

Pondération des données du volet 2013

Karine Dion et Catherine Fontaine
Direction de la méthodologie et de la qualité
Institut de la statistique du Québec
9 juillet 2014

Le présent rapport a pour but de décrire la méthode de pondération utilisée pour les données de l'Étude longitudinale du développement des enfants du Québec (ÉLDEQ) au volet 2013¹. Dans ce rapport, le lecteur est invité à consulter les sections 1 à 3 pour connaître les étapes ayant mené au calcul de la pondération et la section 4 pour savoir comment utiliser la pondération. Plus précisément,

- la section 1 propose une description des étapes ayant mené au choix de la stratégie de pondération ;
- la section 2 présente l'analyse de la non-réponse totale ayant donné lieu à la création de la pondération ;
- la section 3 détaille le taux de réponse pondéré obtenu ; et
- la section 4 renseigne l'utilisateur sur le fichier de pondération ainsi que sur la façon d'utiliser les poids échantillonnaires dans les analyses statistiques. Elle comporte également quelques mises en garde concernant l'utilisation de ces poids.

1. Stratégie de pondération

1.1 Admissibilité à l'enquête au volet 2013

Parmi les 2 120 répondants au volet initial, on compte 31 familles ayant quitté définitivement le Québec et 3 familles dont l'enfant-cible est décédé entre les volets 1998 et 2013. Les enfants-cibles de ces familles, considérés comme inadmissibles à l'enquête, ne sont plus visés par l'enquête en ce sens qu'ils ne font plus partie de la population sur laquelle porte l'inférence. La population visée est par conséquent composée des enfants survivants qui sont demeurés au Québec entre les volets 1998 et 2013 ou qui ont quitté la province temporairement.

Les enfants-cibles des familles n'ayant pu être retracés, ayant refusé de répondre ou ayant été dans l'impossibilité de le faire sont tous considérés comme admissibles à l'enquête. Bien que l'on sache que, parmi les familles n'ayant pu être retracées, certaines pourraient avoir déménagé définitivement hors du Québec, leur nombre est trop petit pour que l'on en tienne compte dans le calcul de la pondération. Sur cette base, l'échantillon admissible à l'enquête au volet 2013 est composé de 2 086 jeunes². Leur répartition, selon la réponse à l'enquête à chacun des volets de 1998 à 2013, est présentée au tableau I.

¹ Les rapports de pondération des volets antérieurs sont disponibles sur le site de l'ÉLDEQ (www.jesuissjeserai.stat.gouv.qc.ca), sous l'onglet « Documentation technique ».

² À partir du volet 2013, les termes « jeune » et « enfant » sont tous deux utilisés dans la documentation technique et dans les bases de données de l'ÉLDEQ pour désigner l'enfant-cible.

Tableau I - Nombre de répondants³ aux volets de 1998 à 2013

| Volets 1998 à 2002 | Volets 2003 à 2011 | Volet 2013 | Nombre de répondants |
|--|----------------------------|------------|----------------------|
| Répondants aux 5 volets | Répondants aux 7 volets | Oui | 991 |
| | | Non | 44 |
| | Répondants à 5 ou 6 volets | Oui | 248 |
| | | Non | 84 |
| | Répondants à 3 ou 4 volets | Oui | 87 |
| | | Non | 100 |
| | Répondants à 1 ou 2 volets | Oui | 70 |
| | | Non | 176 |
| | Répondant à 0 volet | Oui | 13 |
| | | Non | 71 |
| Répondants à 3 ou 4 volets | Répondants aux 7 volets | Oui | 11 |
| | | Non | 0 |
| | Répondants à 5 ou 6 volets | Oui | 23 |
| | | Non | 7 |
| | Répondants à 3 ou 4 volets | Oui | 11 |
| | | Non | 7 |
| | Répondants à 1 ou 2 volets | Oui | 6 |
| | | Non | 16 |
| | Répondant à 0 volet | Oui | 6 |
| | | Non | 25 |
| Répondants à 2 volets | Répondant à 3 ou 4 volets | Oui | 0 |
| | | Non | 1 |
| | Répondant à 1 ou 2 volets | Oui | 0 |
| | | Non | 2 |
| | Répondant à 0 volet | Oui | 0 |
| | | Non | 34 |
| Répondants à 1 volet | Répondant à 0 volet | Oui | 0 |
| | | Non | 53 |
| Nombre total de jeunes admissibles à l'enquête au volet 2013 | | | 2 086 |

Note : Il n'y a pas eu de collecte de données en 2007, 2009 et 2012.

³ Aux volets 1998 à 2002, les répondants ont tous rempli le QIRI; à partir de 2003, les répondants ont rempli au moins un instrument de collecte au volet concerné.

1.2 Répondants au volet 2013

La pondération est un outil qui permet d'inférer à la population visée les estimations produites à partir des données fournies par les répondants. Cette pondération est requise puisque, en plus d'avoir des probabilités de sélection initiales variables, les répondants diffèrent en général des non-répondants. Ainsi, pour une analyse donnée, toute la non-réponse observée devrait idéalement être traitée, c'est-à-dire que la pondération utilisée pour cette analyse devrait avoir fait l'objet d'un ajustement pour compenser toute perte de répondants.

Au fil des volets et considérant la pluralité des instruments d'enquête, les possibilités d'analyse se multiplient. Il est de ce fait impossible de fournir une pondération adéquate pour toutes les situations d'analyse potentielles. Ainsi, pour le volet 2013, une seule pondération principale a été créée. Celle-ci permet l'analyse des variables du volet 2013 portant sur l'ensemble des jeunes ayant répondu à ce volet d'enquête (avec peu de données manquantes pour ces variables ou des variables d'autres volets incluses dans l'analyse).

Tout comme au volet précédent, il a été décidé de créer un poids qui refléterait le fait d'avoir complété au moins un instrument de collecte au volet 2013, au lieu de produire une pondération spécifique au Questionnaire informatisé rempli par l'intervieweur (QIRI), comme ce fut le cas pour les volets 1998 à 2005. Dans ce contexte, le QIRI est considéré au même titre que les autres instruments, c'est-à-dire que lorsque des variables du QIRI sont incluses dans l'analyse, il faut évaluer au préalable l'ampleur de la non-réponse pour laquelle aucun ajustement n'a été fait à la pondération. Soulignons qu'au volet 2013, l'écart entre le nombre de répondants au QIRI et le nombre de répondants à ce volet est plus grand que celui calculé aux deux volets précédents (voir tableau II). Ce qui s'explique principalement par la stratégie de collecte du volet 2013 qui était axée sur le jeune. En effet, contrairement aux volets précédents où les actions de fidélisation ciblaient la famille, plusieurs actions ont été mises en place au volet 2013 pour s'assurer le plus grand nombre possible de répondants au QELJ⁴.

⁴ Questionnaire en ligne au jeune

Tableau II - Nombre de répondants⁵ à certains volets de 1998 à 2013

| | volet 1998 | volet 1999 | volet 2000 | volet 2001 | volet 2002 | volet 2003 | volet 2004 | volet 2005 | volet 2006 | volet 2008 | volet 2010 | volet 2011 | volet 2013 |
|--|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|---------------|
| Nombre de répondants au QIRI pour un volet donné | 2 120 | 2 045 | 1 997 | 1 950 | 1 944 | 1 759 | 1 492 | 1 528 | 1 451 | 1 334 | 1 396 | 1 290 | 1 400 |
| Nombre de répondants au QELJ pour un volet donné | - | - | - | - | - | - | - | - | - | - | - | - | 1446 |
| Nombre de répondants pour un volet donné | 2 120 | 2 045 | 1 997 | 1 950 | 1 944 | 1 775 | 1 529 | 1 537 | 1 528 | 1 402 | 1 415 | 1 312 | 1 466 |
| Nombre de répondants longitudinaux (pour un volet donné et ses précédents) | 2 120 | 2 045 | 1 985 | 1 924 | 1 894 | 1 723 | 1 462 | 1 355 | 1 287 | 1 186 | 1 121 | 1 035 | 991 |

Tout comme au volet 2011, il n'y a pas de pondération longitudinale distincte pour les 991 répondants longitudinaux de 1998 à 2013. Par contre, dans la situation, moins fréquente, où une analyse impliquerait des variables de tous les volets d'enquête, soit de 1998 à 2013, la pondération transversale 2013 ne serait pas appropriée, car ce poids ne tient pas compte de la non-réponse survenue à au moins un volet avant 2013. En effet, le nombre de jeunes participant au volet 2013 qui étaient non participants à au moins un volet précédent, n'est pas négligeable (475 jeunes sur 1 466, soit une proportion d'environ 32 %). De même, la pondération longitudinale de 1998 à 2008 ne serait pas appropriée pour une telle analyse⁶, car elle ne tient pas compte de la non-réponse survenue à au moins un volet de 2010 à 2013. En effet, le nombre de jeunes ayant un tel poids et étant non participants à au moins un volet entre 2010 et 2013 est aussi non négligeable (195 jeunes sur 1 186, soit une proportion d'environ 16 %). Une pondération spécifique devrait donc être créée pour une telle situation d'analyse.

S'il y a lieu, les autres situations d'analyse devraient être évaluées afin de déterminer si la pondération principale est appropriée. Dans le cas contraire, une pondération sur mesure doit être produite. Ce sera le cas lors de l'analyse des variables du questionnaire autoadministré de la mère/conjointe (QAAM) au volet 2013. Le poids transversal calculé pour l'ensemble des jeunes ayant répondu au volet 2013 comporte une grande portion de non-réponse au QAAM qui n'a pas été prise en compte (proportion pondérée d'environ 20 %, voir le tableau III). Mentionnons que le QIRI présente une plus grande proportion pondérée de non-réponse au volet 2013 (4,5%) qu'au volet 2011 (1,6%). Cette hausse de la non-réponse à cet instrument peut s'expliquer en partie par le changement de la stratégie de collecte qui est maintenant axée sur le jeune.

⁵ Aux volets 1998 à 2002, les répondants ont tous rempli le QIRI; à partir de 2003, les répondants ont rempli au moins un instrument de collecte au volet concerné.

⁶ C'est en 2008 que la pondération longitudinale générale la plus récente fut créée.

1.3 Choix du volet de référence pour l'ajustement pour la non-réponse

Le choix de la stratégie de pondération s'appuie sur différents critères. Ceux-ci permettent de choisir le volet 2002 comme année de référence⁷ pour le calcul de la pondération du volet 2013 plutôt que les volets 2011 ou 2010. Le choix de l'année 2002 comme année de référence permet de s'appuyer sur la dernière année de la première phase de l'ÉLDEQ, comme ce fut le cas pour tous les volets de la deuxième phase, soit de 2003 à 2011. En outre, il a été démontré lors de l'analyse des pondérations de 1998 à 2008 que l'utilisation du volet 2002 permettait d'atteindre un meilleur niveau de cohérence longitudinale pour les 4 caractéristiques liées à l'érosion⁸. Enfin, ce choix évite les multiples ajustements de non-réponse entre 2002 et 2013, qui peuvent entraîner une incohérence longitudinale (Ferland, Tremblay et Simard, 2006).

1.3.1 Ajustement de la non-réponse au niveau transversal

Un poids transversal général a ainsi été créé pour les 1 446 répondants au QELJ, de même que pour 20 jeunes additionnels ayant répondu à au moins un autre instrument de collecte⁹ au volet 2013, soit un total de 1 466 jeunes. La méthode de pondération sera décrite plus en détail à la section 2.

La modélisation de la non-réponse au volet 2013 comporte quatre étapes :

1. Ajustement de l'inverse des probabilités de sélection pour la non-réponse à l'enquête au volet 1998 → pondération QIRI du volet 1998.
2. Ajustement des poids QIRI du volet 1998 pour la non-réponse à l'enquête au volet 2000 parmi les répondants du volet 1998 toujours admissible à l'enquête au volet 2013 → pondération QIRI du volet 2000.
3. Ajustement des poids transversaux du volet 2000 pour la non-réponse à l'enquête au volet 2002 parmi les répondants du volet 2000 toujours admissibles à l'enquête au volet 2013 → pondération QIRI du volet 2002.
4. Ajustement des poids transversaux du volet 2002 pour la non-réponse à l'enquête au volet 2013 parmi les répondants du volet 2002 toujours admissibles à l'enquête au volet 2013 → pondération générale transversale du volet 2013.

Afin d'obtenir une pondération transversale pour l'ensemble des 1 466 répondants du volet 2013, les jeunes qui étaient répondants à au moins un volet à partir de l'année 2002 se sont vu attribuer un poids QIRI pour le volet 2002, ce dernier constituant le poids de base de la dernière étape d'ajustement selon la stratégie de pondération décrite précédemment.

La pondération transversale ainsi créée peut être utilisée pour l'analyse des variables qui

⁷ L'année de référence fournit la pondération de base qui fera l'objet d'un ajustement pour la non-réponse survenue ultérieurement.

⁸ Voir l'article de Fontaine et Courtemanche (2009) portant sur l'étude de l'érosion dans l'ÉLDEQ (disponible sur demande).

⁹ Des données sur le jeune provenant du QIRI sont disponibles pour ces 20 jeunes.

prennent une valeur pour l'ensemble des 1 466 répondants du volet 2013. Cette pondération peut également être utilisée pour une analyse de variables où une petite proportion de jeunes présenterait des valeurs manquantes¹⁰.

1.4 Les autres instruments de collecte

Tout comme pour le volet précédent (2011), il n'y a pas de pondération spécifique qui a été créée pour le volet 2013. Il est prévu que des pondérations sur mesure soient produites pour des analyses spécifiques lorsque cela est nécessaire. Ces pondérations sur mesure devront subir un ajustement pour la non-réponse à un instrument et pour la non-réponse partielle à une question, et ce, pour tous les instruments et variables en cause dans l'analyse.

Le tableau III présente le nombre de répondants obtenus pour chacun des instruments de collecte. La proportion pondérée de répondants au QAAM est calculée avec comme dénominateur le nombre estimé de mères ou conjointes présentes dans le ménage en 2013¹¹.

Tableau III - Nombre de répondants par instrument au volet 2013

| | Nombre de répondants | Proportion pondérée de répondants parmi les répondants au volet 2013(%) |
|---|-----------------------------|--|
| Questionnaire en ligne du jeune - QELJ | 1 446 | 98,3 % |
| Questionnaire informatisé rempli par l'intervieweuse - QIRI | 1 400 | 95,5 % |
| Questionnaire autoadministré aux mères/conjointes - QAAM | 1 167 | 80,5 % |

2. Analyse de la non-réponse

2.1 Pondération transversale des données du volet 2013

2.1.1 Démarche générale d'analyse

La création de pondérations ajustées pour la non-réponse est basée sur la formation de classes de pondération. C'est la méthode du score qui a été utilisée pour créer les classes de pondération (pour plus de détails sur la méthode, voir Haziza et Beaumont, 2007 et Eltinge et Yansaneh, 1997). Cette méthode crée des groupes homogènes selon la valeur d'un score, celui-

¹⁰ Règle générale, on considère comme négligeable une proportion de jeunes avec des données manquantes inférieure à environ 5 %. Entre 5 % et 10 %, il est souhaitable de faire une analyse de biais avant d'interpréter les résultats. Au-delà de 10 %, il est recommandé de produire une pondération sur mesure par un ajustement additionnel sommaire afin de tenir compte de la non-réponse différenciée.

¹¹ Le nombre de mères ou conjointes présentes dans le ménage en 2013 doit être estimé puisque cette information provient du QIRI et que pour 66 jeunes, le QIRI n'a pas été rempli en 2013. Le dénominateur utilisé est de 1 426 jeunes.

ci étant issu d'un modèle de régression logistique. C'est la réponse à l'enquête qui a été analysée à l'aide de ce modèle et la probabilité estimée de réponse constitue le score. Par la suite, la création des groupes s'effectue à l'aide d'une méthode de classification. Enfin, pour un jeune donné, l'ajustement de la pondération consiste à diviser le poids de référence par la proportion pondérée de jeunes répondants observée au sein du groupe auquel il appartient. Pour plus de détails concernant cette démarche d'analyse, consulter l'annexe A.

Pour tenir compte de la non-réponse au volet 2013, un ajustement a été fait à partir de la pondération modifiée du volet 2002 (section 2.1.2). Cet ajustement est requis puisque les répondants au volet 2013 présentent des caractéristiques différentes des non-répondants. On minimise ainsi les risques de biais dus à la non-réponse dans les estimations qui seront produites. La nouvelle variable de pondération transversale (PEGENT16) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 466 jeunes ayant répondu à l'enquête au volet 2013.

2.1.2 Conversion de non-répondants au volet 2002

Au total, ce sont 1 977 jeunes répondants au volet 2002 (ou considérés comme répondants) qui forment la base à partir de laquelle l'analyse de la non-réponse au volet 2013 est effectuée. La pondération transversale du volet 2013 vise à attribuer un poids aux 1 466 répondants de ce volet, parmi ces 1 977 jeunes, à partir du poids QIRI du volet 2002. Les 1 977 jeunes se répartissent de la manière suivante :

- 1 934 répondants du volet 2002, toujours admissibles au volet 2013, sont associés à un poids QIRI du volet 2002.
- 27 jeunes répondants au volet 2013 qui n'étaient pas répondants au volet 2002 et n'ont, de ce fait, aucun poids de référence du volet 2002. Aux fins de la pondération transversale du volet 2013, ces jeunes ont été considérés comme répondants au volet 2002. Un nouveau poids est calculé pour l'ensemble des répondants au volet 2002, incluant ces 27 jeunes. On les nomme pseudo-répondants au volet 2002.
- 16 jeunes répondants qui n'ont répondu ni au volet 2002 ni au volet 2013, mais qui ont répondu à au moins un volet de 2003 à 2011¹². Ils ont également été considérés comme répondants au volet 2002, de manière à obtenir un poids transversal au volet 2002 pour ces jeunes en vue d'une utilisation potentielle dans le calcul des pondérations des volets ultérieurs. Cette décision est justifiée par le fait que ces jeunes n'ont pas cessé de répondre à l'enquête au volet 2002. On les nomme aussi pseudo-répondants au volet 2002.

Pour effectuer le calcul des poids du volet 2002 pour les répondants (et pseudo-répondants), les classes de pondération définies au volet 2002 ont été conservées; seules les proportions pondérées de répondants ont été recalculées. Pour les variables servant à créer les classes de pondération, des valeurs ont été imputées pour les non-répondants du volet 2000, aux seules fins de la pondération.

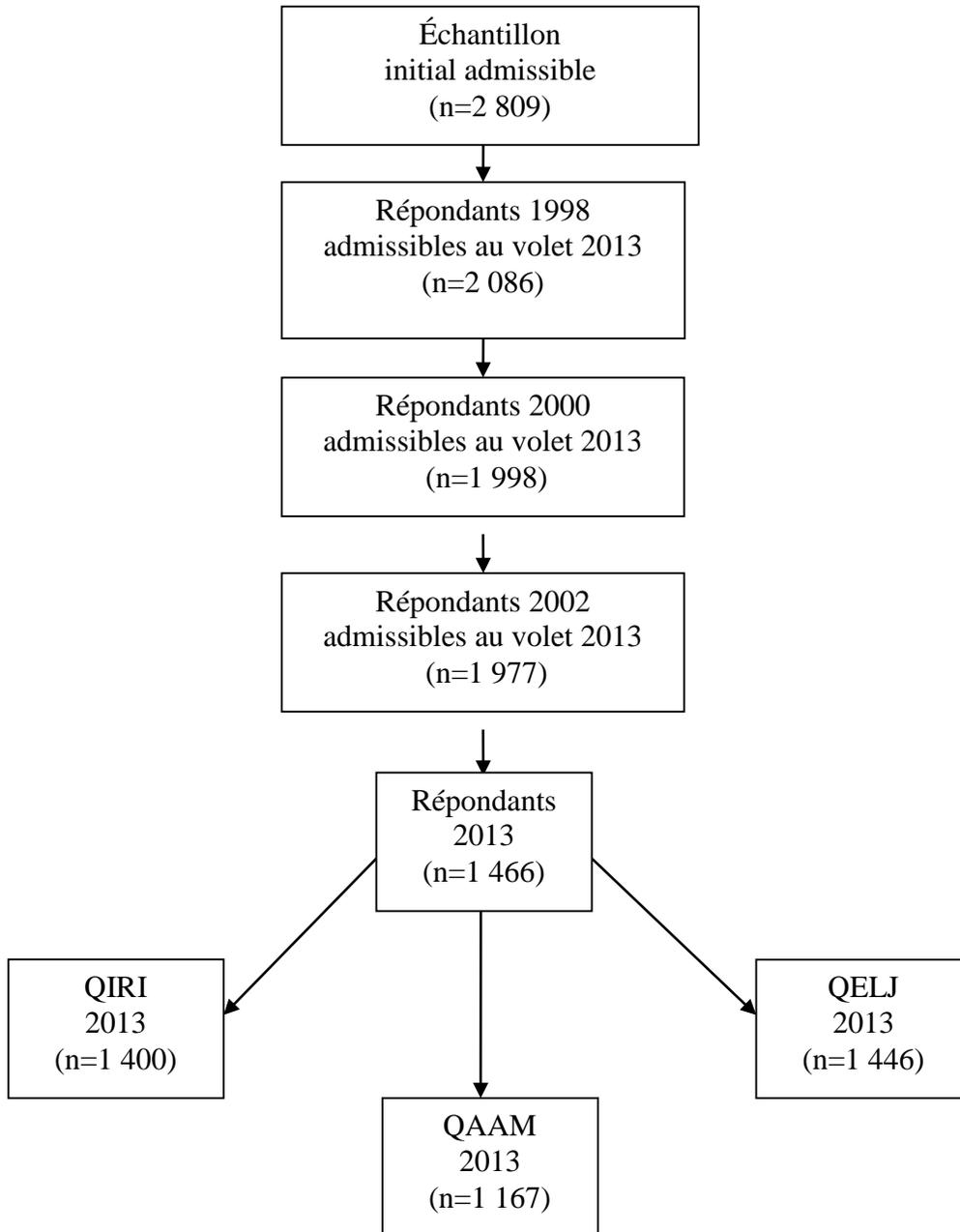
¹² Ils ont rempli le QIRI seulement pour les volets 2003 et 2004.

Il est à noter que parmi ces 1 977 jeunes visés par l'analyse de la non-réponse, il y en a 168 qui ne faisaient pas partie de ceux qui ont été sollicités pour la collecte de données au volet 2013. En effet, à la fin de chaque collecte, un examen des unités enquêtées pour ce volet est effectué. Les enfants-cibles des familles qui n'ont pas répondu depuis un certain temps à l'enquête ou qui ont signifié un refus définitif, par exemple, ne sont plus sollicités au volet suivant. Ainsi, ces 168 jeunes n'avaient aucune possibilité de participer au volet 2013, contrairement aux autres jeunes non-répondants au volet 2013. Ces différents types de non-répondants ont tout de même été modélisés ensemble.

La hiérarchie des répondants admissibles au volet 2013 est illustrée au schéma I.

Schéma I.

Hiérarchie des répondants aux différents instruments de collecte, admissibles au volet 2013



Note : Aux volets 2000 et 2002, le nombre indiqué inclut les pseudo-répondants (voir section 2.1.2).

2.1.3 Variables considérées et résultats

Les variables considérées pour la modélisation sont principalement de nature socioéconomique. Elles portent sur la mère de l'enfant-cible ou sur sa famille et sont tirées du QIRI du volet 2002. Des variables dites longitudinales ont également été étudiées en créant un indice à partir de la même mesure prise de 1998 à 2002. Ces variables sont :

- le revenu du ménage (revenu faible à au moins un des 5 volets (soit moins de 10 000 \$), versus autres; revenu faible à au moins un des 5 volets (soit moins de 15 000 \$), versus autres);
- le type de famille (monoparentalité à au moins un volet versus autres; monoparentalité ou nouveau conjoint à au moins un volet versus autres);
- la présence du père biologique (le père biologique est absent du ménage à au moins un volet versus autres);
- le niveau de suffisance du revenu du ménage (insuffisance du revenu à au moins un volet versus autres);
- le travail de la mère au cours des douze derniers mois (n'a pas travaillé au cours des douze mois précédant l'enquête pour plus d'un volet versus autres);
- la principale source de revenu du ménage (aide sociale comme principale source de revenus à aucun volet, à 1 ou 2 volets, à 3 volets ou plus); et
- la situation en emploi des parents (aucun parent en emploi à aucun volet, à 1 ou 2 volets, à 3 volets ou plus).

Compte tenu de la disponibilité de microdonnées administratives¹³ provenant du ministère de l'Éducation, du Loisir et du Sport du Québec (MELS), des variables supplémentaires ont été considérées pour la pondération du volet 2013. Ces variables sont reliées au cheminement scolaire du jeune : existence d'un plan d'intervention actif, sexe, niveau scolaire, langue d'enseignement, langue maternelle, type de classe spéciale (s'il y a lieu) et code de difficulté. Ce fichier permet aussi de créer une variable qui exprime la mobilité du jeune entre 2011 et 2013. Cette variable a été créée en comparant les adresses de la résidence du jeune pour les années 2011 et 2013.

Parmi l'ensemble des variables considérées, voici celles qui ont été retenues pour le modèle final de régression logistique :

- le plus haut niveau de scolarité de la mère/conjointe en 2002 (EEDMD01)
- le nombre de frères/sœurs de l'enfant-cible en 2002 (EREED01)
- le sexe de l'enfant (sexe)
- la langue d'enseignement en 2013 (lan_enseignement)
- l'existence d'un plan d'intervention actif en 2013 (plan)

Une méthode de classification non hiérarchique a permis de regrouper les probabilités estimées en 4 classes de pondération. Le tableau IV présente les proportions pondérées de répondants au volet 2013 parmi les répondants au volet 2002 pour ces 4 groupes. De plus, il présente le nombre de répondants, parmi les 1 466 répondants, à qui la proportion pondérée sera appliquée en guise de correction de la non-réponse. Par exemple : il y a 55 répondants au

¹³ Microdonnées provenant du fichier daté de février 2014 pour les enfants inscrits à l'école au 30 septembre 2012.

volet 2013 dont le poids de référence sera ajusté par l'inverse de la proportion pondérée de la troisième classe de pondération, qui est de 53,6 %.

Tableau IV - Proportions pondérées de répondants et nombre de répondants par classe de pondération (transversal)

| Classe de pondération | Proportions pondérées de répondants au volet 2013 (en %) | Nombre de répondants |
|-----------------------|--|----------------------|
| 1 | 32,4 | 23 |
| 2 | 53,6 | 55 |
| 3 | 71,0 | 738 |
| 4 | 83,1 | 650 |

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondants varie de 32 % à 83 % (relativement à une proportion globale de 72,9 %). La proportion la plus faible est observée dans une classe où on retrouve, en proportion, un plus grand nombre de jeunes de sexe masculin; un plus grand nombre de jeunes dont la langue d'enseignement est le français, un plus grand nombre de jeunes ayant au moins un frère ou une sœur ; et un moins grand nombre d'enfants dont la mère avait un diplôme universitaire.

2.1.4 Ajustement de la pondération à l'aide de données administratives

Suite à une entente avec le ministère de l'Éducation, du Loisir et du Sport du Québec (MELS), l'Institut de la statistique du Québec a pu obtenir de cet organisme des statistiques agrégées pour la population visée par l'ÉLDEQ¹⁴ (N=69 890). Ces statistiques agrégées sont en fait des totaux pour des caractéristiques choisies par l'Institut et disponibles au MELS. Par exemple : la répartition des 69 890 jeunes selon la région de résidence. Ces statistiques agrégées obtenues pour les données du volet 2013 permettent d'évaluer la pertinence d'effectuer un ajustement à la pondération transversale. Cet ajustement, appelé « calage », est défini comme un redressement des poids d'enquête afin que les estimations s'ajustent à des totaux connus (Lavallée et Durning, 1993). Ce redressement peut aussi être utilisé dans le but de pallier la non-réponse.

L'objectif du calage au volet 2013 est d'effectuer une correction supplémentaire pour diminuer le biais dû à la non-réponse, et ce, à l'aide de caractéristiques reliées aux mesures principales de l'enquête, c'est-à-dire des variables liées à la réussite scolaire. C'est dans cet esprit que les caractéristiques ont été choisies à partir de l'ensemble des variables administratives disponibles¹⁵. En effet, l'Institut reçoit annuellement, en plus des statistiques agrégées, un fichier de microdonnées administratives pour l'essentiel de l'échantillon de départ de la cohorte¹⁶. Les variables provenant des données administratives du MELS sont donc disponibles :

¹⁴ L'ensemble des enfants nés au Québec entre le 1er octobre 1997 et le 30 septembre 1998 qui fréquentent le système scolaire québécois au cours de l'année scolaire 2012-2013.

¹⁵ Pour limiter l'ampleur de la production de statistiques agrégées au MELS, un choix devait être fait.

¹⁶ Fichier en date de février 2014 pour les jeunes inscrit à l'école au 30 septembre 2012.

- Au niveau des microdonnées, pour l'ensemble de l'échantillon de l'ÉLDEQ¹⁷ ;
- Au niveau des macro-données (totaux) pour l'ensemble de la population visée par l'ÉLDEQ.

Les variables administratives considérées pour le calage sont :

- Sexe du jeune;
- Langue maternelle du jeune;
- Existence d'un plan d'intervention actif pour le jeune à l'école;
- Niveau scolaire du jeune ;
- Code de difficulté du jeune.

Une comparaison a d'abord été effectuée entre la distribution pondérée de l'ensemble des répondants au volet 2013 et la distribution de la population visée (pour ces cinq variables en excluant les valeurs manquantes). L'objectif était de vérifier si les proportions pondérées étaient près des proportions calculées pour la population. En effet, si l'écart est négligeable, cela signifie que l'ajustement de calage n'est pas nécessaire puisque le biais est faible.

C'est la variable qui exprime le retard scolaire qui présente un écart le plus important entre les deux distributions. Cette variable identifie les jeunes inscrits à l'école, en 2013, au niveau primaire ou au premier cycle du secondaire, alors que les jeunes qui ont suivi un parcours scolaire dit normal sont en première année du deuxième cycle du secondaire. La proportion pondérée de jeunes en situation de retard scolaire est de 16,7 % dans l'échantillon (avant imputation) comparativement à 21,7 % dans la population de l'ÉLDEQ. La variable sexe du jeune présentait également un écart entre les deux distributions. La proportion pondérée après imputation est de 47,5 % garçons dans l'échantillon comparativement à 51,1 % garçons dans la population de l'ÉLDEQ. Ainsi, l'ajustement des poids s'est effectué selon ces deux variables afin de rendre la distribution pondérée des répondants selon les variables du retard scolaire et du sexe, semblable à la distribution dans la population visée. Quelques variables du volet 2013 ont été choisies pour vérifier l'impact de cet ajustement apporté aux poids. Il est possible de conclure que les proportions pondérées calculées pour des caractéristiques généralement associées à une moins grande réussite scolaire ont légèrement augmenté suite à l'ajustement apporté aux poids¹⁸. Ce constat va dans le sens attendu, à savoir que l'échantillon de l'ÉLDEQ se rapproche de la population visée, cette dernière comprenant davantage d'élèves ayant des caractéristiques liées à une moins grande réussite scolaire. Ce sont les variables du déficit de l'attention (avec ou sans hyperactivité) et celle identifiant les élèves en difficulté selon le fichier du MELS qui sont le plus touchées par l'ajustement de calage.

¹⁷ De nouvelles microdonnées pour l'ensemble de l'échantillon sont disponibles, dont les résultats scolaires aux épreuves, diplomation, etc. Pour le prochain volet, il serait intéressant d'obtenir les macro-données de certaines de ces nouvelles variables.

¹⁸ Par exemple, dans le QIRI pour la variable qui identifie les jeunes avec un problème chronique de déficit d'attention (avec ou sans hyperactivité).

3. Taux de réponse

Le tableau V présente le taux de réponse pondéré transversal obtenu au volet 2013. Ce taux est obtenu en multipliant les taux obtenus aux différentes étapes de pondération, selon le cas. Mentionnons que nous avons obtenu un meilleur taux de réponse pondéré transversal au volet 2013, soit à 51,6% comparativement à 46% au volet 2011.

Tableau V - Taux de réponse pondéré transversal au volet 2013

| | |
|--|----------------------|
| Taux de réponse au volet 1998 | 75,3 % (n=2 809) |
| Proportion de répondants (incluant les pseudo-répondants) au volet 2000 parmi les répondants au volet 1998 admissibles au volet 2013 | 95,0 % (n=2 086) |
| Proportion de répondants au volet 2002 parmi les répondants au volet 2000 admissibles au volet 2013 (incluant les pseudo-répondants) | 98,95 % (n=1 998) |
| Proportion de répondants au volet 2013 parmi les répondants au volet 2002 admissibles au volet 2013 (incluant les pseudo-répondants) | 72,9 % (n=1 977) |
| Taux de réponse transversal au volet 2013 | 51,6 % |

Note : le chiffre présenté entre parenthèses représente le dénominateur à partir duquel le calcul est effectué.

4. Utilisation de la pondération par les utilisateurs des données du volet 2013

4.1 L'importance de la pondération

Les utilisateurs des données du volet 2013 sont fortement encouragés à utiliser la pondération lors des analyses des données de l'ÉLDEQ. La pondération est le résultat du traitement de la non-réponse. Elle permet d'inférer les résultats à la population visée tout en minimisant les biais dans les estimations.

La non-réponse peut survenir à différents niveaux : au niveau du volet d'enquête, au niveau de l'instrument de collecte et au niveau des variables présentes dans les analyses. Ce document discute du traitement pour la non-réponse survenue au volet d'enquête 2013 au niveau transversal. Un second document traite de la non-réponse partielle à une question¹⁹. Au niveau transversal, le taux de réponse au volet 2013 est de l'ordre de 51,6 % (voir tableau V). Ce faible taux confirme l'importance du traitement effectué lors de la pondération.

La stratégie de pondération mise en œuvre pour créer la pondération principale du volet 2013 utilise des méthodes statistiques complexes afin de créer des sous-groupes d'enfants à partir de certaines caractéristiques. Ces caractéristiques sont définies à partir de variables

¹⁹ Voir le document « Étude de la non-réponse partielle au volet 2013 » par Dion et Fontaine (2014).

disponibles à des volets antérieurs pour chacun des enfants. Des variables administratives provenant du MELS ont aussi été considérées au volet 2013 lors du traitement de la non-réponse. Par la suite, la correction tenant compte de la non-réponse est appliquée à l'intérieur de ces sous-groupes.

4.2 Tests statistiques

Le fichier POIDS1601 contient la variable de pondération PEGENT16 (poids général transversal du volet 2013). C'est un poids échantillonnal, c'est-à-dire un poids qui a été multiplié par une constante de sorte que la somme des poids soit égale à la taille de l'échantillon. Ce poids doit faire partie de toute analyse des données du volet 2013, comme indiqué à la section 4.1. Des logiciels statistiques, tels que SUDAAN, SAS ou STATA, permettent l'intégration de la pondération dans les différentes procédures offertes. En plus d'intégrer la pondération afin de minimiser les biais dans les estimations, le plan de sondage peut aussi être pris en compte lors des analyses. Le logiciel SUDAAN le permet, ainsi que certaines procédures du logiciel SAS. Ainsi c'est la variance qui est correctement estimée (pour les estimations et les tests statistiques).

Si les logiciels utilisés ne tiennent pas compte du plan de sondage complexe, le poids PEGENT16 peut être utilisé pour faire des tests approximatifs.

Afin de pallier le caractère approximatif des tests statistiques réalisés à l'aide de poids échantillonnals, il est recommandé d'adopter une approche conservatrice en abaissant le seuil théorique des tests. Par exemple, si l'on souhaite faire des tests au seuil théorique de 0,05, on peut choisir de n'interpréter que les résultats significatifs au seuil 0,01. Par exemple, il serait possible de conclure, avec un seuil observé de 0,005 obtenu d'un test statistique, que l'hypothèse nulle du test est rejetée au seuil théorique de 0,05 (étant donné que 0,005 est inférieur à 0,01).

Dans le cas particulier de tests du khi-deux sur un tableau de fréquences, l'utilisation des poids échantillonnals divisés par un effet de plan moyen égal à 1,3²⁰ demeure appropriée pour obtenir un test approximatif. Il n'est alors pas nécessaire d'abaisser le seuil des tests. Un résultat pour lequel le seuil observé est près de 0,05 devrait néanmoins être interprété avec nuances.

L'utilisation de poids échantillonnals comporte toutefois certaines limites. En fait, les poids ramenés à la taille de l'échantillon permettent d'obtenir des proportions estimées non biaisées par rapport au plan de sondage ainsi qu'une taille d'échantillon global égale à la taille réelle. Ces poids ne préservent toutefois pas la taille d'échantillon de chacune des catégories d'une variable, c'est-à-dire des sous-groupes au sein de la population. En présence de poids peu variables, la somme des poids échantillonnals pour un sous-groupe est approximativement égale à la taille de celui-ci; l'utilisation de ces poids permet de faire des tests approximatifs valides. Dans le cas contraire, la somme des poids échantillonnals peut différer de façon importante de la taille d'échantillon pour un sous-groupe. Cela a pour conséquence d'invalider

²⁰ Il est important de mentionner que cet effet de plan moyen a été estimé au volet 1998 et réévalué aux volets 2000 et 2008. L'effet de plan est demeuré de l'ordre de 1,3.

les tests statistiques, à moins qu'ils ne soient réalisés à l'aide d'un logiciel qui permet de tenir compte de l'effet du plan de sondage dans l'estimation des paramètres ainsi que de leur variance. Ainsi, il se pourrait que l'on déclare significatifs des écarts entre les sous-groupes qui ne sont pas réels, ou l'inverse selon le cas.

Dans ce contexte, il faudrait plutôt faire une analyse pour chacun des sous-groupes séparément en réajustant les poids de telle sorte que la somme des poids pour chaque sous-groupe soit égale à la taille d'échantillon. Il suffit pour ce faire de diviser les poids par la moyenne des poids pour un sous-groupe. Cette recommandation vaut pour toute analyse portant sur un sous-groupe. Il est important dans ces cas de s'assurer que la somme des poids est approximativement égale à la taille d'échantillon de ce sous-groupe; autrement, un ajustement des poids est requis.

4.3 Choix de la pondération

Les possibilités d'analyse incluant des données du volet 2013 sont innombrables. Ainsi, en raison de la non-réponse qui varie selon les instruments de collecte et les volets considérés, le choix d'une pondération adéquate nécessite un examen cas par cas. **En précisant la population visée, de même que les instruments et les volets considérés pour l'analyse, l'Institut peut évaluer si une pondération appropriée est disponible. Dans le cas contraire, une pondération sur mesure peut être requise.** Il s'agirait alors pour l'Institut de faire un ajustement de la pondération existante, de manière à minimiser les biais potentiels qui pourraient être induits par une non-réponse non prise en compte.

En sus des problèmes dus à la non-réponse au volet et/ou à un instrument de collecte, la perte d'unités d'analyse due à la non-réponse partielle provenant de chacune des variables considérées pour la modélisation doit être étudiée. Si cette non-réponse est importante, les estimations pourraient être entachées d'un biais additionnel; l'interprétation des résultats devrait par conséquent en tenir compte, s'il y a lieu.

En résumé, le choix d'une pondération appropriée doit tenir compte tant de la perte d'unités d'analyse due à l'absence de poids pour ces unités que de la qualité de l'ajustement pour la non-réponse. En effet, au moyen d'un ajustement adéquat, une pondération devrait généralement tenir compte de la non-réponse observée pour l'échantillon d'analyse. Le lecteur est invité à consulter des exemples qui illustrent la démarche à suivre pour évaluer la situation. Ceux-ci se retrouvent dans les rapports de pondération des volets antérieurs.

5. Références bibliographiques

Eltinge, J. L. et Yansaneh, I.S. (1997). Diagnostics for formation of nonresponse adjustment cells, with an application to income nonresponse in the U.S. Consumer Expenditure Survey, *Techniques d'enquête*, vol. 23, no. 1, pages 33-40.

Ferland, M., Tremblay, M. et Simard, M. (2006). Dealing with nonresponse in longitudinal social surveys. Soumis au Journal of Official Statistics pour un numéro spécial portant sur la conférence des méthodes d'enquêtes longitudinales (MOLS), Essex, Angleterre, 2006.

Fontaine, C. et Courtemanche, R. (2009). Analyse de l'érosion de l'Étude Longitudinale sur le Développement des Enfants du Québec (ÉLDEQ) de 1998 à 2008, actes du 25^{ème} Symposium international sur les questions de méthodologie de Statistique Canada, Ottawa, octobre 2009 (à paraître).

Dion, K. et Fontaine, C. (2014). Étude de la non-réponse partielle au volet 2013, document interne, Institut de la statistique du Québec.

Haziza, D. et Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, **75**, 25-43

Lavallée, P. et Durning. A. (1993). Estimateur jackknife de la variance pour l'estimation par calage sur marges, article extrait de la présentation faite dans le cadre du congrès de l'Association canadienne français pour l'avancement des sciences (ACFAS) en 1993.

ANNEXE A

Les étapes de la création d'une pondération générale

Voici la description de la séquence des étapes de création de la pondération transversale pour les participants au volet 2013.

Étape 1 :

Analyses bivariées pour réduire le nombre de variables considérées pour la modélisation (environ 50 variables). Les variables ayant les seuils observés les plus faibles sont conservées.

Étape 2 :

Modélisation préliminaire avec la régression logistique afin d'identifier les variables retenues à l'étape 1 qui présentent un problème de multicollinéarité. Plusieurs essais de modélisation ont été effectués afin de ne retenir qu'un sous-ensemble de variables. Celles-ci ne présentent pas de problème de multicollinéarité entre elles, ni de taux de non-réponse partielle élevée, ni de seuils observés très élevés.

Étape 3 :

Estimation de la taille du modèle par la minimisation du critère d'Akaike (à titre indicatif).

Étape 4 :

Détermination d'un modèle de régression logistique avec SUDAAN pour prédire la probabilité de réponse, en excluant les enfants pour lesquels il y a présence de non-réponse partielle combinée

Étape 5 :

Création d'une catégorie de valeurs manquantes pour les variables du modèle retenu à l'étape 4. La validation de ce modèle est effectuée et un modèle final est retenu.

Étape 6 :

Création des classes de pondération effectuée à l'aide de la méthode du score, ce dernier étant la probabilité de réponse estimée à l'aide du modèle. La détermination du nombre de classes et le regroupement sont effectués à l'aide d'une méthode de classification non hiérarchique. Ceci étant fait, les poids de base sont ajustés selon la proportion pondérée de répondants par classe.

Étape 7 :

Ajustement des poids afin que la distribution pondérée des répondants s'ajuste à celle de la population de l'ÉLDEQ, selon une variable déterminée. Ainsi, la pondération 2013 est constituée.