

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

Correction des biais de sélection dans une enquête non probabiliste à deux phases sur les paiements

par Heng Chen et John Tsang

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Correction des biais de sélection dans une enquête non probabiliste à deux phases sur les paiements

Heng Chen et John Tsang¹

Résumé

Le présent article applique l'estimateur du pseudo-maximum de vraisemblance (PMV) à l'échantillonnage non probabiliste à deux phases lorsque des renseignements auxiliaires pertinents peuvent être tirés de l'échantillon d'enquête probabiliste et de l'échantillon d'enquête non probabiliste. Pour tenir compte de divers ajustements de poids et estimer la variance au-delà des totaux et des moyennes, tels que les médianes et les quantiles, une procédure simplifiée de pseudo-population selon la méthode bootstrap est proposée pour estimer approximativement la variance de la deuxième phase. Plus précisément, la simplification ignore la variabilité d'échantillonnage de la deuxième phase (c'est-à-dire qu'elle est traitée comme fixe, alors qu'elle est en fait aléatoire), si la fraction d'échantillonnage de la première phase de l'échantillon non probabiliste est négligeable. À l'aide de la deuxième vague de l'Enquête de 2020 sur les solutions de rechange à l'argent comptant de la Banque du Canada, la performance de la méthode proposée est comparée à celle d'autres méthodes, lesquelles soit ne modélisent pas explicitement la probabilité de sélection (c'est-à-dire la méthode itérative du quotient), soit ignorent les renseignements utiles de la phase 1 (c'est-à-dire la phase 2 uniquement). Selon les résultats obtenus, l'approche fondée sur le PMV fonctionne mieux que la méthode itérative du quotient et les estimations de la phase 2 uniquement dans le cadre de la réduction des biais de sélection des variables à deux phases liées aux paiements, particulièrement en ce qui concerne le groupe des jeunes dont le taux de réponse est faible. Les variances estimées des estimations basées sur le PMV sont stables.

Mots clés : Intégration des données; échantillon d'enquête non probabiliste; méthodes de rééchantillonnage; échantillonnage à deux phases; estimation de la variance.

1. Introduction

Dans le cadre d'un échantillonnage non probabiliste, un plan en deux phases permet de recueillir des informations auxiliaires importantes pour caractériser le mécanisme de sélection et la variable étudiée lorsque ces renseignements provenant de la base de sondage sont limités. Parmi les applications empiriques et les enquêtes sur les paiements réalisées dans ce cadre, on peut citer Henry et col. (2022) et Welte et Wu (2023). La section 2 explique comment produire des systèmes de pondération pour les première et deuxième phases d'un échantillon d'enquête non probabiliste à deux phases. Dans la section 3, nous proposons une procédure de pseudo-population selon la méthode bootstrap (PPB) pour l'estimation de la variance avec les systèmes de pondération de la section 2. La section 4 applique les méthodes décrites dans les sections 2 et 3. Plus précisément, nous utilisons la deuxième vague de l'Enquête de 2020 sur les solutions de rechange à l'argent comptant de la Banque du Canada (ESRACV2) pour présenter ces méthodes. Nous tentons également de remédier au biais de sélection dû au faible taux de réponse de la tranche d'âge de 18 à 34 ans dans l'ESRACV2. Par rapport au système de pondération actuel de l'ESRACV2, notre proposition permet de mieux réduire les biais de sélection. Les estimations obtenues grâce à la méthode proposée sont stables. La section 5 conclut et discute des recherches futures.

¹ Heng Chen, Monnaie, Banque du Canada, 234, rue Wellington O., Ottawa (Ontario), Canada, K1A 0G9 (hchen@bank-banque-canada.ca); John Tsang, Département de mathématiques et de statistique, Université d'Ottawa, 150, rue Louis-Pasteur, Ottawa (Ontario), Canada, K1N 6N5 (John.Tsang@uottawa.ca).

2. Pondération d'échantillons d'enquêtes non probabilistes en deux phases par intégration de données

L'échantillon de la phase 1 $S_{NP,1}$ provient d'une population finie $U = \{1, 2, \dots, N\}$ de la taille N ($S_{NP,1} \subset U$), tandis que l'échantillon de la phase 2 $S_{NP,2}$ provient de la première phase ($S_{NP,2} \subset S_{NP,1}$). L'ensemble de données $S_{NP,1}$ contient les variables auxiliaires $\mathbf{x}_k$, les variables de l'étude $\mathbf{y}_{1k}$ et l'indicateur i_{2k} permettant d'établir que l'unité $k \in S_{NP,1}$ a fait partie de la phase 2 : $\{(\mathbf{x}_k, \mathbf{y}_{1k}, i_{2k}) : k \in S_{NP,1}\}$. Après avoir recueilli des données au cours de la phase 1, $\mathbf{y}_{1k}$ est disponible comme information auxiliaire pour la phase 2. Par conséquent, les variables auxiliaires disponibles pour la phase 2 sont $\mathbf{z}_k = (\mathbf{x}_k, \mathbf{y}_{1k})$, $k \in S_{NP,1}$. Avec les variables de l'étude $\mathbf{y}_{2k}$ de chaque unité $k \in S_{NP,2}$, l'ensemble de données de $S_{NP,2}$ est $\{(\mathbf{z}_k, \mathbf{y}_{2k}) : k \in S_{NP,2}\}$. Nous disposons également d'un échantillon probabiliste de référence S_p au niveau individuel dont l'ensemble de données est $\{(\mathbf{x}_k, d_k) : k \in S_p\}$, où d_k est le poids (déterminé par le plan de sondage) de l'unité $k \in S_p$ et $\mathbf{x}_k$ désigne l'ensemble des variables (à l'exclusion de $\mathbf{y}_{1k}$) que S_p et $S_{NP,1}$ ont en commun.

Dans cette configuration, nous utilisons la méthode du pseudo-maximum de vraisemblance (PMV) pour estimer les probabilités de chaque unité échantillonnée k qui seront sélectionnées dans la phase 1 (π_{1k}) et qui seront sélectionnées dans la phase 2, à condition que k soit sélectionnée dans la phase 1 ($\pi_{2|1k}$). Pour faciliter l'estimation, nous utilisons des modèles de régression logistique pour toutes les probabilités. Ensuite, nous utilisons la pondération de probabilité inverse (PPI) pour obtenir des systèmes de pondération afin d'estimer des quantités de population finies, telles que les moyennes et les médianes, pour les phases 1 et 2.

2.1 Système de pondération de la phase 1

Dans la phase 1, nous utilisons la méthode du PMV pour estimer π_{1k} en intégrant $S_{NP,1}$ et S_p , puis obtenons un système de pondération à partir de la PPI (Chen et coll., 2020). Cette méthode est valable sous trois hypothèses principales. 1) La sélection de la phase 1 et la variable d'étude y_1 sont indépendantes conditionnellement aux variables auxiliaires. 2) Chaque unité de la population peut être sélectionnée. 3) Les unités échantillonnées sélectionnées pour la phase 1 sont indépendantes, conditionnellement selon leurs variables auxiliaires. $S_{NP,1}$ et S_p sont également indépendantes. Selon ces hypothèses, si nous utilisons une régression logistique pour modéliser les probabilités de sélection, la méthode du PMV conduit au système d'équations d'estimation suivant.

$$\sum_{k \in S_{NP,1}} \mathbf{x}_k^T \boldsymbol{\alpha} = \sum_{k \in S_p} d_k \log[1 + \exp(\mathbf{x}_k^T \boldsymbol{\alpha})]$$

Ce système d'équations correspond aux premiers moments des variables auxiliaires $\mathbf{x}_k$ de $S_{NP,1}$ et de S_p . La solution $\hat{\boldsymbol{\alpha}}$ à ce système conduit au système de pondération du PMV de la phase 1 : $\{w_{1k}^{\text{PMV}} = \pi_1(\hat{\boldsymbol{\alpha}}; \mathbf{x}_k) = \hat{\pi}_{1k} : k \in S_{NP,1}\}$.

2.2 Système de pondération de la phase 2

Chaque hypothèse de la phase 2 est analogue à celle de la phase 1. Ensuite, le PMV de la phase 2 conditionnellement à la phase 1 se réduit au système suivant d'équations d'estimation et équilibre les premiers moments des variables $\mathbf{z}_k$ des phases 1 et 2.

$$\sum_{k \in S_{NP,2}} \mathbf{z}_k \boldsymbol{\beta} = \sum_{k \in S_{NP,1}} \log[1 + \exp(\mathbf{z}_k^T \boldsymbol{\beta})],$$

Le système de pondération du PMV en deux phases est $\{w_{2k}^{\text{PMV}} = [\pi_1(\hat{\boldsymbol{\alpha}}; \mathbf{x}_k) \pi_{2|1}(\hat{\boldsymbol{\beta}}; \mathbf{z}_k, \mathbf{i}_1)]^{-1} = (\hat{\pi}_{1k} \hat{\pi}_{2|1k})^{-1} : k \in S_{NP,2}\}$, où $\hat{\boldsymbol{\beta}}$ est la solution à ce système, $\mathbf{i}_1$ est un vecteur d'indicateurs de sélection de la phase 1 et $\hat{\pi}_{1k}$ provient de la section 2.1.

3. Estimation de la variance en générant une pseudo-population selon la méthode bootstrap

En règle générale, une procédure de pseudo-population selon la méthode bootstrap (PPB) utilise d'abord le poids de chaque unité échantillonnée pour estimer la population. La population estimée est appelée pseudo-population. Ensuite, nous tirons des échantillons bootstrap de la pseudo-population et calculons des statistiques bootstrap à partir de ces échantillons. Mashreghi et coll. (2016) fournissent une excellente vue d'ensemble de l'estimation de la variance dans le cadre de l'approche de PPB. La procédure que nous proposons s'inscrit dans la même logique. Les paragraphes suivants décrivent étape par étape comment utiliser la PPB pour estimer approximativement les variances des quantités estimées $\hat{\theta}_1$ de la phase 1 et $\hat{\theta}_2$ de la phase 2.

Étape 1 : Répéter l'unité $k \in S_p$ [d_k] fois pour créer la pseudo-population U_p^* .

Étape 2 : Répéter l'unité $k \in S_{NP1}$ [w_{1k}^{PMV}] = [$\hat{\pi}_{1k}^{-1}$] fois pour créer une pseudo-population U_{NP}^* , y compris $\mathbf{z}_k$ et i_{2k} .

Étape 3 : Répéter les étapes 3a) et 3b) B fois pour créer B^2 paires d'échantillons bootstrap. Chaque paire comporte un échantillon bootstrap de chaque pseudo-population. Pour $b = 1, 2, 3, \dots, B$,

- **Étape 3a :** Tirer des échantillons bootstrap $S_p^{*(b)}$ de U_p^* selon un échantillonnage de Poisson avec des probabilités d'inclusion d_k^{-1} .
- **Étape 3b :** Tirer des échantillons bootstrap $S_{NP,1}^{*(b)}$ de U_{NP}^* selon un échantillonnage de Poisson avec des probabilités d'inclusion $\hat{\pi}_{1k}$. Nous ne mettons pas à jour la valeur de i_{2k} de chaque unité sélectionnée.

Étape 4 :

- En ce qui concerne la phase 1, utilisez le PMV de la phase 1 (section 2.1) pour produire un système de pondération à partir de chaque paire d'échantillons bootstrap. Puis, estimer θ_1 à partir de tous les systèmes de pondération B^2 , puis utiliser la formule habituelle de variance pour l'estimation de la variance.
- En ce qui concerne la phase 2, utiliser le PMV de la phase 2 (section 2.2) pour produire un système de pondération à partir de chaque paire d'échantillons bootstrap. Puis, estimer θ_2 à partir de tous les systèmes de pondération B^2 , puis utiliser la formule habituelle de variance pour l'estimation de la variance.

La procédure ci-dessus peut servir de solution unique pour estimer la variance de quantités de population finies telles que les totaux, les moyennes, les médianes et autres quantiles, tout en intégrant les ajustements de poids finaux tels que le calage. La variance estimée à l'aide de la procédure ci-dessus est une approximation due aux trois hypothèses simplificatrices ci-dessous :

1. À l'étape 1, nous avons supposé un échantillonnage de Poisson pour S_p . S'ils sont disponibles, nous devrions plutôt utiliser les poids de rééchantillonnage bootstrap de S_p .
2. Comme Beaumont et coll. (2015) ont montré, si nous supposons que la fraction d'échantillonnage de la première phase ($|S_{NP,1}|/N$) est négligeable et que la phase 2 est fixe (alors que la phase 2 est effectivement aléatoire), la contribution de la randomisation omise à la variance totale est négligeable. Par conséquent, à l'étape 3, nous étendons cette idée des échantillons probabilistes à deux phases aux échantillons non probabilistes à deux phases. Dans le contexte d'une méthode bootstrap de pseudo-population, les variances de la phase 2 peuvent être estimées approximativement même si les échantillons bootstrap des échantillons non probabilistes sont basés uniquement sur la randomisation de la phase 1.
3. Aux étapes 1 et 2, nous ignorons la variabilité de la partie fractionnaire des poids d_k et w_{1k}^{PMV} , tout en créant les pseudo-populations U_p^* et U_{NP}^* . Étant donné que, dans le deuxième point ci-dessus, nous avons supposé que la fraction d'échantillonnage de la phase 1 est négligeable, la variabilité omise du fait de cette simplification est très faible.

4. Application : Deuxième vague de l'Enquête de 2020 sur les solutions de rechange à l'argent comptant

La Banque du Canada a mené la deuxième vague de l'Enquête de 2020 sur les solutions de rechange à l'argent comptant afin de recueillir des renseignements sur les avoirs liquides et l'utilisation quotidienne de l'argent comptant par les Canadiens. Il s'agit d'une enquête non probabiliste en deux phases. La première phase est un questionnaire d'enquête (QE). La seconde phase est un instrument d'enquête par journal (IEJ) de trois jours. L'IEJ est un sous-échantillon strict du QE. Dans le système de pondération actuel de l'ESRACV2, le QE et l'IEJ ont chacun leur propre système de pondération produit par méthode itérative par le quotient. Ces deux systèmes de pondération considèrent chaque phase comme des échantillons d'enquête distincts. Les lecteurs peuvent consulter Chen et coll. (2021) pour de plus amples renseignements sur l'ESRACV2.

Comme beaucoup d'autres enquêtes, l'ESRACV2 affiche également un faible taux de réponse dans le groupe d'âge le plus jeune (de 18 à 34 ans). Le taux de participation du QE à l'IEJ pour ce même groupe d'âge est également faible. Nous utilisons les systèmes de pondération de la section 2 pour réduire le biais de sélection qui en résulte. Pour estimer les probabilités de sélection de la phase 1 avec le PMV, nous choisissons la Série d'enquêtes sur les perspectives canadiennes 5 (SEPC 5) menée par Statistique Canada comme échantillon probabiliste de référence. La SEPC 5 a été choisie parce que l'ESRACV2 et la SEPC 5 sont semblables en termes de population cible, de mode d'enquête et de période de collecte de données. Dans cet article, nous nous concentrons sur l'estimation de la médiane et de la moyenne des liquidités à partir du QE (phase 1) et de l'IEJ (phase 2).

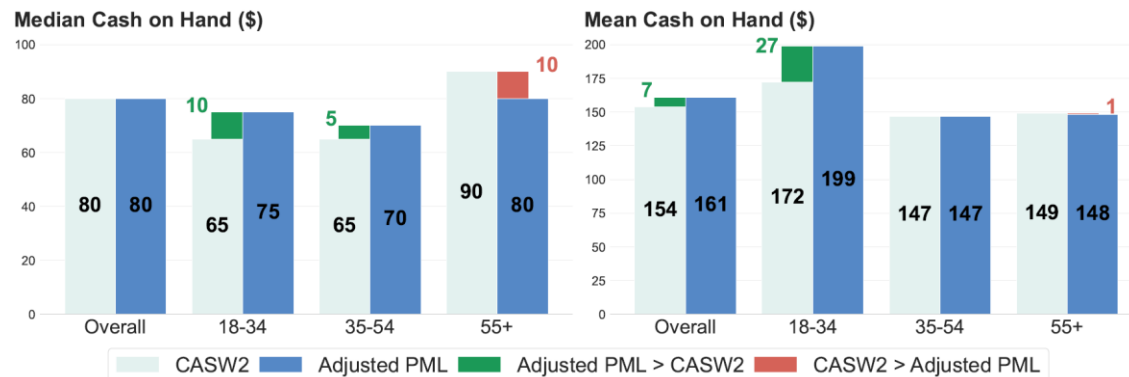
4.1 Phase 1 : Le questionnaire d'enquête

Dans l'application, après avoir estimé les poids avec le PMV de la phase 1 (section 2.1), nous utilisons la méthode itérative pour caler ces poids sur les totaux de population connus du recensement (PMV ajusté) afin de réduire davantage le biais et d'améliorer l'efficacité. Nous avons validé ces poids de PMV ajustés et le système de pondération actuel de l'ESRACV2 en comparant les estimations de huit variables relatives au comportement d'achat en ligne pendant la pandémie de COVID-19 avec celles de la SEPC 5, à la fois au niveau global et au niveau des groupes d'âge.

En utilisant ces estimations de la SEPC 5 comme données repères, les poids du PMV ont un meilleur rendement que le système de pondération de l'ESRACV2 actuel en termes d'écart relatif moyen et maximal dans tous les groupes d'âge, en particulier dans le groupe d'âge des 18 à 34 ans qui a peu répondu aux questions. Ce biais plus important du système de pondération de l'ESRACV2 actuel a mené à sous-estimer de 14 % les liquidités pour le groupe d'âge de 18 à 34 ans (figure 4.1-1). La variation estimée avec la PPB (section 3) montre que les estimations globales et par groupe d'âge du PMV suivi d'une méthode itérative sont stables (tableau 4.1-1).

Figure 4.1-1

Estimation des avoirs liquides selon le groupe d'âge : PMV suivi d'une méthode itérative du quotient par rapport à la méthode itérative du quotient pour la phase 1 de l'ESRACV2



Note : Le PMV ajusté désigne le PMV de la phase 1 suivi de la méthode itérative du quotient.

Tableau 4.1-1

Coefficients de variation des liquidités estimées par le PMV suivi de la méthode itérative du quotient, le questionnaire d'enquête, au niveau global et au niveau des groupes d'âge, en pourcentage.

	Moyenne	Médiane
Global	3,92	3,66
18 à 34 ans	8,90	9,03
35 à 54 ans	6,25	7,11
55 ans et plus	5,14	5,53

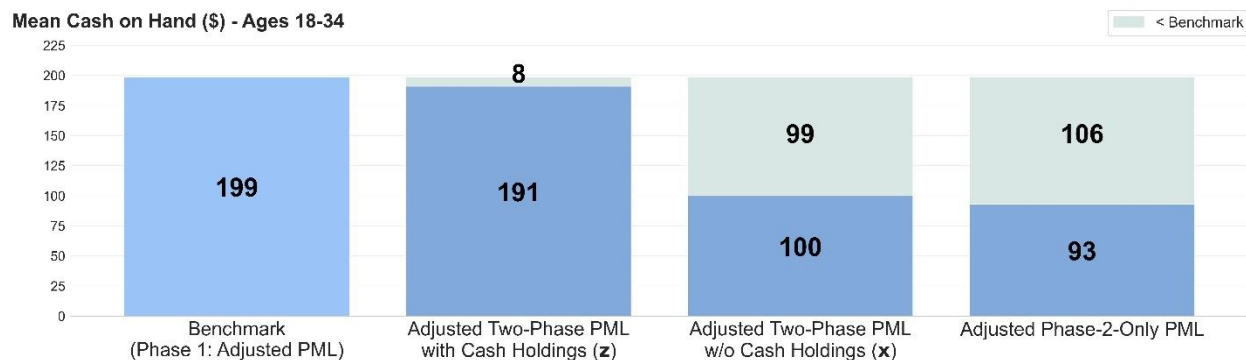
4.2 Phase 2 : L'instrument d'enquête par journal

Dans la phase 2, nous examinons quatre systèmes de pondération possibles. Chacun d'entre eux est suivi d'une méthode itérative du quotient. Les deux premiers systèmes de pondération traitent la deuxième phase indépendamment de la première : 1) le système de pondération de l'ESRACV2 actuel et 2) le système de pondération résultant de l'application directe du PMV à la phase 2 (PMV de la phase 2 uniquement). Les deux autres systèmes sont basés sur le PMV en deux phases (section 2.2). La différence entre ces systèmes restants est qu'ils incluent ou non les variables de l'étude de la phase 1 dans le modèle $\pi_{2|1k}$ (PMV en deux phases avec ou sans liquidités).

Pour comparer ces quatre systèmes de pondération, nous utilisons les liquidités estimées par le PMV ajusté de la phase 1 comme données repères. La comparaison montre que le PMV en deux phases avec les liquidités surpasse les trois autres candidats avec des marges significatives (figure 4.2-1). Par conséquent, le PMV en deux phases avec les variables d'étude de la phase 1 peut améliorer l'effort de réduction du biais pour la deuxième phase. En outre, les estimations globales et selon le groupe d'âge du PMV en deux phases avec les liquidités sont stables, comme le montrent les variances estimées à partir de la PPB (tableau 4.2-1).

Figure 4.2-1

Estimation des liquidités selon le groupe d'âge : PMV en deux phases avec différents ensembles de variables dans le modèle de probabilité de sélection de la deuxième phase et PMV de la phase 2 uniquement



Note : Les systèmes de pondération ajustés désignent les systèmes de pondération corrigés par la méthode itérative du quotient.

Tableau 4.2-1

Coefficients de variation pour les liquidités estimées par le PMV en deux phases (avec liquidités) suivi de la méthode itérative par le quotient, l'instrument d'enquête par journal, au niveau global et au niveau des groupes d'âge, en pourcentage.

	Moyenne	Médiane
Global	8,31	6,81
18 à 34 ans	23,11	14,48
35 à 54 ans	6,72	9,73
55 ans et plus	5,84	7,33

5. Discussion et prochaines étapes

S'appuyant sur l'approche du PMV, nous créons des systèmes de pondération de phase 1 et de phase 2 pour les échantillons non probabilistes à deux phases. Nous utilisons la deuxième vague de l'Enquête de 2020 sur les solutions de rechange à l'argent comptant de la Banque du Canada pour mettre en évidence la réduction des biais apportée par notre proposition et la comparer à d'autres solutions. Les trois principales conclusions empiriques sont les suivantes : i) les poids basés sur le PMV sont meilleurs que les poids obtenus par la méthode itérative du quotient (de l'ESRACV2) pour réduire le biais de sélection, en particulier lorsque les probabilités de sélection sont hétérogènes entre les différents groupes démographiques; ii) le PMV en deux phases, qui modélise séparément les probabilités de sélection de la phase 1 et de la phase 2 (conditionnellement à la phase 1), fournit des estimations comportant moins de biais que le système de pondération de rechange de la phase 2 uniquement, ce qui consiste à ignorer les renseignements de la première phase; et iii) il est souhaitable d'inclure les variables pertinentes de la phase 1 dans le modèle de sélection de la phase 2, tel qu'il est indiqué dans la section 4.2. Les estimations des sections 4.1 et 4.2 sont stables.

Nous prévoyons recueillir davantage de paradosnées, telles que le temps de réponse et la durée, pendant la collecte de données de la phase 1, afin de mieux caractériser la volonté des répondants de participer à la phase 2 (Liu et Netzer, 2023). En prévision de recherches futures, nous nous employons à déterminer comment appliquer l'intégration des données à d'autres enquêtes de la Banque du Canada afin de surmonter des probabilités de sélection inconnues ou compliquées, telles que l'échantillonnage indirect sans utiliser la méthode généralisée de partage des poids.

6. Remerciements

Ces travaux ont bénéficié du soutien financier de la Banque du Canada et de Mitacs dans le cadre du programme Accélération de Mitacs. Les opinions exprimées sont celles des auteurs et ne représentent pas nécessairement le point de vue officiel de la Banque du Canada. Les estimations présentées dans ce document sont préliminaires. Toutes les erreurs restantes relèvent de la seule responsabilité des auteurs.

Bibliographie

- Beaumont, J. F., Béliveau, A., et Haziza, D. (2015), "Clarifying some aspects of variance estimation in two-phase sampling", *Journal of Survey Statistics and Methodology*, 3(4), p. 524-542.
- Chen, H., Engert, W., Felt, M. H., Huynh, K., Nicholls, G., O'Habib, D., et Zhu, J. (2021), "Cash and COVID-19: The impact of the second wave in Canada", (No. 2021-12), Banque du Canada.
- Chen, Y., Li, P., et Wu, C. (2020), "Doubly robust inference with nonprobability survey samples", *Journal of the American Statistical Association*, 115(532), p. 2011-2021.

Henry, C., Shimoda, M., et Zhu, J. (2022), *2021 Methods-of-Payment Survey Report* (No. 2022-23), document de travail de la Banque du Canada.

Liu, S., et Netzer, N. (2023), “Happy times: Measuring happiness using response times”, *American Economic Review*, 113(12), p. 3289-3322.

Mashreghi, Z., Haziza, D., et Léger, C. (2016), “A survey of bootstrap methods in finite population sampling”, *Statistics Surveys*, 10, p.1-52.

Welte, A., et Wu, J. (2023). *The 2021–22 Merchant Acceptance Survey Pilot Study* (No. 2023-1), Banque du Canada.