête sur la santé dans les collectivités canadiennes	Longue	Moyen	Non	Moyen	52,8 %
Enquête sociale canadienne	Moyen	Moyen	Non	Moyen	50,4 %

Le tableau 3.2-2 porte sur les logements sélectionnés au moins deux fois. Chaque colonne compare les taux de réponse des logements dont la première sélection était pour une enquête donnée à ceux dont la deuxième sélection était pour cette enquête. Bien qu'il soit possible de sélectionner des logements deux fois pour la même enquête, il est beaucoup plus courant que les logements figurant dans le tableau aient été sélectionnés pour deux enquêtes différentes en raison de la coordination de l'échantillon entre les itérations de la même enquête.

Tableau 3.2-2
Taux de réponse parmi les logements sélectionnés au moins deux fois, par instance de sélection

Instance de sélection	Mesure	Enquête			
		Enquête sur la population active	Enquête nationale sur les voyages	Enquête sur la santé dans les collectivités canadiennes	Enquête sociale canadienne
1	Nombre de logements	9 905	162 039	7 547	10 598
	Taux de réponse	81,9 %	27,1 %	58,5 %	44,6 %
2	Nombre de logements	54 999	118 182	31 599	22 031
	Taux de réponse	81,3 %	25,8 %	54,6 %	49,4 %

Les logements sélectionnés au moins deux fois n'ont pas les mêmes caractéristiques que l'ensemble de l'échantillon; ils sont plus susceptibles de se trouver dans des provinces moins peuplées et dans des groupes échantillonnés plus fréquemment, comme les ménages avec enfants. Par conséquent, afin de déterminer si l'instance de sélection a une incidence, il est important de limiter l'analyse au groupe de logements sélectionnés au moins deux fois, comme le fait le tableau 3.2-2. En effet, il existe généralement des différences plus importantes entre les taux de réponse au tableau 3.2-2 et les taux de réponse globaux fournis au tableau 3.2-1 que par instance de sélection au sein d'une enquête. Cela reflète les différences dans la composition des groupes sur lesquels les taux de réponse sont calculés.

Au sein du groupe de logements sélectionnés au moins deux fois, peu de preuves indiquent que le fait d'être sélectionné une deuxième fois pour une enquête donnée (c.-à-d., après avoir été sélectionné pour une première enquête) ait un effet sur la probabilité de réponse. Une légère diminution des taux de réponse a été enregistrée entre les logements sélectionnés pour l'Enquête sur la population active à la première et la deuxième sélection. Pour l'Enquête nationale sur les voyages et l'Enquête sur la santé dans les collectivités canadiennes, les diminutions des taux de réponse ont été légèrement plus importantes (mais toujours petites) pour les logements sélectionnés en premier et ceux sélectionnés en deuxième. Toutefois, les taux de réponse ont augmenté entre ceux sélectionnés pour la première et la deuxième sélection de l'Enquête sociale canadienne. Le tableau 3.2-3 va plus loin en examinant les logements sélectionnés au moins trois fois et en comparant les taux de réponse pour ces mêmes enquêtes à la première, deuxième et troisième sélection.

Tableau 3.2-3
Taux de réponse parmi les logements sélectionnés au moins trois fois, par instance de sélection

Instance de sélection	Mesure	Enquête			
		Enquête sur la population active	Enquête nationale sur les voyages	Enquête sur la santé dans les collectivités canadiennes	Enquête sociale canadienne
1	Nombre de logements	351	13 961	136	617
	Taux de réponse	81,8 %	21,9 %	61 %	33,7 %
2	Nombre de logements	2 703	19 949	1 203	2 910
	Taux de réponse	81,6 %	21,8 %	59,4 %	45,5 %
3	Nombre de logements	5 858	11 214	1 944	2 523
	Taux de réponse	81,2 %	22,8 %	51 %	43,8 %

Pour les logements sélectionnés au moins trois fois, l'histoire est semblable à ce que présente le tableau 3.2-2, en ce sens qu'aucune tendance constante dans ces enquêtes ne montre une diminution des taux de réponse lors de l'augmentation des instances de sélection. Il est donc naturel d'envisager une autre approche tenant compte de facteurs supplémentaires pouvant influencer la probabilité de réponse, au-delà du nombre de fois qu'un logement est sélectionné.

3.3 Régression logistique

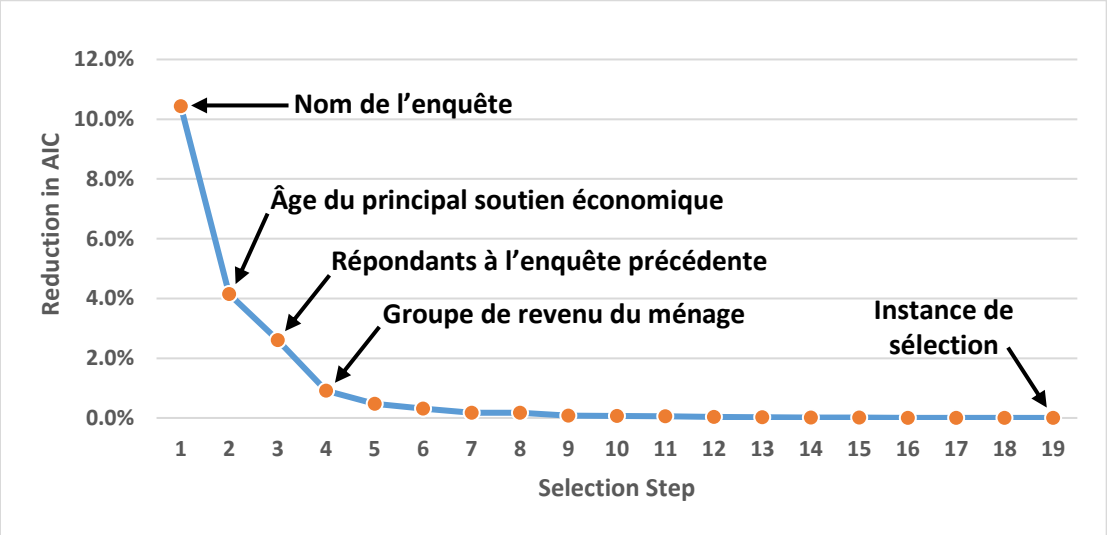
Afin de déterminer les facteurs influençant la propension à répondre et l'importance relative de ces facteurs, on a construit un modèle de régression logistique pour prédire la réponse parmi les logements sélectionnés au moins deux fois. La régression logistique a été choisie en raison de la nature catégorielle des variables et parce qu'elle est couramment utilisée pour prédire la propension à répondre dans les enquêtes sociales. Plusieurs ensembles de variables auxiliaires ont été examinés dans le but de trouver un modèle unique tenant le mieux compte de l'interaction entre les variables, ainsi que des différences dans le taux de réponse et la taille de l'échantillon entre les enquêtes. L'utilisation de certains de ces modèles a entraîné différents défis liés à la colinéarité, même si les conclusions tirées étaient semblables pour tous les modèles comparés. Pour nous aider à nous concentrer sur les conclusions clés, le modèle présenté dans ce travail est le plus simple. L'ensemble de données utilisé comme intrant pour ce modèle contenait 698 444 enregistrements, représentant les sélections individuelles aux enquêtes pour les logements sélectionnés au moins deux fois. De façon générale, les variables du modèle peuvent être catégorisées en variables liées à la sélection pour l'enquête ainsi qu'en variables géographiques et démographiques. Le tableau 3.3-1 présente une liste de ces variables et de la valeur-p associées à la statistique du khi carré pour chacune de ces variables à titre de prédicteurs de la réponse.

Tableau 3.3-1
Valeur-p associée à la statistique du khi carré pour les variables du modèle

Variable	Valeur-p	Variable	Valeur-p
Nom de l'enquête actuelle	< 0,0001	Province	< 0,0001
Instance de sélection	< 0,0001	Type de bâtiment	< 0,0001
Année de sélection	< 0,0001	Taille de la collectivité/influence métropolitaine	< 0,0001
Trimestre de sélection	< 0,0001	Taille du ménage	< 0,0001
Saison de sélection	< 0,0001	Groupe de revenu du ménage	< 0,0001
Nom de l'enquête précédente	< 0,0001	Types de ménage	< 0,0001
Indicateur de même enquête	< 0,0001	Groupe d'âge du principal soutien économique du ménage	< 0,0001
Indicateur de réponse à l'enquête précédente	< 0,0001	Sexe du principal soutien économique du ménage	< 0,0001
Indicateur de refus à l'enquête précédente	< 0,0001	Langue du principal soutien économique du ménage	< 0,0001
Période de repos (par périodes de six mois)	< 0,0001		

Toutes les variables ont une valeur-p inférieure à 0,0001; ce qui n'est pas surprenant compte tenu du grand nombre d'observations utilisées pour construire le modèle. Pour comprendre l'importance relative de ces variables, une approche n'étant pas fondée sur la valeur-p a été utilisée. Le critère d'information d'Akaike (en anglais Akaike information criterion ou AIC) peut être utilisé pour évaluer l'adéquation du modèle à un ensemble de données donné. La formule de l'AIC est *2k-2 (log-vraisemblance)*, où *k* est le nombre de paramètres dans le modèle. L'ajout de la constante 2*k* pénalise la complexité, et un test commun de l'importance lors de la comparaison de deux modèles appliqués aux mêmes données est de comparer les valeurs d'AIC associées. Si l'AIC d'un modèle avec un prédicteur supplémentaire est inférieur de plus de deux unités au modèle de base, alors ce modèle convient mieux aux données et explique une plus grande variation, malgré l'augmentation de la variance causée par une augmentation de la complexité. Pour obtenir les valeurs d'AIC associées aux données sur le fardeau, le même modèle de régression logistique a été généré dans SAS à l'aide d'un algorithme de sélection de variables par étapes fondé sur l'AIC. L'AIC donne un aperçu de l'importance relative des variables à chaque étape. La figure 3.3-1 montre la réduction en pourcentage de l'AIC entre chaque étape de l'algorithme de sélection des variables.

Figure 3.3-1
Réduction en pourcentage de l'AIC à chaque étape de sélection



La majeure partie de la réduction de l'AIC est attribuable à l'ajout des quatre premières variables, soit l'enquête actuelle pour laquelle le logement est sélectionné, le groupe d'âge du principal soutien économique du ménage

sélectionné, le fait que le logement ait répondu ou non à l'enquête précédente pour laquelle il a été sélectionné, et le groupe de revenu du ménage sélectionné. Ces quatre variables représentent 17,2 % de la réduction; toutes les autres variables réduisant le montant de l'AIC de seulement 1,4 % supplémentaire. En particulier, l'instance de sélection d'un logement a été introduite en dernier dans le modèle et a entraîné la réduction la plus faible de l'AIC parmi les variables choisies. Cette variable ne fournit pas beaucoup plus d'information au modèle une fois toutes les autres variables prises en compte, et la plus grande partie de la puissance de prédiction du modèle provient d'un nombre relativement limité de variables. Cela reflète les constatations des tableaux 3.3-2 et 3.3-3, qui suggèrent que la sélection pour une enquête précédente n'a pas d'incidence mesurable sur le plan opérationnel sur les taux de réponse aux enquêtes subséquentes.

4. Conclusion

L'élaboration de plans d'échantillonnage reflétant les expériences de diverses populations est un processus complexe qui peut nécessiter le suréchantillonnage de sous-groupes particuliers. Il est nécessaire de concilier les objectifs de l'enquête avec des considérations éthiques et des contraintes pratiques. Dans ce contexte, les règles de coordination d'échantillon peuvent faire en sorte que les sous-groupes suréchantillonnés ailleurs soient sous-représentés dans les enquêtes auprès de la population générale et constituent donc l'un des facteurs devant être soupesés dans le cadre du processus de plan d'échantillonnage.

La présente étude suggère que l'assouplissement des règles de coordination n'aurait pas une grande incidence sur les taux de réponse. Qu'il s'agisse d'examiner les taux de réponse ou les résultats d'un modèle de régression logistique, le fait qu'un logement ait été échantillonné plusieurs fois ne semble pas être un facteur influent dans la prédiction de la réponse, par rapport à d'autres caractéristiques démographiques et aux métadonnées sur la sélection de l'enquête. Cela suggère qu'il pourrait être possible d'assouplir les règles d'exclusion pour les logements sélectionnés précédemment, en particulier les règles liées au nombre de fois qu'un logement a été sélectionné et au délai entre ces sélections, sans diminuer encore les taux de réponse. Il s'agit donc d'une avenue qui sera explorée à Statistique Canada.

Néanmoins, les taux de réponse ne sont pas la seule raison de gérer le fardeau de réponse. D'un point de vue pratique, la sélection d'un même logement pour deux enquêtes recueillies en même temps porte à confusion et est difficile à gérer sur le plan opérationnel. De plus, compte tenu du temps et des efforts demandés aux répondants, il est plus équitable de veiller à ce que la sélection de l'échantillon ne soit pas trop concentrée sur un sous-groupe de la population, que ce soit par conception ou par sélection aléatoire, pourvu que toute incidence sur la représentativité puisse être gérée. La coordination des échantillons demeurera donc nécessaire, même si elle peut être assouplie. L'analyse présentée ici ayant été effectuée selon des règles de coordination d'échantillon qui limitaient de multiples sélections, les conséquences de l'utilisation de règles de coordination d'échantillons allégées seront réévaluées avec des données supplémentaires lorsque de futures enquêtes auront fait l'objet d'une collecte à l'aide de règles plus souples.

5. Remerciements

Les auteurs aimeraient remercier les collègues suivants pour les discussions et commentaires utiles au sujet de ce travail : David Laferriere, Kenza Sallier, Anne Mather, Cindy Ubartas et Elyse Lalonde.

Bibliographie

Sinibaldi, J. et A.Ö. Karlsson. (2017), « The Effect of Rest Period on Response Likelihood », *Journal of Survey Statistics and Methodology*, 5, p. 70-83.

Yan, T. et D. Williams. (2022), « Response Burden – Review and Conceptual Framework », *Journal of Official Statistics*, 38(4), p. 939-961.