PONDÉRATION DES DONNÉES DU VOLET 2008

Catherine Fontaine, Luc Belleau, Nathalie Plante et Robert Courtemanche
Direction de la méthodologie et de la qualité
Institut de la statistique du Québec
30 juillet 2009

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 2008. La section 1 présente une description des étapes ayant mené au choix de la stratégie de pondération. Les taux de réponse pondérés obtenus sont détaillés à la section 2. À la section 3 se trouve l'analyse de la non-réponse totale ayant donné lieu à chacune des pondérations. Finalement, la section 4 renseigne l'utilisateur sur le fichier de pondération ainsi que sur la façon d'utiliser les poids échantillonnaux dans les analyses statistiques. Cette section comporte également quelques mises en garde sur l'utilisation de ces poids. Finalement, **on retrouve en annexe deux exemples d'évaluation des pondérations disponibles pour une situation d'analyse donnée, ainsi que les étapes détaillées de la création d'une pondération.**

1. Stratégie de pondération :

1.1 Admissibilité à l'enquête au volet 2008

Parmi les 2 120 répondants au volet initial, on compte 26 familles ayant quitté définitivement le Québec entre 1998 et 2008 et trois familles dont l'enfant cible est décédé entre les volets 1998 et 2008. Ces familles, considérées comme inadmissibles à l'enquête, ne sont plus visées par l'enquête en ce sens qu'elles 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 2008 ou qui ont quitté la province temporairement.

Les familles n'ayant pu être retracées, ayant refusé de répondre ou ayant été dans l'impossibilité de le faire sont toutes considérées admissibles à l'enquête. Bien que l'on sache que parmi celles n'ayant pu être retracées il pourrait y avoir des familles déménagées définitivement hors du Québec, leur nombre est trop petit pour que l'on en tienne compte dans la pondération. Sur cette base, l'échantillon admissible à l'enquête au volet 2008 est composé de 2 091 familles. Leur répartition, selon la réponse à l'enquête à chacun des volets de 1998 à 2008, est présentée au tableau I.

Tableau I - Nombre de répondants[1] aux volets de 1998 à 2008

Volet 1998	Volet 1999	Volet 2000	Volet 2001	Volet 2002	Volet 2003	Volet 2004	Volet 2005	Volet 2006	Volet 2008	Nombre de répondants
Oui	Répondant aux 4 volets				Répondants aux 4 volets				Oui	1 186
Oui									Non	99
Oui					Répondants à 2 ou 3 volets				Oui	145
Oui									Non	179
Oui					Répondants à 1 volet				Oui	18
Oui									Non	162
Oui					Répondant à 0 volet				Oui	3
Oui									Non	96
Oui	Répondant à 2 ou 3 volets				Répondants aux 4 volets				Oui	17
Oui									Non	7
Oui					Répondants à 2 ou 3 volets				Oui	28
Oui									Non	8
Oui					Répondant à 1 volet				Oui	4
Oui									Non	12
Oui					Répondant à 0 volet				Oui	0
Oui									Non	37
Oui	Répondant à 1 volet				Répondant à 2 ou 3 volets				Oui	1
Oui									Non	0
Oui					Répondant à 1 volet				Oui	0
Oui									Non	2
Oui					Répondant à 0 volet				Oui	0
Oui									Non	34
Oui	Répondant à 0 volet				Répondant à 1 volet				Oui	0
Oui									Non	1
Oui					Répondant à 0 volet				Oui	0
Oui									Non	52
Nombre total d'enfants admissibles à l'enquête au volet 2008										2 091

[1] Aux volets 1998 à 2002, les répondants à un volet ont tous complété le QIRI; à partir du volet 2003, les répondants ont complété au moins un instrument de collecte au volet concerné.

1.2 Répondants au volet 2008

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 2008, il a été décidé de créer seulement deux pondérations principales, de manière à couvrir, à tout le moins, les trois situations d'analyse suivantes :

1. Analyse des variables du volet 2008 portant sur l'ensemble des enfants ayant répondu à l'enquête (avec peu de données manquantes pour ces variables ou des variables d'autres volets incluses dans l'analyse).

2. Analyse des scores aux tests cognitifs du volet 2008 pour l'ensemble des enfants ayant participé à ces tests (avec peu de scores manquants et peu de données manquantes pour les autres variables considérées pour l'analyse).

3. Analyse des variables des volets 1998 à 2008 portant sur l'ensemble des enfants ayant répondu à chacun des volets de l'enquête, soit de 1998 à 2008 (avec peu de données manquantes).

Concrètement, les situations 1 et 2 requièrent une pondération transversale unique, puisque la non-réponse aux tests cognitifs, parmi les répondants à l'enquête au volet 2008, est négligeable, et la situation 3 nécessite une pondération longitudinale.

Le nombre de familles qui ont complété au moins un instrument de collecte au volet 2008 et qui étaient non répondantes à au moins un volet de 1999 à 2006 est relativement important (216 familles sur 1 402, soit environ 15%), d'où la nécessité de créer des pondérations transversale et longitudinale distinctes. S'il y a lieu, les autres situations d'analyse doivent être évaluées afin de déterminer si l'une des pondérations principales est appropriée. Dans le cas contraire, une pondération sur mesure doit être produite.

Notons qu'au volet 2008 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 2008 l'écart entre le nombre de répondants au QIRI et le nombre de répondants à ce volet est important (voir tableau II).

Tableau II- Nombre de répondants aux volets de 1998 à 2008

	volet 1998	volet 1999	volet 2000	volet 2001	volet 2002	volet 2003	volet 2004	volet 2005	volet 2006	volet 2008
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
Nombre de répondants pour un volet donné	2 120	2 045	1997	1 950	1 944	1 776	1 529	1537	1 526	1 402
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	1 286	1 186

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[2] pour le calcul de la pondération du volet 2008 plutôt que les volets 2005 ou 2006. 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 les volets de 2003 à 2006. En outre, l'utilisation du volet 2002 permet un meilleur niveau de cohérence longitudinale pour les 5 caractéristiques liées à l'attrition[3]. Enfin, ce choix évite les multiples ajustements de non-réponse entre 2002 et 2008, 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 1 334 répondants au QIRI, de même que pour 68 enfants additionnels ayant répondu à au moins un autre instrument de collecte[4] au volet 2008, soit un total de 1 402 enfants. La méthode de pondération sera décrite plus en détail à la section 3.

La modélisation de la non-réponse au volet 2008 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 admissibles à l'enquête au volet 2008 → pondération QIRI du volet 2000

[2] L'année de référence fournit la pondération de base qui fera l'objet d'un ajustement pour la non-réponse survenue ultérieurement.
[3] Voir le document de Fontaine et Courtemanche (2009) portant sur l'étude de l'attrition dans l'ÉLDEQ.
[4] Des données sur l'enfant (provenant du QPAE ou du fichier des jeux) sont disponibles pour ces 68 enfants.

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 2008 → pondération QIRI du volet 2002

4. Ajustement des poids transversaux du volet 2002 pour la non-réponse à l'enquête au volet 2008 parmi les répondants du volet 2002 toujours admissibles à l'enquête au volet 2008 → pondération générale transversale du volet 2008

Afin d'obtenir une pondération transversale pour l'ensemble des 1 402 répondants du volet 2008, les enfants qui étaient répondants à au moins un volet à partir de l'année 2002 et se sont vu attribuer un poids QIRI pour le volet 2002[5], ce dernier constituant le poids de base[6] 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 1 402 enfants ayant répondu à l'enquête au volet 2008. Cette pondération peut également être utilisée pour une analyse de variables où une petite proportion d'enfants présenterait des valeurs manquantes[7].

1.3.2 Ajustement de la non-réponse au niveau longitudinal[8]

Dans le cas de la pondération longitudinale, la pondération QIRI du volet 2002 a également servi de référence pour faire l'ajustement pour la non-réponse. Ces poids sont dans ce cas-ci ajustés pour la non-réponse à l'un ou l'autre des volets de 2003 à 2008. L'année de référence pour la pondération longitudinale a été déterminée selon les mêmes critères que pour la pondération transversale.

La modélisation de la non-réponse au volet 2008 comporte également 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 1999 ou 2000 parmi les répondants du volet 1998 toujours admissibles à l'enquête au volet 2000 → pondération longitudinale QIRI des volets 1998 à 2000

[5] La méthode utilisée pour attribuer un poids QIRI à ces enfants aux volets antérieurs sera décrite à la section 3.

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

[7] 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.

[8] L'ajustement de la non-réponse au niveau transversal et celui au niveau longitudinal diffèrent quant à l'admissibilité des répondants du volet de référence. Cette différence a peu d'impact dans le calcul des poids.

3. Ajustement des poids longitudinaux des volets 1998 à 2000 pour la non-réponse à l'enquête au volet 2001 ou 2002 parmi les répondants des volets 1998 à 2000 toujours admissibles à l'enquête au volet 2002 → pondération longitudinale QIRI des volets 1998 à 2002.

4. Ajustement des poids longitudinaux des volets 1998 à 2002 pour la non-réponse à l'enquête à au moins un volet entre 2003 et 2008 parmi les répondants des volets 1998 à 2002 toujours admissibles à l'enquête au volet 2008 → pondération générale longitudinale pour les volets de 1998 à 2008.

La pondération longitudinale ainsi créée peut être utilisée pour l'analyse des variables qui prennent une valeur pour l'ensemble des 1 186 enfants ayant répondu à l'enquête à chacun des volets de 1998 à 2008. Encore une fois, la présence de quelques données manquantes pour les variables d'analyse n'invalide pas la pondération.

1.4 Les autres instruments de collecte

Aucune pondération pour un instrument particulier n'a été effectuée au volet 2008. En effet, il est prévu que des pondérations sur mesure soient produites pour des analyses spécifiques lorsque nécessaire. Ces pondérations devraient avoir subi un ajustement pour la non-réponse à un instrument et 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 alors que la hiérarchie des répondants est illustrée au schéma I.

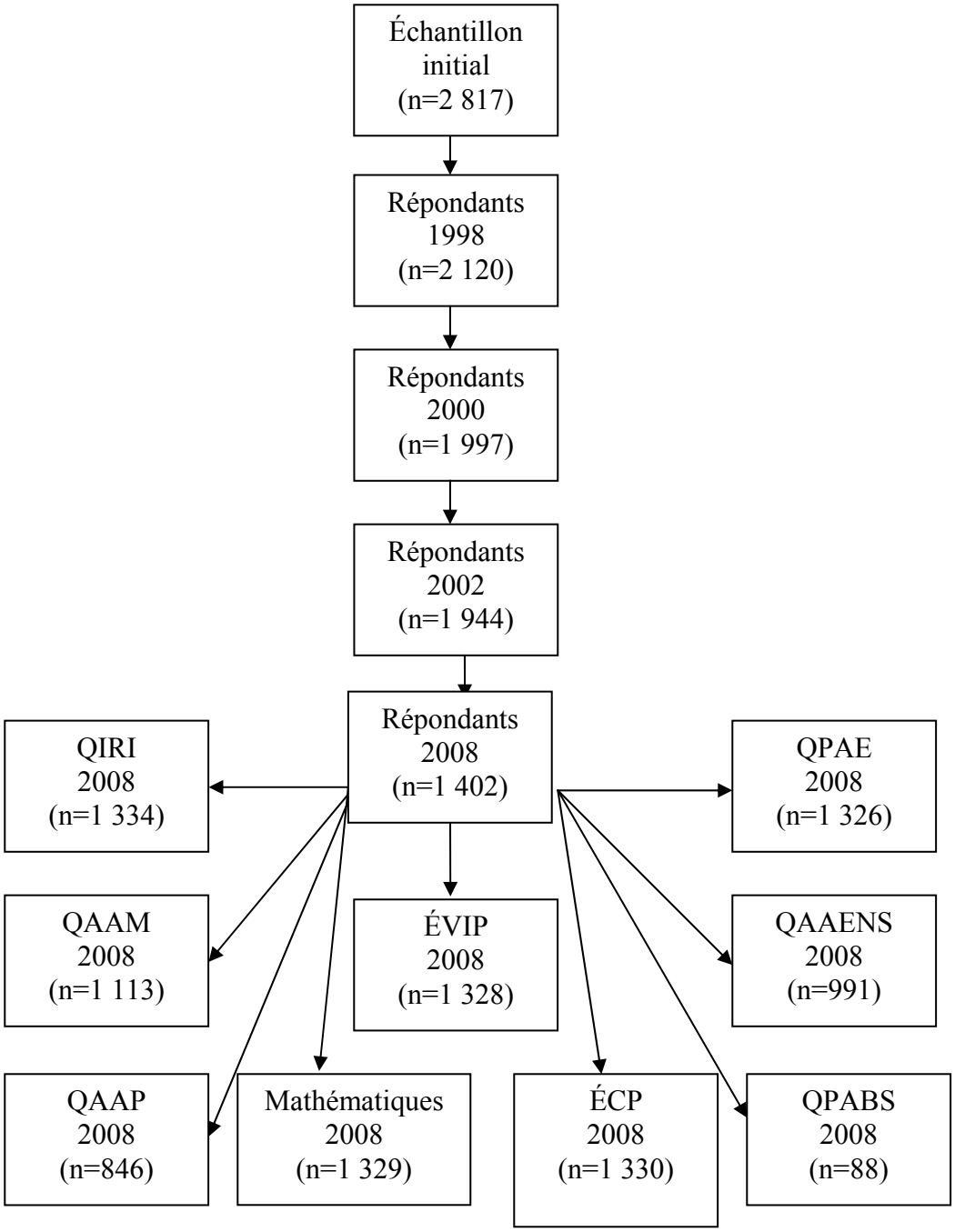
Tableau III - Nombre de répondants par instrument au volet 2008

	Nombre de répondants	Proportion pondérée de répondants parmi les répondants au volet 2008
QIRI	1 334	94,6
QAAM[9]	1 113	77,5
QAAP[10]	846	67,9
QPAE[11]	1 326	93,7
QAAENS[12]	991	68,1
QPABS[13]	88	-
Tests cognitifs : ÉVIP	1 328	93,9
Tests cognitifs : mathématiques	1 329	94,1
Évaluation de la condition physique (ÉCP)	1 330	94,1

La proportion pondérée de répondants au QAAM est calculée avec comme dénominateur le nombre estimé de mères ou conjointes présentes dans le ménage en 2008. La même considération s'applique pour le QAAP. Quant au QPABS, la proportion n'a pas été calculée puisque le nombre de pères biologiques absents du ménage est indéterminé pour 2008.

[9] Questionnaire auto-administré de la mère/conjointe.
[10] Questionnaire auto-administré du père/conjoint.
[11] Questionnaire papier administré à l'enfant.
[12] Questionnaire auto-administré de l'enseignant
[13] Questionnaire papier pour les pères biologiques absents.

Schéma I. Hiérarchie des répondants aux différents instruments de collecte



2. Taux de réponse :

Le tableau IV présente le taux de réponse pondéré transversal obtenu au volet 2008. Ce taux est obtenu en multipliant les taux obtenus aux différentes étapes de pondération, selon le cas. La démarche de modélisation ainsi que les résultats obtenus pour chacune des pondérations sont présentés à la section 3.

Tableau IV - Taux de réponse pondéré transversal au volet 2008

Taux de réponse au volet 1998	75,3% (n=2 809)
Proportion de répondants au volet 2000 parmi les répondants au volet 1998 admissibles au volet 2008 (incluant les nouveaux répondants[14])	95,0 % (n=2 091)
Proportion de répondants au volet 2002 parmi les répondants au volet 2000 admissibles au volet 2008 (incluant les nouveaux répondants)	98,5 % (n=2 003)
Proportion de répondants au volet 2008 parmi les répondants au volet 2002, admissibles au volet 2008	69,4 % (n=1 976)
Taux de réponse transversal au volet 2008	48,9 %

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

[14] Cette proportion diffère légèrement de la proportion des « vrais répondants » telle que présentée dans les documents antérieurs puisque, aux fins de la pondération du volet 2008, on redéfinit ici certains enfants répondants au volet 2000 alors qu'ils ne l'étaient pas en réalité, mais qui étaient répondants à un volet ultérieur (voir section 3.2).

Le tableau V présente le taux de réponse pondéré longitudinal aux volets 1998 à 2008. Ces taux ont également été calculés en concordance avec la démarche de pondération.

Tableau V - Taux de réponse pondéré longitudinal pour les volets de 1998 à 2008

Taux de réponse au volet 1998	75,3% (n=2 809)
Proportion de répondants aux volets 1998 à 2000 parmi les répondants au volet 1998, admissibles au volet 2000	93,7 % (n=2 101)
Proportion de répondants aux volets 1998 à 2002 parmi les répondants aux volets 1998 à 2000, admissibles au volet 2002	95,1 % (n=1 979)
Proportion de répondants aux volets 1998 à 2008 parmi les répondants aux volets 1998 à 2002, admissibles au volet 2008	60,6 % (n=1 888)
Taux de réponse longitudinal pour les volets 1998 à 2008	**40,7 %**

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

Les taux de réponse présentés aux tableaux IV et V peuvent sembler faibles. Par contre, il convient de rappeler que ces taux sont obtenus dans le cadre d'une enquête longitudinale où un très grand nombre de variables historiques sont disponibles pour effectuer les ajustements de la non-réponse. À taux de réponse égal, les risques de biais sont moindres s'ils sont obtenus lors d'une enquête longitudinale que lors d'une enquête transversale.

3. Analyse de la non-réponse :

3.1 Démarche générale d'analyse

La création de pondérations ajustées pour la non-réponse est basée sur la formation de classes de pondération. Contrairement aux volets précédents, c'est la méthode du score qui a été utilisée pour créer les classes de pondération (lire entre autres des détails sur la méthode dans Haziza et Beaumont, 2007). Elle a été retenue puisqu'elle améliorait le niveau de cohérence longitudinale des estimations[15]. 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 B.

[15] Une description détaillée des analyses ayant mené à ce changement se retrouve dans le document : Modélisation de la non-réponse globale du volet 2006 pour l'Étude Longitudinale sur le Développement des Enfants du Québec (ÉLDEQ) à l'aide de la méthode du score par Fontaine et Courtemanche, 2009.

3.2 Pondération transversale des données du volet 2008

3.2.1 Conversion de non-répondants au volet 2002

La pondération transversale du volet 2008 vise à attribuer un poids aux 1 402 répondants de ce volet, à partir du poids QIRI du volet 2002. Aux 1 938 répondants du volet 2002, toujours admissibles au volet 2008, était déjà associé un poids QIRI du volet 2002. Cependant, 26 enfants répondants au volet 2008 n'étaient pas répondants au volet 2002 et n'ont de ce fait aucun poids de référence. Aux fins de la pondération transversale du volet 2008, ces enfants ont été considérés répondants au volet 2002, de manière à recalculer un nouveau poids pour l'ensemble des répondants au volet 2002, incluant ces derniers[16].

Pour ce faire, 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 2002, aux seules fins de la pondération.

3.2.2 Variables considérées et résultats

Pour tenir compte de la non-réponse au volet 2008, un ajustement a été fait à partir de la pondération modifiée du volet 2002 (section 3.2.1). Cet ajustement est requis puisque les répondants au volet 2008 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 (PEGENT11) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 402 enfants ayant répondu à l'enquête au volet 2008.

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 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).

[16] Douze enfants supplémentaires qui n'ont répondu ni au volet 2002, ni au volet 2008, mais qui ont répondu à au moins un volet de 2003 à 2006 ont également été considérés répondants au volet 2002, de manière à obtenir un poids transversal au volet 2008 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.

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

- la langue parlée à la maison par la mère/conjointe (ESDMD6A)
- le plus haut niveau de scolarité de la mère/conjointe (EEDMD01)[17]
- le nombre de frères/sœurs de l'enfant cible (EREED01)[18]
- la situation en emploi des parents (variable longitudinale - nombre de volets avec les parents sans emploi)

Une méthode de classification non hiérarchique a permis de regrouper les probabilités estimées en 5 groupes de pondération. Le tableau VI présente les proportions pondérées de répondants au volet 2008 parmi les répondants au volet 2002 pour ces 5 groupes.

Tableau VI: Proportions pondérées de répondants par classe de pondération (transversal)

Classe de pondération	Proportions pondérées de répondants au volet 2008 (en %)
1	60,8
2	69,8
3	48,3
4	76,1
5	87,6

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondant varie de 48 % à 88 %. La proportion la plus faible est observée dans la classe où on compte, en proportion, un plus grand nombre d'enfants dont la mère est peu scolarisée (D.E.S. ou moins) ; dont la mère ne parle ni le français, ni l'anglais à la maison ; dont le nombre de frères ou sœurs est de 0 ou de 4 et plus ; et dont les parents étaient sans emploi à au moins un volet entre 1998 et 2002.

3.3 Pondération longitudinale des données des volets de 1998 à 2008

La non-réponse aux volets de 2003 à 2008[19], parmi les répondants longitudinaux du volet 2002[20], a été modélisée à l'aide des mêmes variables que celles étudiées pour la pondération transversale. Parmi l'ensemble des variables considérées, voici celles qui ont été retenues pour le modèle final de régression logistique :

- la langue parlée à la maison par la mère/conjointe (ESDMD6A)
- le plus haut niveau de scolarité de la mère/conjointe (EEDMD01)
- le nombre de frères/sœurs de l'enfant cible (EREED01)
- le revenu du ménage (variable longitudinale – seuil de 10 000 $)

Par la même méthode, les probabilités estimées ont été regroupées en cinq groupes de pondération. Le tableau VII présente les proportions pondérées de répondants aux volets de 1998 à 2008 parmi les répondants aux volets de 1998 à 2002 pour les 5 groupes.

[17] Une catégorie pour les valeurs manquantes a été créée pour cette variable dans le modèle final.
[18] Voir note 17.
[19] Cette non-réponse est ici définie comme la non-réponse à au moins un volet de 2003 à 2008.
[20] Seuls sont ici considérés les enfants répondants du volet 2002 qui étaient toujours admissibles à l'enquête au volet 2008.

Tableau VII : Proportions pondérées par groupe de pondération (longitudinal)

Classe de pondération	Proportions pondérées de répondants aux volets de 1998 à 2008 (en %)
1	31,9
2	65,9
3	61,3
4	51,3
5	80,8

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondant varie de 32 % à 81 %. La proportion la plus faible est observée dans la classe où on compte, en proportion, un plus grand nombre d'enfants dont la mère est peu scolarisée (D.E.S. ou moins) ; dont la mère ne parle ni le français, ni l'anglais à la maison ; dont le nombre de frères ou sœurs est de 4 ou plus ; et dont les revenus du ménage étaient faibles (moins de 10 000 $) à au moins un volet entre 1998 et 2002.

La nouvelle variable de pondération longitudinale (PEGENL11) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 186 enfants ayant répondu à l'enquête à chacun des volets de 1998 à 2008.

4. Fichier de pondération, mises en garde et recommandations aux fins de l'analyse

4.1 Fichier de pondération

Le fichier SAS POIDS1101 contient les variables de pondération suivantes: PEGENT11 (poids général transversal du volet 2008) et PEGENL11 (poids général longitudinal des volets 1998 à 2008). Ces poids doivent être utilisés dans les analyses afin d'inférer les résultats à la population visée tout en minimisant les biais dans les estimations.

4.2 Tests statistiques

Les poids contenus dans le fichier POIDS1101 sont des poids échantillonnaux, c'est-à-dire des poids qui ont été multipliés par une constante de sorte que la somme des poids soit égale à la taille de l'échantillon. Ces poids peuvent par conséquent être utilisés pour faire des tests approximatifs avec des logiciels qui ne tiennent pas compte du plan de sondage complexe dans l'estimation de la variance et les tests statistiques.

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.

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.

4.3 Choix de la pondération

Les possibilités d'analyse incluant des données du volet 2008 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. Deux exemples illustrant la démarche à suivre pour évaluer la situation sont présentés à l'annexe A.

5. Références bibliographiques :

Belleau. L., Fontaine, C. et Courtemanche, R. (2009). Étude de la non-réponse partielle au volet 2008, document interne, Institut de la statistique du Québec.

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). Modélisation de la non-réponse globale du volet 2006 pour l'Étude Longitudinale sur le Développement des Enfants du Québec (ÉLDEQ) à l'aide de la méthode du score, document interne, Institut de la statistique du Québec.

Fontaine, C. et Courtemanche, R. (2009). Étude de l'attrition pour l'Étude Longitudinale sur le Développement des Enfants du Québec de 1998 à 2008, 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

ANNEXE A

Le choix d'une pondération – deux exemples

On peut choisir d'utiliser la pondération qui minimise la perte d'unités d'analyse afin d'optimiser la précision des estimations et la puissance des tests statistiques. En plus de diminuer la puissance statistique, une perte d'unités d'analyse pourrait entraîner certains biais dans les estimations. Mise à part la perte d'unités d'analyse, le choix d'une pondération doit également considérer la qualité de l'ajustement pour la non- réponse totale ou partielle. En effet, un ajustement incomplet pour la non-réponse étant source de biais, il faut s'assurer que la part de non-réponse pour laquelle aucun ajustement n'a été fait est négligeable. Une proportion de 5 % ou moins peut être considérée faible, voire négligeable. Entre 5 % et 10 %, la non-réponse pourrait être négligée si elle ne présente pas de lien important avec les variables étudiées. Une étude sommaire de biais devrait être réalisée en ce sens afin de juger de la situation. 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.

Exemple 1

Supposons que l'on veuille étudier la relation entre des variables provenant du QAAENS (2008), du QIRI (2008) et un score à l'épreuve de mathématiques de 2006 (ICAESTOT). Si l'on suppose qu'il n'y a aucune non-réponse partielle pour les variables considérées, on compterait 889 unités pour cette analyse. Toutes ces unités ont un poids transversal du volet 2008 (PEGENT11). Par contre, 805 enfants ont un poids longitudinal des volets 1998 à 2008 (PEGENL11), pour une perte d'unités de 9,4% ((889-805)/889).

Quant à la part de non-réponse qui n'est pas prise en compte avec la pondération transversale 2008, elle est d'environ 37% ((1402-889)/1402). Cette part diminue avec l'utilisation de la pondération longitudinale 1998 à 2008, mais demeure importante à 32% ((1186-805)/1186).

Si l'on convient qu'il est préférable de minimiser la perte d'unités, c'est la pondération transversale 2008 qui est à privilégier. Toutefois, la part importante de non-réponse non expliquée fait en sorte qu'une pondération sur mesure est à envisager pour minimiser les biais d'analyse.

Exemple 2

Supposons que l'on veuille mettre en relation des variables provenant de l'ÉVIP (2008), du test de mathématiques (2008) et des variables provenant du QIRI de chacun des volets suivants : 2003, 2005 et 2008. Si l'on suppose qu'il n'y a aucune non-réponse partielle pour les variables considérées, on compterait 1 165 unités pour cette analyse.

On retrouve 1 101 unités qui ont un poids longitudinal pour les volets 1998 à 2008 (PEGENL11), pour une perte d'unités non pondérée d'environ 5% ((1165-1101)/1165). Toutes les unités d'analyse ont un poids transversal du volet 2008 (PEGENT11).

Quant à la part de non-réponse qui n'est pas prise en compte avec la pondération longitudinale des volets 1998 à 2008, elle est d'environ 7% ((1186-1101)/1186). Avec l'utilisation de la pondération transversale de 2008, la part de non-réponse non prise en compte demeure importante à près de 17% ((1402-1165)/1402).

Si l'on convient qu'il est préférable de minimiser la perte d'unités, c'est la pondération transversale de 2008 qui est à privilégier. Toutefois, l'utilisation de cette pondération pourrait entraîner des biais importants puisque la part de non-réponse non prise en compte n'est pas négligeable.

Une pondération sur mesure serait donc à envisager. Par contre, si la perte d'unités qu'entraîne l'utilisation de la pondération longitudinale de 1998 à 2008 n'est pas un obstacle important à l'analyse, cette dernière pourrait être utilisée. Par contre, une analyse des biais potentiels devrait être effectuée et si les caractéristiques des non-répondants ne sont pas liées aux variables à l'étude, la non-réponse pourrait être ignorée.

ANNEXE B

Les étapes de la création d'une pondération

La séquence des étapes de création des deux pondérations du volet 2008, c'est-à-dire la pondération transversale et la pondération générale, est la même. En voici la description.

Étape 1 : Analyses bivariées pour réduire le nombre de variables considérées pour la modélisation (environ 40 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.

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 : Imputation des données manquantes ou 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 : La création des classes de pondération s'effectue 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 de méthodes de classification hiérarchiques et non-hiérarchiques. Ceci étant fait, les poids de base sont ajustés selon la proportion pondérée de répondants par classe, pour ainsi constituer la pondération 2008.