Pondération des données du volet 2011

Catherine Fontaine et Robert Courtemanche
Direction de la méthodologie et de la qualité
Institut de la statistique du Québec
26 juillet 2012

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 2011. Il s'inspire en grande partie de rapports rédigés à des volets précédents[1]. Pour savoir comment utiliser la pondération, le lecteur est invité à consulter la section 4 de ce document. Cette section renseigne l'utilisateur sur le fichier de pondération ainsi que sur la façon d'utiliser les poids échantillonnaux dans les analyses statistiques. Elle comporte également quelques mises en garde concernant l'utilisation de ces poids. Par ailleurs, le lecteur est invité à consulter les sections 1 à 3 pour connaître les étapes ayant menées au calcul de la pondération du volet 2011. La section 1 propose une description des étapes ayant mené au choix de la stratégie de pondération. La section 2, pour sa part, présente l'analyse de la non-réponse totale ayant donné lieu à la création de la pondération. Le taux de réponse pondéré obtenu est, quant à lui, détaillé à la section 3.

1. Stratégie de pondération

1.1 Admissibilité à l'enquête au volet 2011

Parmi les 2 120 répondants au volet initial, on compte 28 familles ayant quitté définitivement le Québec et 3 familles dont l'enfant cible est décédé entre les volets 1998 et 2011. Les enfants cible 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 2011 ou qui ont quitté la province temporairement.

Les enfants cible des familles n'ayant pu être retracées, 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 il pourrait y en avoir qui sont déménagées 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 2011 est composé de 2 089 enfants. Leur répartition, selon la réponse à l'enquête à chacun des volets de 1998 à 2011, est présentée au tableau I.

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

Tableau I - Nombre de répondants[2] aux volets de 1998 à 2011

Volets 1998 à 2002	Volets 2003 à 2006, 2008 et 2010	Volet 2011	Nombre de répondants
Répondants aux 5 volets	Répondants aux 6 volets	Oui	1 035
		Non	86
	Répondants à 4 ou 5 volets	Oui	171
		Non	151
	Répondants à 2 ou 3 volets	Oui	50
		Non	157
	Répondants à 1 volet	Oui	13
		Non	136
	Répondant à 0 volet	Oui	3
		Non	84
Répondants à 3 ou 4 volets	Répondants aux 6 volets	Oui	11
		Non	4
	Répondants à 4 ou 5 volets	Oui	21
		Non	10
	Répondants à 2 ou 3 volets	Oui	4
		Non	15
	Répondants à 1 volet	Oui	3
		Non	12
	Répondant à 0 volet	Oui	1
		Non	32
Répondants à 2 volets	Répondant à 4 ou 5 volets	Oui	0
		Non	1
	Répondant à 1 volet	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 d'enfants admissibles à l'enquête au volet 2011			2 089

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

2 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 2011

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 2011, une seule pondération principale a été créée. Celle-ci permet l'analyse des variables du volet 2011 portant sur l'ensemble des enfants 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).

Notons qu'au volet 2011 il a été décidé de créer un poids qui refléterait le fait d'avoir complété au moins un instrument de collecte, 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 2011 l'écart entre le nombre de répondants au QIRI et le nombre de répondants à ce volet est du même ordre de grandeur que celui calculé au volet 2010 (voir tableau II).

Tableau II- Nombre de répondants[3] aux volets de 1998 à 2011

	volet 1998	volet 1999	volet 2000	volet 2001	volet 2002	volet 2003	volet 2004	volet 2005	volet 2006	volet 2008	volet 2010	volet 2011
Nombre de répondants au QIRI pour un volet donné	2 120	2 045	1997	1 950	1 944	1 759	1 492	1528	1 451	1 334	1 396	1 290
Nombre de répondants pour un volet donné	2 120	2 045	1997	1 950	1 944	1 775	1 529	1537	1 528	1 402	1 415	1 312
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	1355	1 287	1 186	1 121	1 035

Tout comme au volet 2010, il n'y a pas de pondération longitudinale distincte pour les 1 035 répondants longitudinaux de 1998 à 2011. Par contre, dans la situation, moins fréquente, où une analyse impliquerait des variables de tous les volets d'enquête, soit de 1998 à 2011, la pondération transversale 2011 ne serait pas appropriée. En effet, le nombre d'enfants participant au volet 2011 qui étaient non participants à au moins un volet précédent n'est pas négligeable (277 enfants sur 1 312, soit une proportion d'environ 21 %). De même, la pondération longitudinale de 1998 à 2008 ne serait pas appropriée pour une telle analyse[4]. En effet, le nombre d'enfants ayant un tel poids et qui étaient non participants au volet 2011 est aussi non négligeable (132 enfants sur 1 186, soit une proportion d'environ 11 %). Donc une pondération spécifique serait à créer pour cette situation d'analyse.

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.

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

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 probablement le cas lors de l'analyse des variables du Questionnaire auto-administré de la mère/conjointe (QAAM) au volet 2011. Le poids transversal calculé pour l'ensemble des enfants ayant répondu au volet 2011 comporte une grande portion de non-réponse au QAAM qui n'a pas été prise en compte (proportion pondérée d'environ 18%, voir le tableau III).

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 référence[5] pour le calcul de la pondération du volet 2011 plutôt que les volets 2010 ou 2008. 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 2010. 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'érosion[6]. Enfin, ce choix évite les multiples ajustements de non-réponse entre 2002 et 2011, 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 1 290 répondants au QIRI, de même que pour 22 enfants additionnels ayant répondu à au moins un autre instrument de collecte[7] au volet 2011, soit un total de 1 312 enfants. La méthode de pondération sera décrite plus en détail à la section 2.

La modélisation de la non-réponse au volet 2011 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 → 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 2011 → pondération QIRI du volet 2002.

4. Ajustement des poids transversaux du volet 2002 pour la non-réponse à l'enquête au volet 2011 parmi les répondants du volet 2002 toujours admissibles à l'enquête au volet 2011 → pondération générale transversale du volet 2011.

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

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

7 Des données sur l'enfant (provenant du Questionnaire informatisé à l'enfant, du QAAENS ou du QAAM) sont disponibles pour ces 22 enfants.

Afin d'obtenir une pondération transversale pour l'ensemble des 1 312 répondants du volet 2011, les enfants 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[8], 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 prennent une valeur pour l'ensemble des 1 312 enfants ayant répondu à l'enquête au volet 2011. Cette pondération peut également être utilisée pour une analyse de variables où une petite proportion d'enfants présenterait des valeurs manquantes[9].

1.4 Les autres instruments de collecte

Pour le volet 2011, il n'y a pas de pondération spécifique qui a été créée. Il est prévu que des pondérations sur mesure soient produites pour des analyses spécifiques lorsque 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.

Tableau III - Nombre de répondants par instrument au volet 2011

	Nombre de répondants	Proportion pondérée de répondants parmi les répondants au volet 2011(%)
Questionnaire informatisé rempli par l'intervieweuse - QIRI	1 290	98,4 %
Questionnaire auto-administré aux mères/conjointes - QAAM	1 080	82,3 %
Questionnaire auto-administré aux pères/conjoints - QAAP	741	65,2%
Questionnaire auto-administré aux pères biologiques absents - QPABS	105	-
Questionnaire informatisé à l'enfant - QIE	1 234	91,1%
Questionnaire auto-administré à l'enseignant - QAAENS	1 056[10]	79,9 %
Questionnaire complété par l'intervieweuse - QCI	1 229	90,9 %

8 La méthode utilisée pour attribuer un poids QIRI à ces enfants aux volets antérieurs sera décrite à la section 2.

9 Règle générale, on considère négligeable une proportion d'enfants avec 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.

10 On retrouve 1 056 enfants pour lesquels le QAAENS a été rempli par un ou deux enseignant(s). Ce nombre ne représente pas le nombre de questionnaires QAAENS complétés par les enseignants

La proportion pondérée de répondants au QAAM est calculée avec comme dénominateur le nombre estimé de mères ou conjointes présentes dans le ménage en 2011[11]. La proportion pondérée de répondants au QAAP est calculée avec comme dénominateur le nombre estimé de pères ou conjoints présents dans le ménage en 2011[12]. Quant au QPABS, la proportion n'a pas été calculée puisque le nombre de pères biologiques absents du ménage est indéterminé pour 2011. Par contre, il est possible de calculer cette proportion pondérée parmi les répondants au QIRI 2011. Celle-ci est de 23% (103 questionnaires complétés parmi les 415 pères biologiques absents selon le QIRI). La proportion pondérée de répondants au QAAENS est calculée avec comme dénominateur le nombre d'enfants admissibles à compléter ce questionnaire, c'est-à-dire les enseignants d'enfants qui fréquentent une école québécoise au moment de la collecte (n=1 286). En effet, le QAAENS n'a pas été envoyé par la poste ou par web aux enseignants des enfants qui reçoivent leur éducation à la maison ou qui habitent à l'extérieur du Québec de façon temporaire.

2. Analyse de la non-réponse

2.1 Pondération transversale des données du volet 2011

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-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 enfant donné, l'ajustement de la pondération consiste à diviser le poids de référence par la proportion pondérée d'enfants 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 2011, 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 2011 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 (PEGENT14) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 312 enfants ayant répondu à l'enquête au volet 2011.

11 Le nombre de mères ou conjointes présentes dans le ménage en 2011 doit être estimé puisque cette information provient du QIRI et que pour 22 enfants, le QIRI n'a pas été rempli en 2011. Le dénominateur utilisé est de 1 285 enfants.

12 Le nombre de pères ou conjoints présents dans le ménage en 2011 doit être estimé puisque cette information provient du QIRI et que pour 22 enfants, le QIRI n'a pas été rempli en 2011. Le dénominateur utilisé est de 1 079 enfants.

2.1.2 Conversion de non-répondants au volet 2002

Au total, ce sont 1 977 enfants répondants au volet 2002 (ou considérés comme répondants) qui forment la base à partir de laquelle l'analyse de la non-réponse au volet 2011 est effectuée. La pondération transversale du volet 2011 vise à attribuer un poids aux 1 312 répondants de ce volet, parmi ces 1 977 enfants, à partir du poids QIRI du volet 2002. Les 1 977 enfants se répartissent de la manière suivante :

- 1 936 répondants du volet 2002, toujours admissibles au volet 2011, sont associés à un poids QIRI du volet 2002.
- 19 enfants répondants au volet 2011 n'étaient, cependant, 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 2011, ces enfants ont été considérés répondants au volet 2002. Un nouveau poids est calculé pour l'ensemble des répondants au volet 2002, incluant ces 19 enfants. On les nomme pseudo-répondants au volet 2002.
- 22 enfants supplémentaires qui n'ont répondu ni au volet 2002, ni au volet 2011, mais qui ont répondu à au moins un volet de 2003 à 2010[13] ont également été considérés répondants au volet 2002, de manière à obtenir un poids transversal au volet 2002 pour ces enfants en vue d'une utilisation potentielle dans le calcul des pondérations des volets ultérieurs. Cette décision est justifiée par le fait que ces enfants n'ont pas cessé de répondre à l'enquête au volet 2002. On les nomme aussi pseudo-répondants au volet 2002.

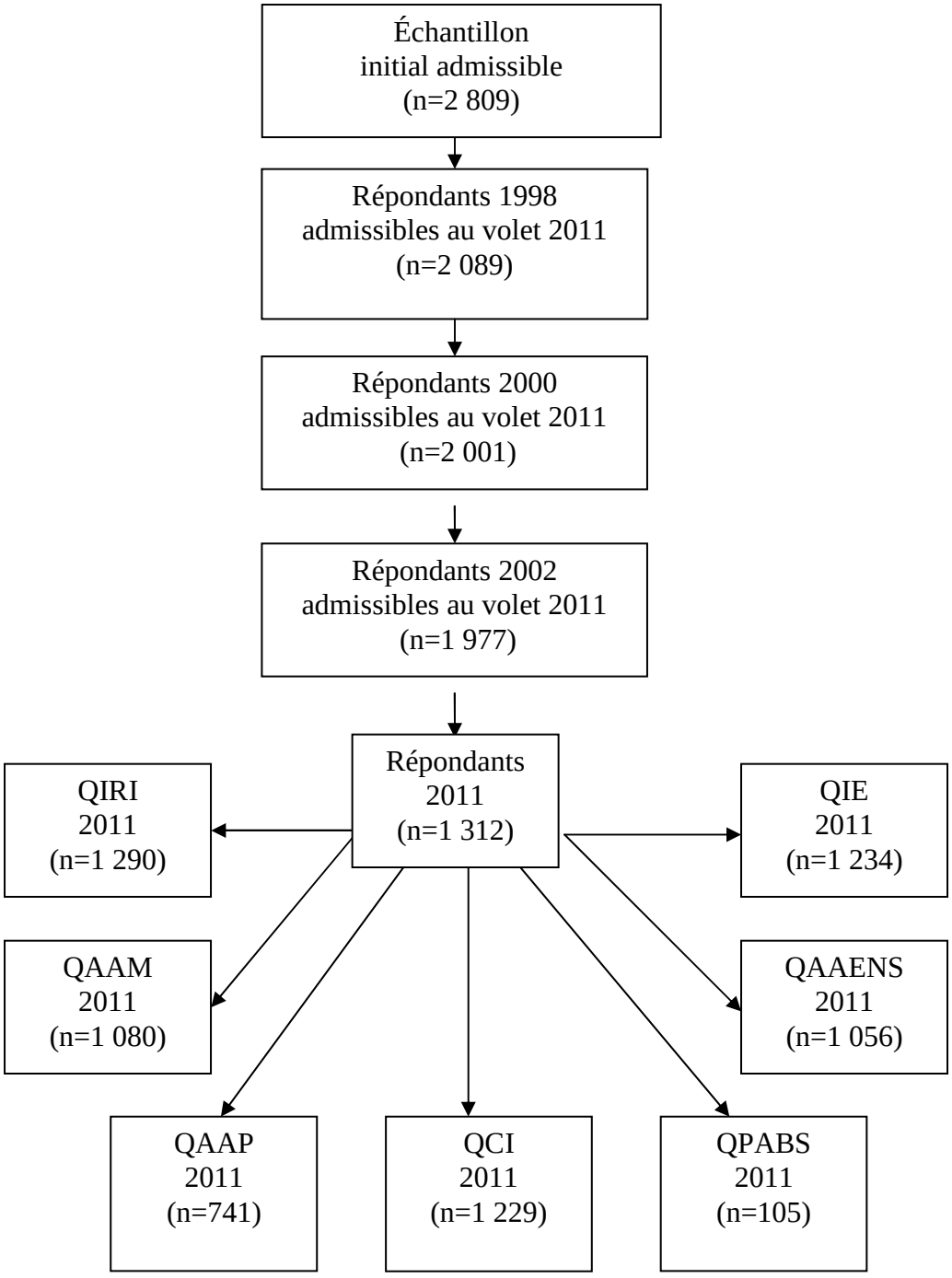
Pour effectuer le calcul des poids du volet 2002 pour les répondants (et pseudo-répondants), les classes de pondération définies au volet 2002 ont été conservées; seules les proportions pondérées de répondants ont été recalculées. Pour les variables servant à créer les classes de pondération, des valeurs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

Il est à noter que parmi ces 1 977 enfants visés par l'analyse de la non-réponse, il y en a 105 qui ne faisaient pas partie de celles qui ont été sollicitées pour la collecte de données au volet 2011. En effet, à la fin de chaque collecte, un examen des unités enquêtées pour ce volet est effectué. Les enfants cible 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 105 enfants n'avaient aucune possibilité de participer au volet 2011, contrairement aux autres enfants non-répondants au volet 2011. Ces différents types de non-répondants ont tout de même été modélisés ensemble.

La hiérarchie des répondants admissibles au volet 2011 est illustrée au schéma I.

13 Ont rempli le QIRI seulement pour les volets 2003 et 2004.

Schéma I. Hiérarchie des répondants aux différents instruments de collecte, admissibles au volet 2011



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 revenu à aucun volet, à 1 ou 2 volets, à 3 volets ou plus); 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 micro-données administratives[14] 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 2011. Ces variables sont reliées au cheminement scolaire de l'enfant : existence d'un plan d'intervention actif, sexe, niveau scolaire, langue d'enseignement, langue maternelle, type de classe spéciale (s'il y a lieu), code de difficulté. Ce fichier permet aussi de créer une variable qui exprime la mobilité de l'enfant de 2010 à 2011. Cette variable a été créée en comparant les adresses de la résidence de l'enfant pour les années 2010 et 2011.

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

- statut d'immigrant de la mère (ESDMD1A)
- le plus haut niveau de scolarité de la mère/conjointe (EEDMD01)
- le nombre de frères/sœurs de l'enfant cible (EREED01)
- le sexe de l'enfant (sexe)
- la langue d'enseignement (lan_enseignement)

Une méthode de classification hiérarchique (Ward) a permis de regrouper les probabilités estimées en 5 classes de pondération. Le tableau IV présente les proportions pondérées de répondants au volet 2011 parmi les répondants au volet 2002 pour ces 5 groupes. De plus, il présente le nombre de répondants, parmi les 1 312, à qui la proportion pondérée sera appliquée en guise de correction de la non-réponse. Par exemple : il y a 493 répondants au volet 2011 dont le poids de référence sera ajusté par l'inverse de la proportion pondérée de la troisième classe de pondération, qui est de 62,5%.

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

Classe de pondération	Proportions pondérées de répondants au volet 2011(en %)	Nombre de répondants
1	29,3	21
2	48,4	91
3	62,5	493
4	71,5	380
5	80,0	327

14 Micro-données provenant du fichier daté d'avril 2011 pour les enfants inscrits à l'école au 30 septembre 2010.

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondants varie de 29 % à 80 % (relativement à une proportion globale de 65,1 %). La proportion la plus faible est observée dans une classe où on retrouve, en proportion, un plus grand nombre d'enfants de sexe masculin; un plus grand nombre d'enfants dont la mère est immigrante; et un moins grand nombre d'enfants dont la mère avait un diplôme universitaire en 2002.

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'ÉLDEQ[15] (N=70 119). Ces statistiques agrégées sont en fait des totaux pour des caractéristiques choisies par l'ISQ et disponibles au MELS. Par exemple : la répartition des 70 119 enfants selon la région de résidence. Ces statistiques agrégées obtenues pour les données du volet 2011 permettent d'évaluer la pertinence d'effectuer un ajustement à la pondération transversale. Cet ajustement, appelé « calage », est définit 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 à la non-réponse.

L'objectif du calage au volet 2011 est d'effectuer une correction supplémentaire pour diminuer le biais dû à la non-réponse 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 disponibles (un choix devait être fait pour limiter l'ampleur de la production de statistiques agrégées au MELS). En effet, l'ISQ reçoit annuellement, en plus des statistiques agrégées, un fichier de micro-données administratives pour l'essentiel de l'échantillon de départ de la cohorte[16]. Les variables provenant des données administratives du MELS sont donc disponibles :

- Au niveau des micro-données, pour l'ensemble de l'échantillon de l'ÉLDEQ ;
- 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 de l'enfant;
- Langue maternelle de l'enfant;
- Existence d'un plan d'intervention actif pour l'enfant à l'école;
- Niveau scolaire (primaire versus secondaire)
- Code de difficulté de l'enfant

Une comparaison a d'abord été effectuée entre la distribution pondérée de l'ensemble des répondants au volet 2011 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, c'est-à-dire les enfants dont le niveau scolaire était le primaire en 2011 alors que les enfants en cheminement régulier sont en secondaire I, qui présente un écart le plus important entre les deux distributions. La proportion pondérée d'enfants en situation de retard scolaire est de 7,5% dans l'échantillon (avant imputation) comparativement à 10,1% dans la population de l'ÉLDEQ. Ainsi, l'ajustement des poids s'est effectué selon cette variable seulement, afin de rendre la distribution

15 L'ensemble des enfants nés au Québec entre le 1er octobre 1997 et le 30 septembre 1998 qui fréquentent le système scolaire québécois au 30 septembre 2010.

16 Fichier en date d'avril 2011 pour les enfants inscrits à l'école au 30 septembre 2010.

pondérée des répondants selon la variable du retard scolaire, semblable à la distribution dans la population visée. Quelques variables du volet 2011 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 poids[17]. 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.

3. Taux de réponse

Le tableau V présente le taux de réponse pondéré transversal obtenu au volet 2011. Ce taux est obtenu en multipliant les taux obtenus aux différentes étapes de pondération, selon le cas.

Tableau V - Taux de réponse pondéré transversal au volet 2011

Taux de réponse au volet 1998	75,3 % (n=2 809)
Proportion de répondants (incluant les nouveaux répondants) au volet 2000 parmi les répondants au volet 1998 admissibles au volet 2011	95,0 % (n=2 089)
Proportion de répondants au volet 2002 parmi les répondants au volet 2000 admissibles au volet 2011 (incluant les nouveaux répondants)	98,7 % (n=2 001)
Proportion de répondants au volet 2011 parmi les répondants au volet 2002 admissibles au volet 2011 (incluant les nouveaux répondants)	65,1 % (n=1 977)
Taux de réponse transversal au volet 2011	46,0 %

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

17 Dans le QAAENS : faible potentiel d'apprentissage scolaire de l'élève, peu d'importance accordée aux études.
Dans le QIRI : problème chronique de déficit d'attention (avec ou sans hyperactivité) et degré de réussite dans l'ensemble est faible ou très faible selon la PCM.

4. Utilisation de la pondération par les utilisateurs des données du volet 2011

4.1 L'importance de la pondération

Les utilisateurs des données du volet 2011 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 2011 au niveau transversal. Un second document traite de la non-réponse partielle à une question[18]. Au niveau transversal, le taux de réponse au volet 2011 est de l'ordre de 46,0 % (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 2011 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 disponibles à des volets antérieurs pour chacun des enfants. Des variables administratives provenant du MELS ont aussi été considérées au volet 2011 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 Fichier de pondération

Le fichier POIDS1401 contient la variable de pondération PEGENT14 (poids général transversal du volet 2011).

4.3 Tests statistiques

Le poids contenu dans le fichier POIDS1401 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 2011, tel qu'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 du fichier POIDS1401 peut être utilisé pour faire des tests approximatifs.

Afin de pallier au caractère approximatif des tests statistiques réalisés à l'aide de poids échantillonnaux, 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).

18 Voir le document « Étude de la non-réponse partielle au volet 2011 » par Fontaine et Courtemanche (2012).

Dans le cas particulier de tests du khi-deux sur un tableau de fréquences, l'utilisation des poids échantillonnaux divisés par un effet de plan moyen égal à 1,3 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 échantillonnaux 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 globale 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 échantillonnaux 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 échantillonnaux peut différer de façon importante de la taille d'échantillon pour un sous-groupe. Cela a pour conséquence d'invalider 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.

4.4 Choix de la pondération

Les possibilités d'analyse incluant des données du volet 2011 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'ISQ 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'ISQ de faire un ajustement sommaire 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 25ème Symposium international sur les questions de méthodologie de Statistique Canada, Ottawa, octobre 2009 (à paraître).

Fontaine, C. et Courtemanche, R. (2012). Étude de la non-réponse partielle au volet 2011, 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 jacknife 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 2011.

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 multicolinéarité. 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 multicolinéarité 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 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 2011 est constituée.