

## Article

Symposium 2008 :  
Collecte des données : défis, réalisations et nouvelles orientations

### **Incidences du suivi des cas de non-réponse dans une enquête sur les conditions de vie des immigrants en Norvège**

par Johan Fosen, Gustav Haraldsen et Øyvind Kleven

2009



## **Incidences du suivi des cas de non-réponse dans une enquête sur les conditions de vie des immigrants en Norvège**

Johan Fosen, Gustav Haraldsen et Øyvin Kleven<sup>1</sup>

### **Résumé**

Dans une enquête où les résultats ont été estimés par des moyennes simples, nous comparons les incidences qu'ont sur les résultats un suivi des cas de non-réponse et une pondération fondée sur les derniers dix pour cent de répondants. Les données utilisées proviennent de l'enquête sur les conditions de vie des immigrants en Norvège qui a été réalisée en 2006.

Mots clés : Biais de non-réponse, suivi, pondération.

### **1. Introduction**

À la fin du processus de collecte des données d'une enquête, nous devons parfois décider s'il faut ou non étendre la période de collecte. Cette période de collecte des données supplémentaire est souvent appelée *deuxième vague de collecte* ou *suivi*, approche qui a été discutée pour la première fois par Hansen et Hurwitz (1946). Le but est habituellement de réduire le biais de non-réponse, mais parfois tout simplement d'accroître l'échantillon net. Dans le cas d'une enquête sur les conditions de vie des immigrants qui a été réalisée en Norvège en 2006, ces éléments ont dû être pris en considération.

L'Enquête sur les conditions de vie des immigrants (*Survey on Living Conditions among immigrants*) devait, au départ, voir la collecte des données se terminer le 1<sup>er</sup> juin 2006. Cependant, comme le taux de réponse était aussi faible que 54 % contrairement à celui prévu de 70 %, il a été décidé de prolonger la période de collecte des données, c'est-à-dire de procéder à un suivi ou à une deuxième vague de collecte des données. Nous décrirons l'enquête plus en détail à la section suivante. À la section 3, nous comparons les résultats pour certaines variables clés avant et après le suivi, en employant la méthode d'estimation simple que Statistics Norway a utilisé pour la diffusion des résultats. Des études antérieures portant sur d'autres enquêtes ont révélé que le suivi peut réduire le biais (p. ex. Thomsen et coll. 2006; Fosen et Kleven 2007) et c'est ce que nous espérons observer pour cette enquête également. D'où la première question à savoir si l'effort de suivi a porté fruit.

À la section 4, nous étudions l'effet du remplacement du suivi par une méthode d'estimation plus perfectionnée, où nous supposons que les derniers répondants arrivés sont ceux qui ressemblent le plus aux non-répondants. Nous appelons cette méthode *pondération fondée sur les répondants tardifs* et utilisons la procédure décrite par Bartholomew (1961) et utilisée, par exemple, dans Thomsen et Siring (1983). À la fin de la section 4, nous examinons l'effet et la rentabilité de l'utilisation combinée du suivi et de la pondération fondée sur les répondants tardifs.

### **2. Enquête sur les conditions de vie des immigrants en Norvège**

La figure 2-1 résume les principales caractéristiques de l'Enquête sur les conditions de vie des immigrants en Norvège (Gulløy 2008). Comme l'indique la figure, l'enquête couvrait les immigrants provenant de dix pays. Les clients s'inquiétaient de la non-réponse et la Division des enquêtes, qui a exécuté l'enquête, a fixé un taux de

---

<sup>1</sup>Johan Fosen, Statistics Norway, P. O. Box 8131 Dep, Norway, N-0033 Oslo (Johan.Fosen@ssb.no);  
Gustav Haraldsen, Statistics Norway, P. O. Box 8131 Dep, Norway, N-0033 Oslo (Gustav.Haraldsen@ssb.no);  
Øyvin Kleven, Statistics Norway, P. O. Box 8131 Dep, Norway, N-0033 Oslo (Oyvvin.Kleven@ssb.no)

réponse cible de 70 %. Toutefois, à la fin de la période originale de collecte à l'été 2006, le taux de réponse n'était que de 54 %.

Dans l'espoir d'accroître les taux de contact et, donc, d'obtenir des taux de réponse plus élevés, d'importantes ressources ont été affectées à une extension de la phase de collecte des données. De nouvelles procédures de suivi ont été introduites et l'unité d'enquête a été réaffectée aux employés présentant les possibilités de contact les plus élevées. Après sept mois supplémentaires, le taux de réponse est passé à 64 % et l'enquête a été clôturée.

**Figure 2-1**

**Certains renseignements importants sur l'Enquête sur les conditions de vie des immigrants en Norvège**

- Échantillon brut = échantillon stratifié de 500 provenant des 10 groupes d'immigrants les plus importants = 5 000
- Échantillon net = 64 %
- Méthodes de collecte des données : intervieweurs multilingues recrutés principalement parmi les groupes d'immigrants et ayant reçu une formation concernant l'enquête.
  - Interviews sur place = 62 %
  - Interviews téléphoniques = 38 %
- Vagues de collecte des données
  - De septembre 2005 à l'été 2006
  - De l'été 2006 à janvier 2007

### 3. Effets du suivi

Quand il a été décidé de procéder à une deuxième vague de collecte afin d'accroître le taux de réponse, l'hypothèse sous-jacente était que cela réduirait aussi le *biais de non-réponse*. Ce dernier correspond à l'erreur systématique dans les résultats d'enquête causée par la non-réponse. Le biais n'a une signification qu'en relation avec une variable particulière et nous pouvons parler de biais de non-réponse pour n'importe quel résultat estimé de l'enquête. Dans le présent article, le terme « biais » fait référence au biais *conditionnel* étant donné l'échantillon brut<sup>2</sup>. Nous le mesurons simplement par l'écart entre le résultat pour l'échantillon net et le résultat pour l'échantillon brut, c.-à-d.

$$\text{biais } \bar{Y} = \bar{Y}_{net} - \bar{Y}_{brut}.$$

Malheureusement, le résultat pour l'échantillon brut est généralement inconnu pour les variables d'une enquête. Cependant, l'écart relatif entre le biais avant et après la deuxième vague est

$$\frac{\text{biais } \bar{Y}_{net,orig} - \text{biais } \bar{Y}_{net,final}}{\bar{Y}_{brut}} = \frac{\bar{Y}_{net,orig} - \bar{Y}_{net,final}}{\bar{Y}_{brut}},$$

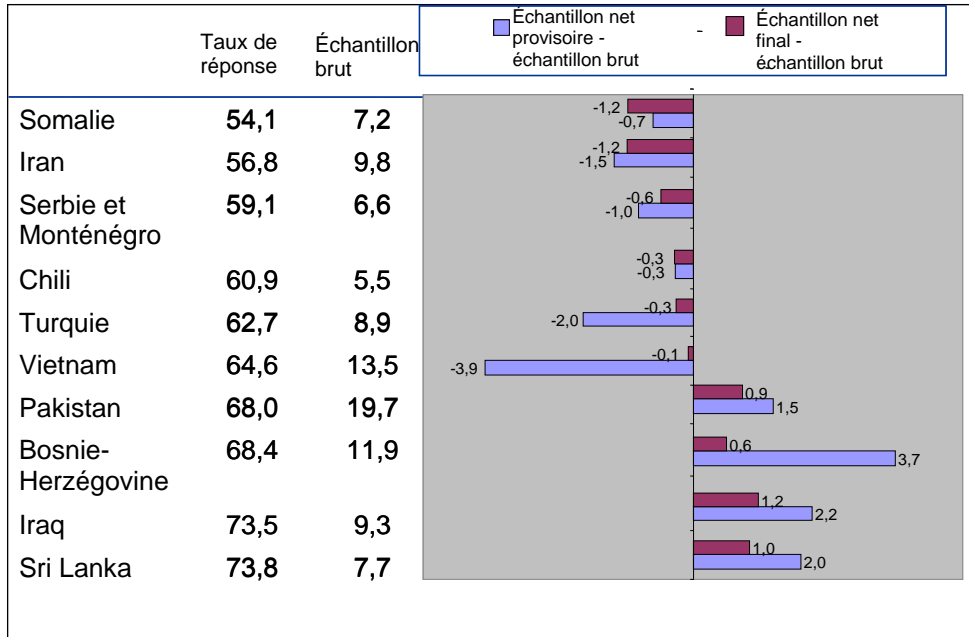
où  $\bar{Y}_{net,orig}$  et  $\bar{Y}_{net,final}$  représentent les moyennes dans l'échantillon net original et final, respectivement. Le biais relatif est proportionnel à la différence entre les deux moyennes d'échantillon net, si bien que la différence proprement dite révèle quelque chose au sujet de la différence relative de biais due au suivi.

Pour certaines variables non incluses dans l'enquête, nous connaissons en fait la moyenne d'échantillon brut. À Statistics Norway, les données provenant de registres administratifs sont potentiellement disponibles, puisque nous possédons une clé d'identification unique que nous pouvons utiliser pour fusionner les données d'enquête avec les données des registres. Nous pouvons considérer chaque variable d'un registre comme une variable d'enquête pour laquelle le taux de réponse est de 100 % et nous pouvons déterminer le biais de la variable du registre dans chacun des échantillons nets.

<sup>2</sup> Le biais *inconditionnel* est l'écart moyen par rapport au résultat de population, calculé sur de nombreux échantillons différents. Puisque nous ne disposons que d'un seul échantillon, nous considérons le biais étant donné cet échantillon brut, quoique le biais inconditionnel aurait pu être étudié en utilisant des techniques de rééchantillonnage.

**Figure 3-1**

**Biais de non-réponse de la distribution des pays d'origine pour l'échantillon net provisoire (avant le suivi) et l'échantillon net final**



La figure 3-1 montre l'effet du suivi sur la distribution des pays d'origine. Les pays sont classés par ordre croissant de taux de réponse, celui-ci variant d'un peu plus de 50 % à près de 75 %. Pour chaque pays, le tableau donne la variable « Proportion de répondants ayant ce pays comme pays d'origine » et la différence entre les barres représente la variation du biais due au suivi. Comme on peut le voir, le biais a diminué pour la plupart des pays, surtout pour la Turquie, le Vietnam et la Bosnie-Herzégovine. Le seul pays pour lequel le biais a augmenté est la Somalie. Par conséquent, il semble raisonnable de conclure que le suivi produit une meilleure répartition entre les pays que l'échantillon net provisoire.

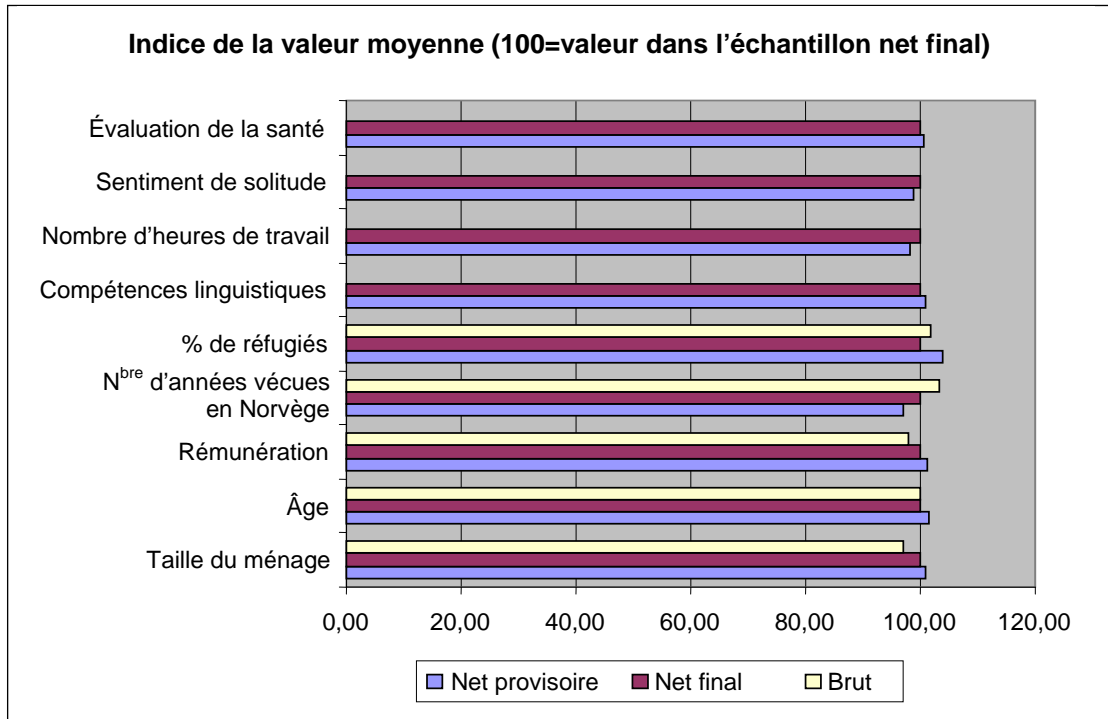
Nous souhaitons comparer les valeurs moyennes avant et après le suivi pour plusieurs variables d'enquête et variables de registre. Dans la mesure du possible, nous désirions également faire la comparaison avec la valeur réelle de l'échantillon brut. Cela a été fait pour deux types de variables de registre. Premièrement, nous avons examiné des variables auxiliaires classiques, comme la taille du ménage, l'âge et la rémunération. Deuxièmement, nous avons choisi certaines variables qui auraient été des variables d'enquête si elles n'avaient pas déjà été disponibles dans les registres. Ces variables étaient le nombre d'années vécues en Norvège et le pourcentage d'immigrants arrivés en Norvège en tant que réfugié (par contraste avec d'autres causes d'immigration). En outre, nous avons examiné certaines variables d'enquête qui ont été choisies de manière à couvrir une grande gamme de types de questions :

- l'autoévaluation de l'état de santé a été choisie parce que la santé était le thème principal de l'enquête;
- le sentiment de solitude a été choisi afin d'inclure une variable dont l'évaluation est subjective;
- le nombre d'heures de travail a été choisi afin d'inclure une variable associée à des profils de comportement;
- les compétences linguistiques ont été choisies comme indicateur de l'intégration sociale.

Basée sur les variables susmentionnées, la figure 3-2 montre la valeur relative des résultats pour l'échantillon provisoire et l'échantillon brut, qui sont tous deux comparés à l'échantillon net final. L'examen des variables qui peuvent être comparées à l'information sur l'échantillon brut révèle que les résultats se sont rapprochés durant la deuxième vague de collecte, mais que même avant le suivi, ils étaient étonnamment proches des valeurs réelles. Nous voyons, par exemple, que la taille moyenne provisoire du ménage était environ 1 % plus élevée que pour l'échantillon net final. Les résultats pour les variables d'enquête ont également changé durant le suivi, mais peu.

**Figure 3-2**

**Valeur moyenne en pourcentage de la valeur pour l'échantillon net final, pour certaines variables, dans les échantillons net provisoire, net final et brut**



Nous allons maintenant examiner de plus près comment les résultats pour certaines de ces variables ont changé durant la période de collecte des données. Les graphiques des figures 3-3 et 3-4 montrent l'évolution des estimations d'enquête à mesure que le taux de réponse a augmenté. Quand les données sont disponibles, la moyenne pour l'échantillon brut est indiquée par une droite (figure 3-3).

Le graphique de gauche de la figure 3-3 révèle un sous-dénombrement des petits ménages, ce à quoi il fallait s'attendre. Il est généralement plus difficile d'entrer en contact avec les membres des petits ménages qu'avec ceux des grands ménages. Ce biais a diminué faiblement, mais régulièrement durant la période de suivi.

Fait peut-être plus étonnant, nous voyons dans le graphique de droite que les personnes ayant vécu en Norvège pendant une longue période étaient plus difficiles à recruter que celles qui n'y vivaient que depuis peu de temps. Ce résultat met en doute la notion voulant que les personnes peu intégrées ont tendance à avoir un taux de non-réponse plus élevé que celles qui sont bien intégrées. Nous constatons que les résultats ont fort peu changé durant la deuxième phase de collecte des données. Néanmoins, ils étaient un peu plus proches de la valeur pour l'échantillon brut après le suivi.

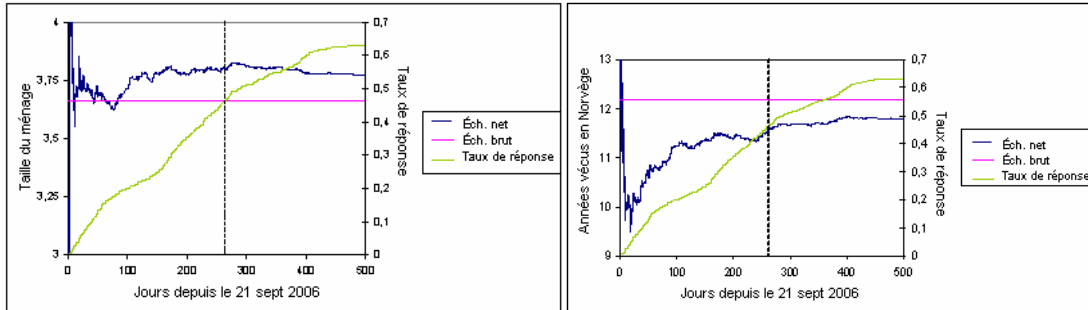
La figure 3-4 donne la valeur moyenne de certaines variables d'enquête au fil du temps. La prévalence moyenne du sentiment de solitude (graphique supérieur gauche) change à peine durant la deuxième vague de collecte. Si nous tenons compte, en particulier, du fait que la valeur moyenne est basée sur l'hypothèse simplificatrice que la solitude est une variable numérique, nous ne pouvons pas déterminer dans quelle direction cette variable évolue durant la deuxième vague.

Avant la période de suivi, 4 % seulement de répondants considéraient que leur santé était mauvaise ou très mauvaise. Si nous examinons, dans le graphique de droite de la figure 3-4, la tendance à mesure que le taux de réponse augmente, nous pourrions soupçonner que l'état de santé n'est pas si positif. Toutefois, de nouveau, la deuxième vague de collecte de données ne semble pas ajouter beaucoup d'information à cet égard.

Comme nous l'avons déjà mentionné, les personnes qui travaillaient de longues heures étaient plus difficiles à rejoindre que les autres, ce que confirme le graphique inférieur de la figure 3-4. Dans ce cas, les estimations semblent s'améliorer légèrement durant la période de collecte.

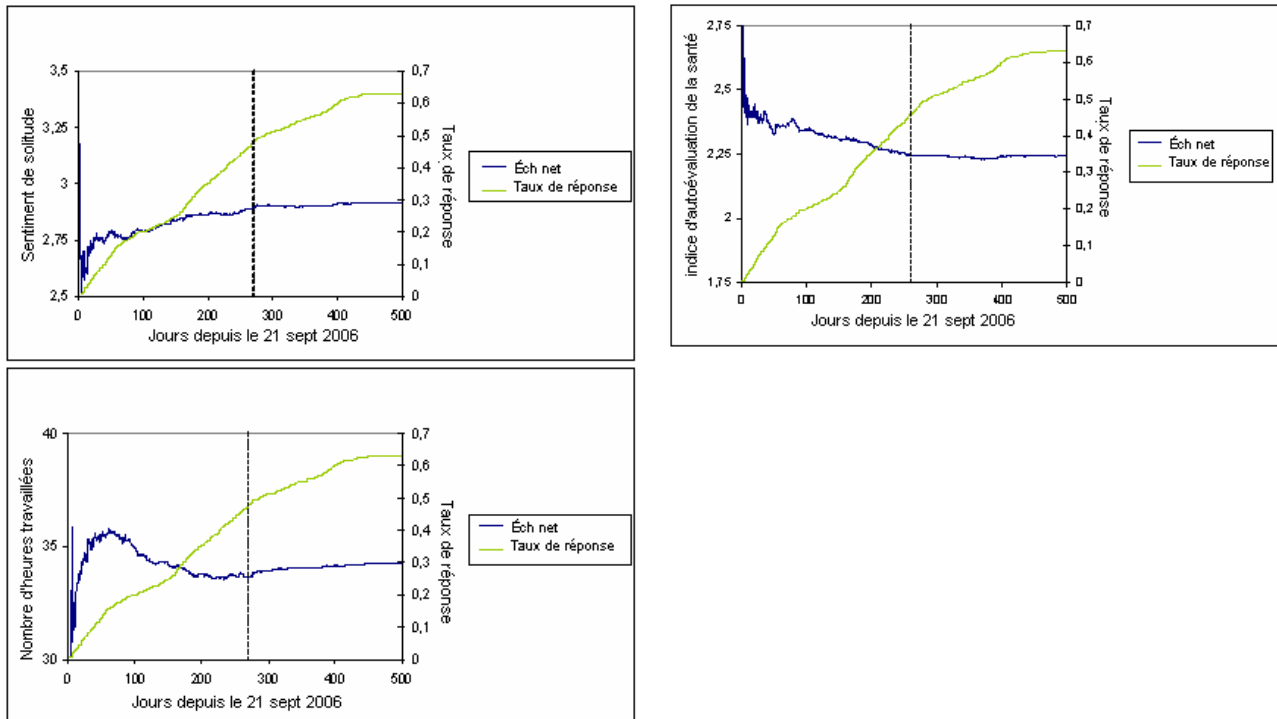
**Figure 3-3**

**Valeur moyenne de la taille du ménage (graphique de gauche), pour chaque échantillon net d'interviews achevées à chaque date (bleue) et pour l'échantillon brut (rose). Taux de réponse (vert) présenté sur le deuxième axe de y.**



**Figure 3-4**

**Valeur moyenne de l'indice de sentiment de solitude (graphique supérieur gauche), de l'indice d'autoévaluation de la santé (graphique de droite) et du nombre d'heures travaillées (graphique inférieur gauche), pour chaque échantillon net d'interviews achevées à chaque date (bleue). Taux de réponse (vert) présenté sur le deuxième axe des y.**



Si nous récapitulons jusqu'ici :

- la répartition entre les pays était meilleure dans l'échantillon net final que dans l'échantillon provisoire;
- les résultats connus pour l'échantillon brut étaient étonnamment proches de ceux obtenus de l'échantillon net, même avant le suivi;

- la plupart des résultats se sont stabilisés avant la période de suivi et très peu de choses sont survenues durant la deuxième vague de collecte des données;
- malgré cela, un biais de non-réponse affectait probablement certains résultats.

#### 4. Pondération fondée sur les répondants tardifs

Afin de justifier l'affectation d'un grand nombre de ressources au recrutement de répondants supplémentaires, nous supposons implicitement que les répondants recrutés durant le suivi seront plus semblables aux non-répondants que ceux de l'échantillon net original. C'est seulement dans ces conditions qu'un accroissement marginal du taux de réponse donnera lieu à une meilleure estimation. Cependant, pour ce qui est du suivi proprement dit, nous n'appliquons cette *hypothèse au sujet des répondants tardifs* que pour justifier son exécution. Tant que nos estimations sont fondées sur des moyennes simples, nous supposons que la moyenne est la même pour les non-répondants que pour les membres de l'échantillon net final (Bartholomew 1961).

La pondération fondée sur les répondants tardifs a pour but d'estimer les résultats sous l'hypothèse au sujet des répondants tardifs. Sous la forme proposée par Bartholomew (1961), nous supposons que les non-répondants sont équivalents aux répondants de la deuxième vague, autrement dit nous prédisons que la valeur moyenne pour les non-répondants est identique à la moyenne pour les répondants de deuxième vague. Cette pondération ne se distingue des moyennes simples que par l'hypothèse au sujet des répondants tardifs, si bien qu'elle donnera de moins bons résultats que les moyennes simples dans les situations où l'hypothèse n'est pas vérifiée.

Nous avons examiné la méthode de pondération mentionnée pour trois variables d'enquête, à savoir l'autoévaluation de la santé, le sentiment de solitude et le nombre d'années vécues en Norvège.

**Tableau 4-1**

**Valeur moyenne de l'indice d'autoévaluation de la santé, de l'indice de sentiment de solitude et du nombre d'années vécues en Norvège, en utilisant la pondération fondée sur les répondants tardifs (colonnes 3 et 4), et en utilisant l'échantillon net provisoire (colonnes 1 et 3) et l'échantillon net final (colonnes 2 et 4)**

	Résultat provisoire non pondéré	Résultat final d'enquête non pondéré	Pondération fondée sur les derniers 10 % avant l'échantillon net provisoire	Pondération basée sur les derniers 10 % avant l'échantillon net final
Autoévaluation de la santé	2,25	2,24	2,21	2,27
Sentiment de solitude	2,88	2,91	2,92	2,95
Nombre d'années vécues en Norvège	11,4	11,8	11,8	12,0
			(Valeur réelle = 12,2)	

Nous avons appliqué la pondération fondée sur les répondants tardifs en définissant les 10 % de répondants arrivés les derniers comme étant les « répondants tardifs ». Nous pouvons alors utiliser la pondération fondée sur les répondants tardifs avec et sans suivi<sup>3</sup>. Cela donne lieu à quatre options de résultats présentées dans les quatre colonnes du tableau 4-1.

<sup>3</sup> Si, au contraire, nous utilisons la pondération fondée sur les répondants tardifs uniquement pour les cas avec suivi, nous aurions défini tous les répondants arrivant durant le suivi comme étant des répondants tardifs

Statistics Norway a utilisé l'option de la colonne 2 pour la publication des résultats. Une question intéressante sur laquelle le tableau 4-1 peut jeter une certaine lumière est celle de savoir dans quelle mesure les résultats (biais) auraient été pires/meilleurs si

- le suivi avait tout simplement été abandonné et que les résultats étaient des moyennes simples basées sur l'échantillon net original (option de la colonne 1);
- le suivi avait été abandonné et remplacé par la pondération fondée sur les répondants tardifs (option 3);
- le suivi avait été exécuté, mais que la pondération par les répondants tardifs avait été utilisée au lieu des moyennes simples (option 4).

Le tableau 4-1 montre que si nous laissons tomber le suivi (option 1), les estimations pour les variables « années vécues en Norvège » et « sentiment de solitude » ne varieraient que peu et, fait plus important, que les résultats auraient pu être obtenus par pondération au lieu d'un suivi (option 3). Puisque cette dernière option est moins coûteuse, elle semble préférable si nous tirons les conclusions pour ces deux variables uniquement. Un résultat encore meilleur serait probablement obtenu en effectuant la pondération en plus du suivi (option 4). Cependant, l'amélioration est assez faible et, par conséquent, le gain de qualité est peu important comparativement au coût du suivi.

Pour l'« autoévaluation de la santé », la situation n'est pas la même que pour les deux autres variables. La caractéristique la plus intéressante est peut-être que, si le simple abandon du suivi ne change presque rien aux résultats (2,25 contre 2,24), la pondération fondée sur les répondants tardifs réduit les estimations si elle est effectuée sans le suivi, mais les *augmente* si elle est effectuée *avec* le suivi. Cela signifie que les répondants tardifs avant le suivi sont différents de ceux recrutés durant le processus de suivi. Dans l'une de ces deux situations, l'hypothèse qui sous-tend la pondération n'est manifestement pas vérifiée.

Nous obtenons une meilleure idée de ce phénomène en retournant au graphique de droite de la figure 3-3. Nous pouvons y voir que la valeur moyenne n'augmente pas ou ne diminue pas de manière monotone pour se rapprocher de la valeur vraie, mais s'accroît dans certains intervalles et baisse dans d'autres. Avant de choisir d'appliquer la pondération fondée sur les répondants tardifs, nous devrions nous assurer que la moyenne est monotone, sauf dans la première partie du graphique. L'omission de cette première partie tient au fait que, dans ce cas, nous pourrions nous attendre à ce que la variation de la moyenne soit causée par des phénomènes aléatoires.

## 5. Conclusion

Nos résultats semblent indiquer que, dans cette enquête, l'amélioration obtenue en procédant à une deuxième vague de collecte des données n'est que marginale. Cependant, nous n'avons étudié qu'une petite partie des variables pour lesquelles des données ont été recueillies. En outre, les résultats étaient, dans une large mesure, publiés pour des sous-groupes. Même si le suivi n'a pas eu d'incidence sur la moyenne de l'échantillon net total, les effets pourraient être différents pour les sous-groupes. Donc, nos données empiriques sont insuffisantes pour suggérer que la deuxième vague de collecte aurait pu être abandonnée. Notre objectif, dans le présent article, est plutôt d'attirer l'attention sur une méthode qui pourrait être étendue et utilisée pour décider si une deuxième vague de collecte des données devrait être exécutée.

La pondération fondée sur les répondants tardifs peut, dans certains cas, représenter un moyen peu coûteux d'améliorer les résultats et accomplira parfois autant que la deuxième vague de collecte. Cependant, l'hypothèse concernant les répondants tardifs n'est pas toujours vérifiée et cette méthode de pondération doit par conséquent être utilisée avec beaucoup de prudence. Certains indices de la validité de l'hypothèse peuvent être dégagés de graphiques tels que ceux de la figure 3-3. Excepté pour la première partie du graphique, la tendance devrait être monotone si nous choisissons d'utiliser cette pondération.

## Bibliographie

Bartholomew, D.J. (1961). A Method of Allowing for 'not-at-home' Bias in Sample Surveys, *Applied Statistics*, 10, 52-59.



Fosen, J. et Kleven, Ø. (2007). Non Response, Bias and Costs in the Norwegian Consumer Expenditure Survey, *18<sup>th</sup> Annual International Workshop on Household Survey Nonresponse*, Southampton, UK.

Gulløy, E. (2008). Living conditions among immigrants 2005/2006. Documentation report (en norvégien), Notater 2008/5, Statistics Norway.

Hansen, M.H. et Hurwitz, W.N. (1946). The Problem of Nonresponse in Sample Surveys, *Journal of the American Statistical Association*, 41, 517-529.

Thomsen, I.O., Kleven, Ø., Wang, J.E. et Zhang, L.C. (2006). Coping with Decreasing Nonresponse Rates in Statistics Norway. Recommended Practice for Reducing the Effect of Nonresponse, Reports 2006/29, Statistics Norway.

Thomsen, I. et Siring, E. (1983). On the Causes and Effects of Nonresponse: Norwegian Experiences, *Incomplete Data in Sample Surveys*, (Eds. W.G. Madow et Olkin) vol. 3, New York: Academic Press.