

LA CONCEPTION DE QUESTIONNAIRE DU RECENSEMENT DANS LE CADRE D'UNE SAISIE AUTOMATISÉE DES DONNÉES

Laurent Roy et Manon Monette¹

RÉSUMÉ

Il existe plusieurs éléments à considérer dans la conception de questionnaire et l'un de ceux-ci est le format du questionnaire. Cet élément requiert beaucoup d'attention dans le cadre d'entrevues auto administrées. Une étude préparatoire au Recensement du Canada de 2006 a porté sur la comparaison de différents formats (séquentiel et matriciel) pour remanier le questionnaire court et le questionnaire long du Recensement. Le remaniement des questionnaires est nécessaire pour la saisie automatisée des données en 2006. Ces formats ont été comparés par rapport aux coûts de collecte, de la qualité des données et du fardeau de réponse. Parmi plusieurs autres conclusions, les résultats de cette étude montrent, pour le questionnaire long, le taux de questionnaires complétés est élevé pour le format matriciel. La proportion de questionnaires complétés pour lesquels il y a au moins une personne avec aucune question répondue est plus grande pour le format séquentiel. De plus, avec le format matriciel, il est plus probable que le répondant commence à répondre aux questions pour chaque membre du ménage. Pour le format séquentiel, lorsque le répondant a commencé à répondre aux questions pour un membre du ménage, il est plus probable qu'il réponde à toutes les questions pour ce membre. Pour le questionnaire court, les études ont montrés qu'un questionnaire de type feuillet permettait de réduire les coûts de saisie de données tout en ne créant pas de problèmes pour les répondants.

MOTS CLÉS : Erreur d'aiguillage; fardeau de réponse; qualité des données; questionnaire séquentiel; questionnaire matriciel.

ABSTRACT

There are many elements to be considered when developing a questionnaire, one of which is the questionnaire format. This element particularly requires a lot of attention in the context of a self administered interview. A study in preparation for the Canadian Census of 2006 was conducted to compare different formats (sequential and matrix) to redesign the long and short questionnaire for the census. The redesign of census questionnaires is necessary in order to use the automated data capture in 2006. These formats have been compared in terms of collection cost, data quality as well as response burden. Among other conclusions, the results of this study show that for the long form, the rate of completed questionnaires is high for the matrix questionnaire. The proportion of completed questionnaires for which there is at least one person with no questions answered is higher for the sequential questionnaire. Also, with the matrix questionnaire, chances are higher that the respondent starts reporting for all members of the household. For the sequential questionnaire, when the respondent starts reporting for another member of the household, chances are higher that he will report entirely for this member. For the short form, studies have shown that a booklet type questionnaire led to reduced costs for data capture while not causing problems for respondents.

KEY WORDS : Branching error, Data quality, Matrix questionnaire, Response burden, Sequential questionnaire.

1-INTRODUCTION

Lorsqu'un questionnaire auto administré est utilisé dans le cadre d'un recensement ou d'une enquête par échantillon, il est important de s'assurer que ce questionnaire est simple et facile à remplir pour le répondant puisqu'il n'y a pas d'intervention de la part d'interviewers. Le format et le design sont des aspects importants pour motiver et guider le répondant. Il existe plusieurs règles quant à la conception du questionnaire pour faciliter la lecture des questions et des instructions et la navigation à travers le questionnaire. Un des éléments à considérer est le format du questionnaire.

Dans le cadre d'un recensement où plusieurs questions doivent être répondues pour chaque membre d'un ménage, il existe plusieurs formats de questionnaire possibles tels un format matriciel et un format séquentiel. Le format matriciel se voit comme une matrice où les rangées sont utilisées pour les questions et les colonnes sont utilisées pour les espaces réponses des membres du ménage (voir figure 1). Le format séquentiel présente toutes les questions pour une même personne regroupées et les questions apparaissent de façon verticale (voir figure 2). Dans le choix d'un format de questionnaire, Jenkins et Dillman (1997) mentionnent que pour le format matriciel, le répondant, pour comprendre la question et y répondre, doit faire le lien entre les titres des lignes et des colonnes et l'intersection de celles-ci. Cette difficulté peut créer un obstacle pour le répondant.

¹ Laurent Roy (Laurent.Roy@statcan.ca), Division des méthodes d'enquêtes sociales, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6., Manon Monette (Manon.Monette@statcan.ca), Direction de la statistique démographique et du recensement, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6..

Dans le cadre de la préparation au Recensement du Canada de 2006, les questionnaires ont été remaniés pour satisfaire les exigences d'une saisie automatisée des données. Des études ont eu lieu pour déterminer le meilleur format. Pour ces études, les formats matriciel et séquentiel ont été considérés et comparés. Les comparaisons ont eu lieu au niveau des coûts de collecte, de la qualité des données et du fardeau de réponse.

2. PRÉPARATION AU RECENSEMENT DU CANADA DE 2006

Le Recensement du Canada est mené tous les cinq ans. Le prochain aura lieu en 2006. Pour le recensement de la population des ménages privés, deux types de questionnaire auto administré sont distribués : un questionnaire court (80% de la population) et un questionnaire long (20% de la population). Le questionnaire court comprend sept questions portant sur des renseignements démographiques (sexe, date de naissance, langue). Le questionnaire long contient près de cinquante questions pour chaque membre du ménage : les questions du questionnaire court et des questions sur l'éducation, l'occupation, la connaissance des langues, le revenu, etc. Il y a aussi une section de seize questions sur les caractéristiques du logement. Depuis le Recensement de 1991, les questionnaires courts et longs étaient de format matriciel avec les rangées utilisées pour les questions et les colonnes utilisées pour les réponses.

Lors du Recensement du Canada de 2006, il est prévu que les techniques de lecture optique et de reconnaissance de texte seront utilisées pour la saisie des données. Le questionnaire court actuel permet de faire cette opération avec quelques changements mineurs. Par contre, lors de la planification de la saisie des données, il a été décidé que des changements possibles au questionnaire court pourraient faciliter cette opération et en réduire les coûts. Diverses études qualitatives ont eu lieu pour étudier différents formats du questionnaire court et les résultats sont présentés à la section 4. La situation est différente pour le questionnaire long. La principale lacune est que les espaces réponses des questions ouvertes n'étaient pas de longueur suffisante pour incorporer la segmentation nécessaire pour obtenir une saisie automatique efficace. Un remaniement du questionnaire long était donc absolument nécessaire. Lors de ce remaniement, les formats matriciel et séquentiel ont été considérés et ont été comparés selon les coûts de collecte, la qualité des données et le fardeau de réponse.

1 NOM
Sélectionnez votre nom de famille. Cliquez sur votre nom de famille le mieux adapté à l'étape 2. Répondez ensuite aux questions suivantes pour chaque personne.

2 SEXE
 Masculin
 Féminin

3 DATE DE NAISSANCE
Exemple : Jour Mois Année
2 3 0 2 1 9 x 4
Il s'agit de votre date de naissance. Répondez à cette question en utilisant le format suivant.

4 ETAT MATRIMONIAL
Cliquez sur l'état marital qui s'applique.

5 CETTE PERSONNE VIT-ELLE AVEC UN PARTENAIRE EN UNION LIBRE?
Cliquez sur Oui ou Non.

6 LIEN AVEC LA PERSONNE 1
Cliquez sur l'état qui décrit le mieux votre lien avec la personne 1.

PERSONNE 1
Nom de famille
Prénoms
Sexe
Date de naissance
État matrimonial
Vie avec un partenaire en union libre
Lien avec la personne 1

PERSONNE 2
Nom de famille
Prénoms
Sexe
Date de naissance
État matrimonial
Vie avec un partenaire en union libre
Lien avec la personne 1

Figure 1 : Exemple du format matriciel standard testé.

PERSONNE 1

1 Quel est le nom de la personne 1?
Nom de famille
Prénoms

2 Quel est le sexe de la personne 1?
 Masculin
 Féminin

3 Quelle est la date de naissance de la personne 1?
Exemple : 2 3 0 2 1 9 x 4
Jour Mois Année

4 Quel est l'état matrimonial de la personne 1?
Cliquez sur l'état marital qui s'applique.

5 La personne 1 vit-elle avec un partenaire en union libre?
Cliquez sur Oui ou Non.

6 Quel lien a la personne 1 avec la personne 1?
Cliquez sur l'état qui décrit le mieux votre lien avec la personne 1.

ACTIVITES DE LA VIE QUOTIDIENNE

7 La personne 1 a-t-elle un emploi à temps plein, à temps partiel, ou est-elle sans emploi?
 Oui, plein-temps
 Oui, temps partiel
 Non

8 Quel est le statut de la personne 1?
 Oui, plein-temps
 Oui, temps partiel
 Non
 Sans emploi

Figure 2 : Exemple du format séquentiel testé

Lors des dernières années, plusieurs études ont eu lieu pour améliorer le questionnaire long et permettre la lecture optique de celui-ci pour la saisie des données. Des études qualitatives à petite échelle (environ 100 répondants) ont été faites dans le but de déterminer des formats convenables du questionnaire long pour la saisie des données et pour les répondants. Lors de ces études, divers formats de questionnaires matriciels et séquentiels ont été mis à l'essai. De façon générale, on a remarqué lors d'observations de répondants que certains ont plus de difficultés à comprendre la structure du format matriciel. Ils hésitent en voyant ce format et doivent réfléchir pour comprendre que chaque colonne correspond à une personne et que chaque ligne correspond à une question. Après hésitation, cette difficulté est surmontée pour la grande majorité d'entre eux. Pour le format séquentiel, les répondants n'ont pas ce moment d'hésitation, mais plusieurs semblent découragés à devoir recommencer la séquence de questions pour les autres membres du ménage lorsqu'ils ont complété les questions du premier membre. Relire les questions et les instructions semble créer un fardeau pour ces répondants, particulièrement pour les ménages de grande taille.

Ces études qualitatives ont permis de retenir trois formats de questionnaire qui semblaient convenables pour les répondants. De ces trois formats, deux sont des questionnaires matriciels. Un de ces formats matriciels est de format standard (pages 8,5 pouces de largeur x 11 pouces de hauteur ou 21,6 cm x 27,9 cm), c'est-à-dire le même format que les recensements précédents, et l'autre est un format dit agrandi (pages 14 pouces de largeur x 11 pouces de hauteur ou 35,6cm x 27,9cm). Selon les répondants, le principal avantage de ce dernier est qu'il y a plus d'espacement entre les questions et les espaces réponses et donc qu'il est plus facile à lire. L'autre format retenu est un format séquentiel où les questions sont regroupées pour chaque membre du ménage. La figure 1 montre un exemple de format matriciel standard testé et la figure 2 montre un exemple de format séquentiel testé.

3. ÉTUDE QUANTITATIVE SUR LE FORMAT DU QUESTIONNAIRE LONG

3.1 Description des formats de questionnaire testés

Les études qualitatives ont permis de déterminer trois formats de questionnaire long pouvant être testés dans une étude de plus grande envergure (deux formats matriciels et un format séquentiel). Pour mieux analyser les résultats d'une telle étude, un questionnaire contrôle, le format du questionnaire long utilisé en 2001, a été inclus dans cette étude. De plus, il a été décidé de tester une variation de la segmentation des espaces réponses du format séquentiel et ainsi créer deux formats séquentiels.

Au total, cinq formats de questionnaires ont été testés et ils se regroupent en deux classes : format matriciel et format séquentiel. Pour tous ces formats, le questionnaire est organisé dans un livret. La page couverture est identique pour chaque format. La question et les instructions sur la liste des membres du ménage sont placées au début pour chaque format. Le tableau 1 résume les caractéristiques des formats. Pour le format matriciel, les questions sur le logement sont placées à la fin. Pour le format séquentiel, à cause des contraintes, la section des questions sur le logement est placée après la liste des membres du ménage et avant les questions personnes. Chaque format a des espaces réponses pour un maximum de cinq personnes, sauf le matriciel 2001 dont les espaces réponses permettent d'inscrire au plus six personnes.

Tableau 1 Caractéristiques des formats de questionnaire testés

Format	Dimension d'une page	Nombre de pages	Description des espaces réponse pour les questions ouvertes
1-Matriciel 2001	8,5 pouces x 11 pouces (21,6cm x 27,9cm)	32	Pas de segmentation.
2-Matriciel standard	8,5 pouces x 11 pouces (21,6cm x 27,9cm)	40	11 cases par ligne réponse
3-Matriciel agrandi	14 pouces x 11 pouces (35,6cm x 27,9cm)	32	15 cases par ligne réponse
4-Séquentiel A	8,5 pouces x 11 pouces (21,6cm x 27,9cm)	40	14 cases séparées par ligne réponse
5-Séquentiel B	8,5 pouces x 11 pouces (21,6cm x 27,9cm)	40	17 cases jointes par ligne réponse

Pour bien mesurer l'effet du format et ne pas avoir d'interférence en comparant ces cinq formats, le contenu du questionnaire est identique entre les formats. De façon générale, le design des questions, des choix de réponses, des instructions et des aiguillages n'a pas été modifié. Les questions sont ordonnées de la même façon pour chaque format, mais pas nécessairement sur les mêmes pages. Pour tous les formats, sauf le matriciel 2001, les espaces réponses des questions ouvertes ont été segmentées. La segmentation est nécessaire pour permettre de façon efficace la saisie des données lors de la lecture optique et reconnaissance de texte. Un trait mince sépare chacune des cases de l'espace réponse segmenté. Comparativement au format matriciel standard, le format matriciel agrandi permet d'avoir plus d'espacement pour les espaces réponses et d'offrir plus de cases pour ces espaces réponses segmentés. Pour le format séquentiel, il a été décidé de tester deux types d'espace réponse segmenté : segmentation avec des cases jointes et segmentation avec des

cases séparées. Pour les cases séparées, un espace, plus large qu'un trait, sépare chacune des cases. Les cases jointes permettent, pour un espace réponse précis, d'avoir plus de cases que si des cases séparées étaient utilisées. Par rapport aux cases jointes, les cases séparées auraient l'avantage d'inciter le répondant à ne pas utiliser de lettres attachées et donc à augmenter la qualité de la saisie des données lors de la reconnaissance de texte. Les cases jointes auraient l'avantage d'offrir plus de cases pour un espace réponse donné, et cela pourrait permettre au répondant d'élaborer sa réponse et fournir plus de détails sans avoir à écrire en dehors des espaces réponses.

3.2 Méthodologie de l'étude

Pour réaliser cette étude, cinq échantillons indépendants de 4 200 logements privés chacun ont été sélectionnés à partir du Registre des adresses de Statistique Canada. La collecte a eu lieu approximativement une année après le Recensement de 2001. Un premier envoi postal a eu lieu simultanément pour chaque format. Des suivis téléphoniques ont été faits dix jours après le début de la collecte pour augmenter le taux de réponse. Ces suivis ont été faits de façon uniforme entre les formats testés. Contrairement au Recensement, la participation était volontaire. La longueur de la période de collecte a été similaire à celle du Recensement et identique pour chaque échantillon. Les poids d'estimation ont été calculés pour les questionnaires retournés et considérés complétés et ils ont été ajustés pour tenir compte de la non-réponse. Tous les résultats présentés ont été calculés en utilisant les poids d'estimation.

3.3 Taux de réponse au niveau ménage

Le taux de questionnaires complétés est un élément déterminant pour les coûts de collecte du Recensement. Il s'agit donc d'un critère important dans la recommandation d'un format de questionnaire pour cette étude. Lors du Recensement, environ 2,4 millions de ménages reçoivent un questionnaire long. Chaque ménage qui ne retourne pas le questionnaire fait l'objet d'un suivi et ces suivis comptent pour une part importante du budget alloué à la collecte. Chaque point de pourcentage pour la proportion de questionnaires retournés représente pour le Recensement environ 24 000 ménages pour lesquels un suivi n'est pas nécessaire. Le tableau 2 montre le taux de questionnaires retournés et considéré complétés pour chaque format. Le format semble être un facteur déterminant pour le taux de questionnaires complétés. Ainsi, il existe deux groupes distincts : format matriciel et format séquentiel. Dans chacun de ces groupes, les taux de questionnaires complétés sont similaires, mais il existe une différence significative entre ces deux groupes, donnant un avantage au format matriciel. Le facteur déterminant dans cette différence est la taille du ménage. Ainsi, en utilisant les tailles de ménages du Recensement de 2001 pour les logements de l'échantillon du test, il nous est possible de comparer le taux de questionnaires complétés selon la taille du ménage. Une telle analyse montre que pour les ménages de cinq personnes ou plus, le taux de questionnaires complétés est 11% plus élevé pour le format matriciel versus le format séquentiel. Pour les ménages de une ou deux personnes, cette différence n'est que de 2%. Dans les deux cas, la différence est significative au seuil de 1%.

Tableau 2 : Proportions de questionnaires retournés et considérés complétés selon le format

Format de questionnaire	Proportions de questionnaires retournés et considérés complétés
1-Matriciel 2001	48,5%
2-Matriciel standard	47,6%
3-Matriciel agrandi	47,6%
4-Séquentiel A	42,3%
5-Séquentiel B	43,9%
Total Matriciel	47,9%
Total Séquentiel	43,1%

Pour expliquer la différence entre le taux de questionnaires complétés des formats séquentiel et matriciel, il faut se rappeler que le format matriciel a été utilisé lors des trois derniers recensements. Il peut donc y avoir un effet d'habitude ou de reconnaissance du format matriciel de la part des répondants. D'autres facteurs peuvent expliquer les différences entre les taux de questionnaires complétés. Il est plausible de penser que pour le format séquentiel, le répondant d'un ménage de plusieurs personnes peut se décourager d'avoir à recommencer la séquence de cinquante questions après la première personne. Dans ce cas, il peut décider de ne pas retourner le questionnaire. Avec le même exemple mais pour un format matriciel, si le répondant décide d'arrêter, les premières questions seront répondues pour tous les membres. Il peut lui sembler plus logique de retourner le questionnaire dans ce cas-ci que dans l'exemple précédent. Le taux de questionnaire remplis selon la taille du ménage semble soutenir cette hypothèse.

La localisation des questions sur le logement peut aussi expliquer la différence entre les taux de questionnaires complétés des formats. Selon Dillman (1978), il est important de placer les questions faciles au début du questionnaire. Si les répondants ont de la facilité à répondre aux premières questions, les craintes potentielles sur le fait que les questions suivantes peuvent être difficiles sont amoindries. La section des questions logement compte environ seize questions dont certaines demandent un effort pour y répondre. Cette section se retrouve avant les questions personnes pour le format séquentiel, donc au début, tandis qu'elle se trouve à la fin pour le format matriciel. Dans le cas où le répondant peut se sentir découragé de répondre à une telle section, l'impact sera différent pour les deux formats sur le taux de questionnaires complétés : pour le format séquentiel, le répondant peut complètement abandonner avant même d'avoir répondu une question personne. Notons que la localisation de ces questions est inhérente à chacun des formats. Des tests précédents ont permis de conclure que pour le format séquentiel, la localisation de cette section à la fin du questionnaire apportait un problème d'incompréhension de l'aiguillage indiquant de poursuivre au besoin pour le membre suivant du ménage ou de continuer avec les questions logements.

3.4 Taux de non-réponse au niveau personne

Un taux de questionnaires complétés plus faible pour le format séquentiel peut s'expliquer par la possibilité qu'un répondant puisse se sentir découragé d'avoir à recommencer la séquence de toutes les questions. S'il arrête de répondre après la première ou les premières personnes, il pourrait décider de ne pas retourner le questionnaire ou remettre à plus tard le retour puisqu'il considère la tâche incomplète. Une façon de soutenir cette hypothèse est de mesurer le taux d'incidence de ce comportement pour les questionnaires retournés. Le taux de questionnaires complétés pour lesquels il y a au moins un membre du ménage pour lequel aucune question n'a été répondue (non réponse totale au niveau personne) a été calculé pour chaque format (voir tableau 3). Ce tableau donne aussi l'estimation du nombre de personnes pour lesquelles aucune question n'a été répondue. À titre d'information, la population totale estimée à partir de chacun des cinq échantillons est de 17 millions.

Tableau 3 : Taux de questionnaires complétés avec de la non-réponse totale au niveau personne selon le format

Format	Taux de questionnaires complétés avec de la non-réponse totale au niveau personne	Estimations du total de personnes avec aucune question répondue (en milliers)
1-Matriciel 2001	0,2%	12
2-Matriciel standard	0,1%	11
3-Matriciel agrandi	0,5%	60
4-Séquentiel A	1,9%	276
5-Séquentiel B	1,8%	199

La différence entre les formats séquentiels A et B est significative avec les formats matriciel 2001 et matriciel standard au seuil de 1%. Cette différence s'explique surtout par les ménages de grande taille (4 personnes ou plus). Le nombre de personnes pour lesquelles aucune question n'a été répondue est beaucoup plus grand pour le format séquentiel et cela semble confirmer l'hypothèse énoncée auparavant.

3.5 Qualité des données pour les questionnaires complétés

Après avoir comparé les taux de réponse, il faut considérer la qualité des données des questionnaires complétés. Celle-ci a été mesurée sous deux aspects : le nombre de questions non répondues par questionnaire complété et le nombre de questionnaires pour lesquels un suivi serait nécessaire à cause d'information incomplète. Le tableau 4 donne le nombre moyen de questions non répondues par questionnaire considéré complété. Le nombre a été calculé en tenant compte des questions nécessitant une réponse selon qu'il y ait un aiguillage ou non et en tenant compte des membres de la liste du ménage pour lesquels aucune question n'a été répondue.

Selon le tableau 4, le format séquentiel a moins de questions non répondues que le format matriciel. Au total, il y a environ deux questions non répondues de plus sur le format matriciel standard par rapport aux deux formats séquentiel. En excluant la section sur le logement, la différence entre le nombre de questions non répondues du format matriciel standard et des formats séquentiel n'est pas significative au seuil de 1%. La différence pour la section des questions sur le logement est probablement attribuable à la localisation de cette section. On pourrait penser qu'il ne s'agit pas d'un effet du format. Pourtant, c'est le format qui impose d'avoir cette section au début pour le questionnaire séquentiel et à la fin pour le questionnaire matriciel. On peut aussi envisager la possibilité de placer les questions sur le logement sur un feuillet séparé. De cette façon, c'est le répondant qui décide s'il répond à ces questions au début ou à la fin.

Il peut sembler surprenant de voir un plus grand nombre de questions non répondues pour le format matriciel agrandi par rapport aux autres formats matriciels. L'observation de répondants lors de tests qualitatifs a permis de constater que le désavantage d'un format matriciel se trouve qu'il est plus difficile de suivre les rangées et les colonnes pour remplir le questionnaire, spécialement pour les personnes #2 à #5. Lors d'observations de répondants, cet élément de difficulté a semblé plus important pour le format matriciel agrandi. La distance entre l'espace réponse et la question est plus importante et cela semble accentuer ce désavantage.

Tableau 4 : Nombre moyen de questions non répondues par questionnaire complété selon le format de questionnaire

Format	Moyenne de questions non répondues par questionnaire considéré complété		
	Total	Questions personne	Questions logement
1-Matriciel 2001	14,4	12,3	2,2
2-Matriciel standard	13,7 ²	11,4 ³	2,2
3-Matriciel agrandi	16,0	13,6	2,4
4-Séquentiel A	12,1	10,6	1,5
5-Séquentiel B	11,5	10,0	1,5

Lors du Recensement de 2001, les suivis ont été déterminés selon certaines règles de vérification. Les règles indiquent qu'un questionnaire complété doit faire l'objet d'un suivi si certaines questions importantes n'ont pas été répondues ou si le nombre total de questions non répondues dépasse une limite pré-établie. Pour notre test, le nombre de questionnaires nécessitant un suivi a été calculé pour chaque format. Le but était de mesurer s'il y avait un format pour lequel moins de suivis seraient nécessaires lors d'une éventuelle collecte. Le tableau 5 donne le nombre total de suivis nécessaires pour les questionnaires complétés selon les spécifications de suivi utilisées lors du Recensement de 2001. Pour cette mesure, les estimations montrent que la proportion de questionnaires nécessitant un suivi est plus faible pour le format séquentiel. Également, en analysant les données, on a noté que la distribution des questions non répondues est différente entre le format matriciel et séquentiel. En résumé, pour le format séquentiel, lorsqu'il y a des questions non répondues, il s'agit plus souvent du cas où aucune question n'est répondue pour un membre du ménage. Pour le format matriciel, lorsqu'il y a des questions non répondues, il s'agit plus souvent du cas où quelques questions ne sont pas répondues pour tous les membres.

Tableau 5 : Taux de questionnaires complétés nécessitant un suivi selon le format

Format de questionnaire	Taux de questionnaires complétés nécessitant un suivi
1-Matriciel 2001	37,6%
2-Matriciel standard	34,5%
3-Matriciel agrandi	41,1%
4-Séquentiel A	29,4%
5-Séquentiel B	28,3%

Le tableau 5 indique aussi un résultat important sur la qualité de cette étude et sa validité. Le taux de 37,6% de suivis nécessaires auprès des questionnaires remplis pour le format matriciel 2001 est similaire à celui obtenu lors du Recensement de 2001 (environ 38%). Cette observation semble indiquer que les répondants du test sont similaires aux répondants du Recensement. Dans le cas où les répondants du test auraient été considérés comme de meilleurs répondants que ceux du Recensement, il aurait été logique de penser que la non-réponse au niveau des questions aurait été plus faible que pour le Recensement. Puisque le taux de suivis est basé sur la non-réponse au niveau des questions et que ce taux est le même pour le Recensement de 2001 et le format matriciel 2001 (format contrôle du test), on peut donc penser que les répondants du test sont représentatifs des répondants du Recensement.

Les résultats précédents permettent de tirer une conclusion par rapport au format. Pour le format matriciel, il est plus probable, par rapport au format séquentiel, que le répondant commence à répondre aux questions pour une personne listée dans le ménage. Pour le format séquentiel, lorsque le répondant a commencé à répondre aux questions pour une personne listée, il est plus probable, par rapport au format matriciel, que celui-ci réponde à toutes les questions. Du point de vue de la qualité des données, le format matriciel semble donner un avantage. Ce format permet d'obtenir plus de personnes pour lesquelles les premières questions sont répondues. Dans le cas du format séquentiel, il y a plus de personnes avec toutes

² "P-values" pour la comparaison Matriciel standard versus chaque format séquentiel A et B sont respectivement 2,6% et 0,2%.

³ "P-values" pour la comparaison Matriciel standard versus chaque format séquentiel A et B sont respectivement 22,2% et 2,9%.

les questions répondues, mais il y a également plus de personnes avec aucune question répondue. Pour la qualité des données, le premier cas est préférable car l'imputation des questions non répondues est plus efficace lorsqu'on a certaines données démographiques (les premières questions) par rapport au cas des personnes pour lesquelles aucune question n'est répondue.

3.6 Fardeau de réponse selon le format

Le fardeau de réponse est un élément important aux yeux d'un répondant. Il mesure l'effort nécessaire pour compléter le questionnaire. Cet effort se mesure sous plusieurs aspects : le temps requis pour compléter le questionnaire, la complexité du questionnaire et l'effort nécessaire pour comprendre la question et y répondre. La compréhension de la question et l'effort pour y répondre n'est pas un élément déterminant de cette étude. Les questions étant identiques pour chaque format, ce type de fardeau devrait donc être identique.

Lors de cette étude, aucune donnée directe n'a été récoltée sur le temps nécessaire à chaque répondant pour compléter le questionnaire. Par contre, des études qualitatives passées ont permis de constater en observant des petits échantillons de répondants que le temps moyen pour compléter un format matriciel est environ cinq minutes de moins que le format séquentiel. Le temps requis peut aussi se mesurer par le nombre de questions répondues en trop. Il y a une question répondue en trop lorsqu'un aiguillage a été manqué par le répondant et qu'une question est répondue alors qu'elle ne devait pas l'être. Ainsi, plus il y a de questions répondues en trop, plus le temps requis pour compléter le questionnaire est important, en supposant que les autres facteurs tels que le nombre de personnes, le contenu des réponses et la non-réponse aux questions soient identiques.

Le questionnaire du Recensement comporte plusieurs aiguillages. Il y a d'abord un aiguillage selon l'âge : les questions #1 à #21 s'adressent à toutes les personnes et les questions #22 jusqu'à #51 s'adressent aux personnes âgées de 15 ans et plus. Pour le format matriciel, dans les pages suivant cet aiguillage, un rappel d'aiguillage figure au haut des pages et donc sous les noms des personnes. Le format séquentiel n'a pas de tel rappel. Il est connu que le format séquentiel facilite la navigation par rapport à un format matriciel pour les aiguillages. Il avait été donc supposé qu'une seule instruction sur cet aiguillage serait suffisante. En plus d'un aiguillage relatif à l'âge, il existe huit questions avec aiguillage : six de ces aiguillages n'impliquent qu'une question seulement (une question à répondre ou à sauter) et deux aiguillages impliquent un groupe de questions. Le nombre de questions répondues en trop a été calculé pour chaque format par rapport à ces aiguillages. Le design des instructions des aiguillages est identique pour chaque format de questionnaire. Ces comparaisons permettent donc de voir l'impact du format par rapport à la navigation entre les questions lorsqu'il y a des aiguillages.

L'analyse des résultats sur l'aiguillage relatif à l'âge montre que la proportion de personnes de moins de 15 ans pour lesquelles cet aiguillage est manqué est 26% pour le format matriciel comparativement à 32% pour le format séquentiel (différence significative au seuil de 5%). Lorsque l'aiguillage est manqué, il y a en moyenne 7 questions répondues en trop pour le format matriciel contre 14 pour le format séquentiel. Selon ces résultats, il était donc erroné de penser que pour le format séquentiel, il n'était pas nécessaire de faire de rappel sur l'aiguillage relatif à l'âge. Ce résultat montre aussi que l'aiguillage est plus visible dans le format matriciel : l'instruction de cet aiguillage est répétée pour chaque personne au haut de chaque page suivant cet aiguillage.

Le tableau 6 montre le total de questions répondues en trop selon les deux types d'aiguillages : l'aiguillage selon l'âge et les autres aiguillages. Pour l'aiguillage de l'âge, le nombre de questions répondues en trop est beaucoup plus élevé pour les formats séquentiels. Pour les autres aiguillages, le nombre de questions répondues en trop est similaire, sauf pour le format matriciel standard. La différence est principalement due à un aiguillage d'une question portant sur l'occupation. Pour le format matriciel standard, les questions reliées à l'aiguillage se retrouvent sur deux pages distinctes, alors que pour les autres formats, ces questions sont sur une même page. Il apparaît donc plus difficile de naviguer avec cet aiguillage pour le format matriciel standard et on retrouve ainsi plus de questions répondues en trop. Il est donc préférable que si un aiguillage fait référence à un groupe de questions, ces questions soient placées sur une même page pour faciliter la navigation. Malgré cela, selon la colonne total, le fardeau de réponse due aux questions répondues en trop est plus faible pour chacun des formats matriciels par rapport aux deux formats séquentiels.

Les résultats sur les aiguillages non relatifs à l'âge montrent que peu importe le format, la navigation avec des aiguillages pose problème de façon égale pour le nombre de questions répondues en trop. Mais qu'en est-il du cas des questions non répondues suite à un aiguillage ? L'analyse a montré que pour le format séquentiel, il y a environ en moyenne 0,12 erreur

d'aiguillage de ce type par personne, alors que pour le format matriciel, cette moyenne est de 0,20. Globalement, les erreurs de aiguillages ne sont pas plus fréquentes pour un format matriciel qu'un format séquentiel.

Tableau 6: Nombre total de questions répondues en trop après des aiguillages manqués selon le format de questionnaire

Format	Total de questions répondues en trop suite à des aiguillages manqués (en millions)		
	Aiguillage relatif à l'âge	Autres aiguillages	Total
1-Matriciel 2001	5,8	26,4	32,2
2-Matriciel standard	4,0	32,6	36,6
3-Matriciel agrandi	5,8	24,8	30,6
4-Séquentiel A	13,6	28,0	41,6
5-Séquentiel B	14,2	26,4	40,6

3.7 Qualité des réponses pour les questions ouvertes.

L'avantage d'un format séquentiel ou matriciel agrandi se retrouve au niveau des questions ouvertes car il y a en général moins de contraintes pour la taille et le design des espaces réponses. Dans un format matriciel standard, on doit s'en tenir à l'espace de la colonne pour chaque personne et cet espace est plutôt restreint comparativement à un format matriciel agrandi. Dillman et Christian (2002) mentionnent que réduire la taille des espaces réponses à des questions ouvertes a pour effet de réduire le nombre de mots des réponses. Cependant, ce résultat n'est pas vrai pour des questions où on s'attend à une réponse brève. Le tableau 1 donne pour chaque ligne réponse de chaque format, le nombre de cases par ligne réponse que le répondant peut utiliser. Donc, en offrant des espaces réponses plus grands, peut-on s'attendre à voir des réponses plus longues et plus détaillées de la part des répondants ? Si les réponses sont plus longues, est-ce que la qualité de celles-ci est meilleure ? Pour tenter répondre à ces questions, deux analyses ont été faites.

Une première analyse a été faite en comparant pour chaque format le nombre de caractères dans les réponses aux questions ouvertes. Les résultats montrent que pour la plupart des questions ouvertes, sauf pour quelques questions portant sur la description de l'emploi, la longueur des réponses est similaire entre les formats. Un élément qui peut expliquer ce résultat est que la plupart des questions ouvertes telles que les questions sur la langue, le lieu de naissance ou la citoyenneté peuvent être répondues par une réponse brève pour laquelle il n'est pas nécessaire d'élaborer.

Une seconde analyse a été faite pour voir la qualité des réponses des questions ouvertes dans le but de mesurer l'impact du design des espaces réponses. Les réponses aux questions ouvertes portant sur le lieu de naissance, la langue maternelle, la connaissance des langues, l'origine ethnique et le principal domaine d'études ont été soumises au système de codification automatisé. Ce système permet d'assigner un code selon la question et la réponse en se référant à un dictionnaire des codes et des réponses possibles correspondantes. Lors de cette analyse, deux éléments ont été étudiés pour chaque format : la proportion de réponses pour lesquelles aucun code n'a été assigné et la distribution des réponses codées. Le premier élément permet de vérifier l'hypothèse que plus l'espace réponse est petit, plus le répondant tend à abrégé sa réponse et cela pourrait affecter la codification automatisée (aucun code assigné si la réponse est trop courte). Le second élément permet de vérifier que le design de l'espace réponse n'a pas d'impact sur le contenu de la réponse après codage.

Pour ces analyses, les résultats pour la plupart des questions montrent que le taux d'insuccès du codage automatisé est similaire entre les formats. Le design et le nombre de cases des espaces réponses n'affectent donc pas la capacité du système à reconnaître les réponses. Donc, même si le répondant utilise une réponse abrégée ou non, le système de codage automatisé peut assigner un code correspondant peu importe le design ou le nombre de cases. De plus, la comparaison de la distribution des codes assignés pour les questions étudiées n'est pas significativement différente entre les formats, c'est-à-dire que le contenu des réponses codées se ressemble. Pour les deux formats séquentiels, les résultats sur le codage automatisé sont similaires entre les deux types de cases. L'utilisation de cases jointes ou cases séparées n'a pas d'impact sur les taux de succès du codage automatisé. Une étude supplémentaire à partir d'un échantillon d'images de questionnaires séquentiels complétés a également permis de conclure que le nombre de réponses en lettres attachées n'est pas significativement différent entre les deux types de cases.

3.8 Conclusion pour l'analyse du format du questionnaire long.

Puisque l'enquête est volontaire, qu'il n'y a pas eu de programme de communication et de publicité comme pour le Recensement et que le taux de réponse est plus faible que pour le Recensement, il est possible de se demander si les répondants ont agité de la même façon que s'il s'agissait du Recensement. De façon générale, les répondants sont

représentatifs de la population, à l'exception des ménages de grande taille et des personnes âgées de 15 à 34 ans sur le marché du travail. Ces groupes étaient sous représentés et un ajustement à la pondération a été fait pour en tenir compte.

Selon l'analyse, le format matriciel agrandi offre, pour certains éléments, d'aussi bons résultats que le format matriciel standard. Par contre, en terme de production et de collecte, il existe certains risques tels que les coûts d'envoi, la nécessité de plier ce questionnaire lors de l'envoi et l'impact possible de ce pli lors de la saisie des données. Pour ces raisons, le format matriciel agrandi n'a pas été retenu puisque les résultats ne lui donnaient pas un avantage marqué. Quant au format séquentiel, il n'y a pas de preuves que ce format permettrait de réduire les coûts de collecte ou d'offrir une meilleure qualité de données. Les résultats tendent à prouver le contraire. De plus, en considérant le fait qu'un format matriciel a été utilisé depuis les trois derniers recensements, le format matriciel standard, sera donc le format utilisé pour le questionnaire long du Recensement du Canada 2006.

4. ÉTUDES QUALITATIVES SUR LE FORMAT DU QUESTIONNAIRE COURT

Lors du Recensement de 2001, le questionnaire court était présenté selon le format d'un livret bilingue avec un format standard pour chaque page, c'est-à-dire des pages de 8,5 pouces x 11 pouces (21,6 cm par 27,9 cm). Dans ce livret, les éléments du questionnaire étaient présentés selon l'ordre suivant : la page couverture avec l'identification géographique du logement, la page 2 avec les instructions sur qui inclure et exclure dans la liste des résidents habituels du logement, la page 3 avec la liste des résidents habituels du logement, les pages 4 et 5 avec les questions pour chaque résident habituel du logement et la page 6 avec un espace pour les commentaires et des renseignements généraux. Le livret avait donc un total de 6 pages du côté français et 6 pages du côté anglais.

Dans le but de faciliter les opérations d'impression et de saisie des données, il a été décidé que la présentation d'un questionnaire court sous format d'un feuillet unilingue serait préférable. Le terme feuillet désigne ici un questionnaire imprimé sur une même feuille et pour lequel il n'y a pas d'assemblage requis de plusieurs feuilles. Il faut noter que même si le feuillet est unilingue, il était convenu que chaque ménage recevrait dans une même enveloppe deux feuillets séparés pour le prochain Recensement : un feuillet français et un feuillet anglais. Le principal avantage d'un feuillet par rapport à un livret, est que, pour l'impression, il n'est pas nécessaire d'avoir plusieurs feuilles à assembler. De plus, lors de la saisie automatisée, il serait nécessaire de guillotiner les pages d'un questionnaire sous format livret pour le passage lors du balayage optique. Cette opération de guillotinage n'est pas nécessaire avec un questionnaire sous forme de feuillet. Ces deux avantages sont assez importants pour avoir un impact considérable sur les coûts d'impression et de saisie automatisée des données d'un format livret par rapport à un format feuillet.

Plusieurs types de questionnaires feuillets ont donc été considérés, et pour chacun, les questions sur les personnes étaient présentées sous un format matriciel. Une étude plus approfondie précédente avait permis de conclure que pour le questionnaire court, le format matriciel était préférable (Gower 1997). Les options considérées portaient donc sur la taille du feuillet et sur la façon dont ce feuillet présentait les éléments du questionnaire. Le design initial du feuillet comportait une contrainte : il devait permettre de présenter tous les éléments du livret utilisé en 2001. À cause de cette contrainte, un seul format de feuillet a été retenu, c'est-à-dire un feuillet imprimé sur une feuille de dimension 26 pouces par 11 pouces (66 cm par 27,9 cm) soit l'équivalent de trois feuilles de taille standard côte à côte (3 x 8,5 x 11 pouces). De façon générale, deux formats de feuillets ont été retenus et testés : feuillet de format accordéon et feuillet de format plié. Les deux types de feuillet ont la même taille, mais la principale différence se trouve dans la séquence des éléments présentés. Le tableau 7 présente pour chaque type de feuillet, cette séquence selon le feuillet et selon le côté de feuille.

Pour chaque format considéré, des tests qualitatifs ont eu lieu pour évaluer les problèmes potentiels que les répondants pouvaient rencontrer. Pour chaque test, environ 80 répondants ont été sélectionnés. Pour chaque répondant, une entrevue à la résidence du répondant avait lieu et le répondant devait compléter le questionnaire. Le but était de connaître la réaction des répondants en complétant et en manipulant le questionnaire. Par la suite, une entrevue était conduite pour obtenir la réaction du répondant face à divers éléments dont le format. Lors de ces tests, les éléments suivants ont été notés par rapport aux deux types de feuillets. Les répondants n'ont pas de problème à ouvrir le feuillet plié, mais les instructions sur qui inclure et exclure sont moins visibles. Donc, lorsque le répondant complète la liste des résidents habituels du logement, certains ne voient pas ces instructions. Pour la majorité des ménages, cela ne pose pas problème car la définition des résidents habituels n'est pas ambiguë. Par contre, pour ceux qui ont besoin de ces instructions, cela pourrait poser des problèmes. L'élément principal noté avec le feuillet accordéon démontre que certains répondants avaient plus de difficulté à repérer le début du questionnaire. Il était moins facile de voir que les deux côtés comportaient des éléments à compléter. Ainsi, en ayant la liste des résidents habituels sur un côté différent de celui des questions sur les résidents, certains répondants avaient tendance à voir qu'un seul de ces deux éléments. Le risque était donc plus élevé lors de

circonstances plus normales, c'est-à-dire sans la présence d'observateurs, qu'un nombre important de répondants ne complètent qu'un seul des deux éléments : la liste des résidents ou les questions sur les résidents. Pour le format plié, ce problème ne peut se poser car ces deux éléments sont sur un même côté.

Tableau 7: Séquence des éléments du questionnaire tel que présenté aux répondants selon le type de feuillet.

Séquence	Feuillet accordéon	Feuillet plié
1	Côté 1 - Page couverture	Côté 1 - Page couverture
2	Côté 1 - Instructions sur qui inclure et exclure	Côté 2 - Liste des résidents habituels
3	Côté 1 - Liste des résidents habituels	Côté 2 - Questions sur les résidents habituels
4	Côté 2 - Questions sur les résidents habituels	Côté 1 - Instructions sur qui inclure et exclure
5	Côté 2 - Commentaires et renseignements généraux	Côté 1 - Commentaires et renseignements généraux
Principal Inconvénient	Liste des résidents et questions sur les résidents ne sont pas sur le même côté de feuille.	Instructions sur qui inclure et exclure et liste des résidents ne sont pas sur le même côté de feuille.

Suite à ces tests, il a été recommandé d'utiliser le format feuillet plié. Pour inciter les répondants à consulter les instructions sur qui inclure et exclure avant de compléter la liste des résidents habituels, une instruction a été ajoutée juste avant cette liste. L'instruction invite donc le répondant à consulter les instructions au besoin. Suite à cette décision, il a été recommandé d'utiliser les données d'un autre test à plus grande échelle pour étudier l'impact de ce feuillet plié par rapport au format de livret utilisé lors du Recensement 2001.

5. RÉFÉRENCES

- DeMaio, T., et Bates, N. (1992). « Redesigning the Census long form: Results from the 1990 alternative questionnaire experiment ». *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 784-789.
- Dillman, D. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: John Wiley and Sons.
- Dillman, D., Clark, J., et Treat, J. (1994). « Influence of 13 design factors on completion rates to decennial census questionnaires ». *Proceedings of 1994 Annual Research Conference*, U.S. Bureau of the Census, Washington, 137-166.
- Dillman, D. (1998). « Mail and other self-administered surveys in the 21st century: The beginning of a new era ». Ébauche non publiée Washington State.
- Dillman, D., et Christian, L. (2002). « The influence of words, symbols, numbers, and graphics on answers to self-administered questionnaires: results from 18 experimental comparisons ». Ébauche non publiée Washington State.
- Gower, A. (1997). « Présentation des questions sous formes séquentielle, matricielle de feuillet unique et de livret pour le questionnaire du Recensement du Canada ». *Recueil : Symposium 97, Nouvelles orientations pour les enquêtes et les recensements*, Statistique Canada, 251-256.
- Jenkins, C., et Dillman, D. (1993). « Combining cognitive and motivational research perspectives for the design of respondent-friendly self-administered questionnaires ». *Paper for the Annual Meetings of the American Association for Public Opinion Research*.
- Jenkins, C., et Dillman, D. (1997). Towards a theory of self-administered questionnaire design, In Lyberg, et. al. (eds.), *Survey Measurement and Process quality*. New York: John Wiley & Sons, Inc.
- Moore, J., et Moyer, L. (2002). « Questionnaire design effects on interview outcomes ». *Research Report Series, Survey Methodology*, U.S. Bureau of the Census.