PONDÉRATION DES DONNÉES DU VOLET 2010

Catherine Fontaine, Luc Belleau et Robert Courtemanche
Direction de la méthodologie et de la qualité
Institut de la statistique du Québec
1er septembre 2011

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 2010. Il s'inspire en grande partie du même rapport rédigé à des volets précédents[1]. Pour savoir comment utiliser la pondération, le lecteur est invité à consulter en tout premier lieu la section 4. Cette 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 concernant l'utilisation de ces poids. Par ailleurs, le lecteur est invité à consulter les sections 1 à 3 pour savoir comment a été effectuée la pondération. La section 1 présente une description des étapes ayant mené au choix de la stratégie de pondération. Le taux de réponse pondéré obtenu est détaillé à la section 2. À la section 3 se trouve l'analyse de la non-réponse totale ayant donné lieu à chacune des pondérations.

1. Stratégie de pondération

1.1 Admissibilité à l'enquête au volet 2010

Parmi les 2 120 répondants au volet initial, on compte 27 familles ayant quitté définitivement le Québec entre 1998 et 2010 et trois familles dont l'enfant cible est décédé entre les volets 1998 et 2010. 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 2010 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 2010 est composé de 2 090 familles. Leur répartition, selon la réponse à l'enquête à chacun des volets de 1998 à 2010, est présentée au tableau I.

1. Voir le site Internet : http://www.jesuisjeserai.stat.gouv.qc.ca/doc_tech.htm

Tableau I - Nombre de répondants[2] aux volets de 1998 à 2010

Volet 1998	Volet 1999	Volet 2000	Volet 2001	Volet 2002	Volet 2003	Volet 2004	Volet 2005	Volet 2006	Volet 2008	Volet 2010	Nombre de répondants
Oui	Répondant aux 4 volets				Répondants aux 5 volets					Oui	1 121
Oui										Non	65
Oui					Répondants à 4 volets					Oui	133
Oui										Non	75
Oui					Répondants à 2 ou 3 volets					Oui	81
Oui										Non	151
Oui					Répondants à 1 volet					Oui	25
Oui										Non	140
Oui					Répondant à 0 volet					Oui	9
Oui										Non	87
Oui	Répondant à 2 ou 3 volets				Répondants aux 5 volets					Oui	15
Oui										Non	2
Oui					Répondants aux 4 volets					Oui	15
Oui										Non	6
Oui					Répondants à 2 ou 3 volets					Oui	11
Oui										Non	15
Oui					Répondant à 1 volet					Oui	1
Oui										Non	11
Oui					Répondant à 0 volet					Oui	4
Oui										Non	33
Oui	Répondant à 1 volet				Répondant à 4 volets					Oui	0
Oui										Non	1
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 2010											2 090

2. 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 2010

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 2010, il a été décidé de créer seulement deux pondérations principales, de manière à couvrir, à tout le moins, les deux situations d'analyse suivantes :

- Analyse des variables du volet 2010 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).

- Analyse des variables du volet 2010 provenant du Questionnaire auto-administré de l'enseignant (QAAENS) (avec peu de données manquantes pour ces variables ou des autres variables incluses dans l'analyse).

Les situations 1 et 2 requièrent chacune une pondération transversale distincte. En effet, le nombre de familles pour lesquelles le QAAENS a été complété est de 1 008, comparativement à 1 415 familles répondantes à au moins un instrument de collecte pour le volet 2010 (proportion non pondérée d'environ 70%). La non-réponse au QAAENS par rapport à l'ensemble des répondants au volet 2010 pourra être prise en compte avec la pondération transversale du QAAENS.

Pour le volet 2010, il n'y aura pas de pondération longitudinale distincte. Les 1 121 répondants longitudinaux de 1998 à 2010 forment un sous-ensemble par rapport aux 1 186 répondants longitudinaux de 1998 à 2008. La non-réponse à au moins un volet de 1998 à 2010 par rapport au sous-ensemble des répondants longitudinaux de 1998 à 2008 est de 5 % (1 186-1 121/1 186). Comme le taux de non-réponse est faible, c'est la pondération longitudinale de 1998 à 2008 qui pourra être utilisée lors d'analyses combinant des variables des volets 1998 à 2010, avec peu de valeurs manquantes pour ces variables.

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. Ce sera probablement le cas lors de l'analyse des variables du Questionnaire auto-administré de la mère/conjointe (QAAM) au volet 2010. Le poids transversal calculé pour l'ensemble des enfants ayant répondu au volet 2010 comporte une grande portion de non-réponse au QAAM qui n'a pas été prise en compte (proportion non pondérée d'environ 11%, voir le tableau III).

Notons qu'au volet 2010 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 2010 l'écart entre le nombre de répondants au QIRI et le nombre de répondants à ce volet est plus faible qu'au volet 2008 (voir tableau II).

Tableau II- Nombre de répondants aux volets de 1998 à 2010

	volet 1998	volet 1999	volet 2000	volet 2001	volet 2002	volet 2003	volet 2004	volet 2005	volet 2006	volet 2008	volet 2010
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
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	1 415
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	1 121

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[3] pour le calcul de la pondération du volet 2010 plutôt que les volets 2008 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 2008. En outre, il a été démontré lors de l'analyse des pondérations de 1998 à 2006 que l'utilisation du volet 2002 permettait d'atteindre un meilleur niveau de cohérence longitudinale pour les 5 caractéristiques liées à l'attrition[4]. Enfin, ce choix évite les multiples ajustements de non-réponse entre 2002 et 2010, 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 1 396 répondants au QIRI, de même que pour 19 enfants additionnels ayant répondu à au moins un autre instrument de collecte[5] au volet 2010, soit un total de 1 415 enfants. La méthode de pondération sera décrite plus en détail à la section 3.

La modélisation de la non-réponse au volet 2010 comporte quatre étapes :

- Ajustement de l'inverse des probabilités de sélection pour la non-réponse à l'enquête au volet 1998 → pondération QIRI du volet 1998.

- Ajustement des poids QIRI du volet 1998 pour la non-réponse à l'enquête au volet 2000 parmi

3. L'année de référence fournit la pondération de base qui fera l'objet d'un ajustement pour la non-réponse survenue ultérieurement.
4. Voir le document de Fontaine et Courtemanche (2009) portant sur l'étude de l'attrition dans l'ÉLDEQ.
5. Des données sur l'enfant (provenant du Questionnaire informatisé à l'enfant ou du QAAENS) sont disponibles pour ces 19 enfants.

les répondants du volet 1998 toujours admissibles à l'enquête au volet 2010 → pondération QIRI du volet 2000.

- 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 2010 → pondération QIRI du volet 2002.

- Ajustement des poids transversaux du volet 2002 pour la non-réponse à l'enquête au volet 2010 parmi les répondants du volet 2002 toujours admissibles à l'enquête au volet 2010 → pondération générale transversale du volet 2010.

Afin d'obtenir une pondération transversale pour l'ensemble des 1 415 répondants du volet 2010, 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[6], 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 1 415 enfants ayant répondu à l'enquête au volet 2010. 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 transversal pour le QAAENS

Dans le cas de la pondération transversale du QAAENS, la pondération transversale générale du volet 2010, décrite à la section 1.3.1, a servi de référence pour faire l'ajustement pour la non-réponse. Ces poids sont dans ce cas-ci ajustés pour la non-réponse au QAAENS par rapport à l'ensemble des participants au volet 2010.

La modélisation de la non-réponse au volet 2010 comporte une étape, c'est-à-dire l'ajustement des poids transversaux généraux du volet 2010 pour la non-réponse au QAAENS au volet 2010 parmi les répondants du volet 2010 → pondération transversale du QAAENS pour le volet 2010.

La pondération transversale du QAAENS ainsi créée peut être utilisée pour l'analyse des variables qui prennent une valeur pour l'ensemble des 1 008 enfants ayant répondu au QAAENS en 2010. Encore une fois, la présence de quelques données manquantes pour les variables d'analyse n'invalide pas la pondération.

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

1.4 Les autres instruments de collecte

Pour le volet 2010, la pondération transversale pour le QAAENS est la seule pondération spécifique pour un instrument particulier 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 alors que la hiérarchie des répondants est illustrée au schéma I.

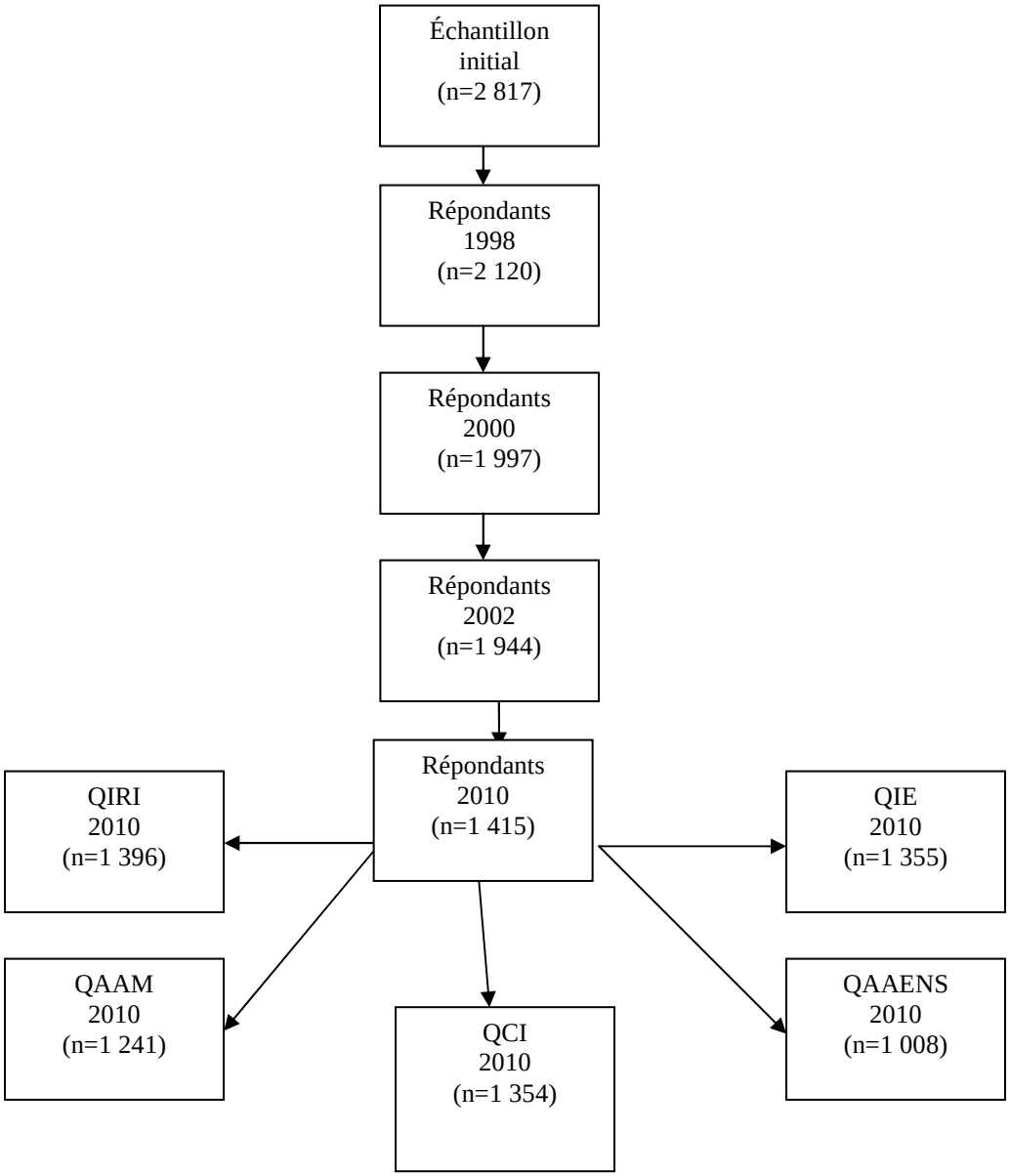
Tableau III - Nombre de répondants par instrument au volet 2010

	Nombre de répondants	Proportion pondérée de répondants parmi les répondants au volet 2010 (%)
QIRI	1 396	98,6 %
QAAM	1 241	88,6 %
QIE[8]	1 355	95,5 %
QAAENS	1 008	71,7 %
QCI[9]	1 354	95,5 %

La proportion pondérée de répondants au QAAM est calculée avec comme dénominateur le nombre estimé de mères ou conjointes présentes dans le ménage en 2010[10]. La proportion pondérée de répondants au QAAENS est calculée avec comme dénominateur le nombre d'enfants admissibles à ce questionnaire, c'est-à-dire les enfants qui fréquentent une école québécoise au moment de la collecte (n=1 380). En effet, le QAAENS n'a pas été envoyé par la poste aux enfants qui reçoivent leur éducation à la maison ou qui habitent à l'extérieur du Québec de façon temporaire.

8. Questionnaire informatisé à l'enfant.
9. Questionnaire complété par l'intervieweure.
10. Le nombre de mères ou conjointes présentes dans le ménage en 2010 doit être estimé puisque cette information provient du QIRI et que 19 familles n'ont pas complété le QIRI en 2010. Le dénominateur utilisé est de 1 390 enfants.

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 2010. 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 2010

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 2010 (incluant les nouveaux répondants[11])	95,0 % (n=2 090)
Proportion de répondants au volet 2002 parmi les répondants au volet 2000 admissibles au volet 2010 (incluant les nouveaux répondants)	98,6 % (n=2 002)
Proportion de répondants au volet 2010 parmi les répondants au volet 2002, admissibles au volet 2010	70,5 % (n=1 977)
Taux de réponse transversal au volet 2010	49,7 %

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

3. Analyse de la non-réponse

3.2 Pondération transversale des données du volet 2010

3.2.1 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 (lire entre autres des détails sur la méthode dans 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.

11. 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 2010, 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).

3.2.2 Conversion de non-répondants au volet 2002

La pondération transversale du volet 2010 vise à attribuer un poids aux 1 415 répondants de ce volet, à partir du poids QIRI du volet 2002. Aux 1 937 répondants du volet 2002, toujours admissibles au volet 2010, était déjà associé un poids QIRI du volet 2002. Cependant, 21 enfants répondants au volet 2010 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 2010, 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[12].

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.

3.2.3 Variables considérées et résultats

Pour tenir compte de la non-réponse au volet 2010, un ajustement a été fait à partir de la pondération modifiée du volet 2002 (section 3.2.2). Cet ajustement est requis puisque les répondants au volet 2010 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 (PEGENT13) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 415 enfants ayant répondu à l'enquête au volet 2010.

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).

Une nouvelle variable a également été considérée pour la pondération transversale générale. Celle-ci exprime la mobilité de l'enfant de 2008 à 2010. Cette variable a été créée en comparant les adresses de la résidence de l'enfant pour les années 2008, 2009 et 2010. La comparaison a permis de créer une variable à deux modalités : déménagement au moins une fois entre 2008 et 2010; aucun déménagement sinon. En présence de données manquantes pour ces coordonnées pour au moins une des trois années disponibles, des règles ont été fixées afin d'imputer la majorité des cas

12. Dix-neuf enfants supplémentaires qui n'ont répondu ni au volet 2002, ni au volet 2010, mais qui ont répondu à au moins un volet de 2003 à 2008 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.

causant des données manquantes pour la variable de mobilité. Cette variable de mobilité qui a été considérée lors de la modélisation montre un grand potentiel pour la réduction du biais dû à la non-réponse lors des analyses bivariées.

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)

La variable de mobilité n'a pas été retenue pour faire partie du modèle final. Cependant, elle a fait partie des dernières variables qui ont été exclues du modèle de régression logistique. Les trois variables conservées présentaient des seuils observés plus faibles que celui de la variable de mobilité.

Une méthode de classification non hiérarchique a permis de regrouper les probabilités estimées en 4 classes de pondération. Le tableau V présente les proportions pondérées de répondants au volet 2010 parmi les répondants au volet 2002 pour ces 4 groupes. De plus, il présente le nombre de répondants, parmi les 1 415, à qui la proportion pondérée sera appliquée en guise de correction de la non-réponse. Par exemple : il y a 165 répondants au volet 2010 dont le poids de référence sera ajusté par l'inverse de la proportion pondérée de la troisième classe de pondération, qui est de 59,2 %.

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

Classe de pondération	Proportions pondérées de répondants au volet 2010 (en %)	Nombre de répondants
1	50,5	19
2	71,3	874
3	59,2	165
4	80,7	357

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondants varie de 51 % à 81 % (relativement à une proportion globale de 70,5 %). 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 est immigrante ; et dont le nombre de frères ou sœurs est de 0 ou de 4 et plus.

3.2.4 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 peut obtenir de cet organisme des statistiques agrégées pour la population visée de l'ÉLDEQ[13] (N=70 101). Ces statistiques agrégées sont en fait des totaux pour des caractéristiques choisies par l'ISQ et disponibles au MELS. Par exemple : la répartition des 70 101 enfants selon la région de résidence. Ces statistiques agrégées obtenues pour les données du volet 2010 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

13. 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 2009.

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.

Après une analyse des variables disponibles pour le calage, il a été décidé de ne pas procéder à cet ajustement de la pondération du volet 2010. Le travail de pondération décrit à la section 3.2.3, ainsi que celui effectué aux volets précédents, a permis de bien effectuer la réduction du biais de non-réponse quant aux mesures de réussite scolaire considérées. Il convient toutefois de mentionner que cette analyse a été compliquée par la présence de données manquantes pour les variables administratives considérées. Le lecteur est invité à consulter l'annexe B pour une description complète de l'analyse effectuée.

3.3 Pondération transversale des données du QAAENS du volet 2010

3.3.1 Démarche générale d'analyse

La création de pondérations ajustées pour la non-réponse dans l'ÉLDEQ est basée sur la formation de classes de pondération. Pour la pondération transversale du QAAENS, c'est la modélisation par segmentation fondée sur l'algorithme CHAID (« Chi-square automatic interaction detection ») mis au point par Kass (1980) qui a été utilisée. Les classes de pondération sont créées sous forme d'arborescence ; elles ne résultent donc pas nécessairement du croisement de toutes les variables considérées pour la modélisation. Pour une famille donnée, l'ajustement de la pondération consiste à diviser le poids de référence par la proportion pondérée de familles répondantes observées au sein de la classe à laquelle elle appartient. Des tests du khi-deux approximatifs sont effectués au préalable à l'aide du logiciel SAS pour choisir un sous-ensemble de variables les plus liées à la non-réponse. Les variables considérées pour la modélisation avec CHAID sont celles dont le seuil est égal ou inférieur à 0,20. Certaines des variables retenues par la modélisation ont par ailleurs fait l'objet d'une analyse supplémentaire. Cette analyse consistait à vérifier si celles-ci étaient reliées aux mesures provenant du QAAENS. En effet, lors de la création des classes de pondération, il faut s'assurer que les variables choisies soient liées non seulement à la probabilité de répondre, mais aussi aux variables mesurées dans l'enquête. Autrement, la réduction du biais potentiel dû à la non-réponse pourrait être négligeable (Beaumont, 2002) ; elle ne compenserait pas dans ce cas la perte de précision due à la perturbation des poids[14].

Cette technique est celle utilisée lors de la pondération du même questionnaire, à des volets antérieurs. Par contre, elle diffère de celle utilisée pour la pondération transversale générale. Les deux méthodes présentent des similitudes, comme par exemple le fait de créer des groupes homogènes, selon certains critères, en ce qui concerne la probabilité de répondre. L'homogénéité est cependant évaluée différemment. En ce qui concerne la modélisation par segmentation, l'algorithme crée des groupes qui diffèrent de manière significative en ce qui concerne la proportion pondérée de répondre, suite à un regroupement optimal des modalités des variables considérées. Quant à la méthode du score, des méthodes de classification sont utilisées pour créer des groupes dont la probabilité de répondre, estimée à partir d'un modèle de régression logistique, est semblable.

14. Puisque les mesures sont manquantes pour les non-répondants, on doit étudier ce lien à partir du sous-ensemble des répondants seulement. Dans ce cas, on fait l'hypothèse que si un lien est détecté pour le sous-ensemble des répondants, il le serait aussi si le même test était effectué à partir de l'échantillon total.

La méthode du score a été retenue pour la pondération générale de 2006 (et les suivantes) puisqu'il a été observé qu'elle améliorait le niveau de cohérence longitudinale des estimations[15]. La méthode par segmentation demeure une option très intéressante pour les pondérations secondaires en raison de sa simplicité d'utilisation.

3.3.2 Variables considérées et résultats

La pondération des données du QAAENS consiste à ajuster le poids général transversal pour tenir compte de la non-réponse au QAAENS. Les variables considérées pour la modélisation proviennent du QIRI du volet 2010. En plus de variables de nature socioéconomique, des variables à propos de l'enfant selon la perception de la PCM ont également été considérées. Par exemple, le niveau de condition physique de l'enfant, les difficultés dans ses relations avec les autres, l'importance d'avoir des bonnes notes, etc. De plus, des variables provenant du fichier administratif du MELS pour l'année 2009-2010 ont été considérées (sexe de l'enfant, langue maternelle de l'enfant, existence d'un plan d'intervention actif, indice du milieu socio-économique et indice du seuil de faible revenu).

Parmi l'ensemble des variables considérées pour l'ajustement de la non-réponse au QAAENS du volet 2010, les variables suivantes ont été retenues :

- l'existence d'un plan d'intervention actif (PLAN_INTERVENTION)
- le niveau de suffisance du revenu (MINFD3A)
- la langue parlée par l'enfant: anglais (MSDEQ6BA)

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondant varie de 51 % à 77 % (relativement à une proportion globale de 71,7 %). La proportion la plus faible est observée dans la classe où on compte, en proportion, un plus grand nombre d'enfants ayant un plan d'intervention actif et dont les revenus de son ménage sont insuffisants.

La nouvelle variable de pondération (PEQAAENS13) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 008 enfants ayant répondu au QAAENS au volet 2010[16].

4. Utilisation de la pondération par les utilisateurs des données du volet 2010

4.1 L'importance de la pondération

Les utilisateurs des données du volet 2010 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 traite

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.
16. Le QAAENS vise les enfants qui fréquentent une école au Québec au moment de la collecte de données. Ainsi, les enfants qui habitent hors du Québec mais qui participent à l'enquête, ainsi que les enfants qui font l'école à la maison, sont exclus de l'inférence faite à partir de la pondération transversale du QAAENS au volet 2010.

du traitement pour la non-réponse survenue au volet d'enquête 2010 au niveau transversal. De plus, il traite de la correction pour la non-réponse survenue à l'instrument QAAENS, par rapport à l'ensemble des familles ayant complété au moins un instrument de collecte en 2010. Un second document traite de la non-réponse partielle à une question[17]. Au niveau transversal, le taux de réponse au volet 2010 est de l'ordre de 50,0 % (voir tableau V). Quant au QAAENS, la proportion pondérée de répondants en 2010 par rapport à l'ensemble des répondants de ce volet est de l'ordre de 70,0 % (voir tableau III). Ces faibles taux confirment l'importance du traitement effectué lors de la pondération.

La stratégie de pondération mise en œuvre pour créer les deux pondérations principales du volet 2010 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 2010 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 SAS POIDS1301 contient les variables de pondération suivantes: PEGENT13 (poids général transversal du volet 2010) et PEQAAENS13 (poids transversal au QAAENS du volet 2010).

4.3 Tests statistiques

Les poids contenus dans le fichier POIDS1301 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 doivent faire partie de toute analyse des données du volet 2010, 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, les poids du fichier POIDS1301 peuvent être utilisés 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.

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

17. Voir le document « Étude de la non-réponse partielle au volet 2010 » par Belleau, Fontaine et Courtemanche (2011).

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 2010 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

Beaumont, J.-F. (2002). Quand est-on en présence de non-réponse non-ignorable? *Le Bulletin d'imputation*, 2-1, pages 2-4.

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, Belleau, L. et Courtemanche, R. (2009). Étude de l'attrition pour l'Étude Longitudinale sur le Développement des Enfants du Québec de 1998 à 2008 : analyse des données administratives du MELS, document interne, Institut de la statistique du Québec.

Fontaine, C., Belleau, L. et Courtemanche, R. (2011). Étude de la non-réponse partielle au volet 2010, document interne, Institut de la statistique du Québec.

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

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied statistics*, 29, pages 119-127.

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.

Särndal, Carl-Erik (2007). La méthode de calage dans la théorie et la pratique des enquêtes. *Techniques d'enquête*, vol. 33, no. 2, pages 113-135.

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 2010.

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 non hiérarchiques. Ceci étant fait, les poids de base sont ajustés selon la proportion pondérée de répondants par classe, pour ainsi constituer la pondération 2010.

ANNEXE B

L'objectif du calage au volet 2010 est d'effectuer une correction supplémentaire pour la diminution du 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[18]. Les variables provenant des données administratives du MELS sont donc disponibles :

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

Les variables administratives disponibles à ces deux niveaux sont importantes pour la réduction du biais. Cependant, il a été décidé d'utiliser les variables administratives au niveau des macro-données seulement pour effectuer cette correction supplémentaire des poids. En effet, celles-ci seraient plus importantes que les variables disponibles au niveau des micro-données pour réduire la variance (Särndal, 2007).

Les variables administratives considérées pour le calage sont :

a. Sexe de l'enfant;
b. Région de résidence de l'enfant;
c. Langue maternelle de l'enfant;
d. Existence d'un plan d'intervention actif pour l'enfant à l'école;
e. Indice du seuil du faible revenu calculé par école (pour l'ordre d'enseignement primaire)[19];
f. Indice du milieu socio-économique calculé par école (pour l'ordre d'enseignement primaire).

Des tests bivariés, pondérés et tenant compte du plan de sondage, ont été effectués afin d'identifier les variables administratives les plus reliées aux mesures de réussite scolaire afin que la réduction du biais dû à la non-réponse soit maximisée pour ces variables. Quatre variables reliées à la réussite scolaire ont été sélectionnées :

a. La motivation intrinsèque au volet 2008 (variable kqeet9aa);
b. La présence d'un problème chronique d'hyperactivité ou d'inattention au volet 2008 (variable khleq45m);
c. Les aptitudes globales d'apprentissage au volet 2008 (variable kaeet04c);
d. Le score global à l'ÉVIP au volet 2008 (variable keves01).

Ces tests ont permis de déterminer que le sexe, l'existence d'un plan d'intervention actif, l'indice du milieu socio-économique et la langue maternelle sont les variables reliées à un plus grand nombre de mesures de la réussite scolaire.

Les analyses effectuées auprès des données administratives doivent prendre en compte la non-réponse partielle qui est présente dans le fichier administratif de micro-données et pour les statistiques agrégées. En ce qui concerne le fichier administratif, il est à noter qu'un certain nombre d'enfants de l'échantillon de départ n'ont pas fait l'objet d'un jumelage réussi. Cela peut être dû à

20. Fichier en date du 20 avril 2010 pour les enfants inscrits à l'école au 30 septembre 2009.
21. http://www.mels.gouv.qc.ca/stat/Indice_defav/index_ind_def.htm

de multiples causes, telles qu'une erreur dans le code permanent (utilisé comme clé d'appariement), une absence de l'enfant dans le système scolaire québécois, etc. En présence de jumelage réussi, il peut tout de même y avoir présence de non-réponse partielle pour les variables administratives considérées. Par exemple : les deux indices de défavorisation sont calculés par école. Par contre, ils ne sont pas calculés pour certaines écoles, telles les écoles privées. Cela constitue l'une des causes de données manquantes pour ces variables. Parmi les 1 415 répondants au volet 2010, la proportion pondérée de non-réponse partielle pour les quatre variables reliées aux mesures de réussite scolaire varie de 2,0 % à 10,0 %.

En ce qui concerne les données manquantes pour les statistiques agrégées, seule la variable de l'indice du milieu socio-économique présente ce problème, parmi les quatre variables retenues. L'absence de calcul de l'indice pour certaines écoles est la cause principale des données manquantes. Cette variable comporte environ 8% de données manquantes.

En conséquence, un travail d'imputation s'imposait avant de pouvoir effectuer le calage basé sur ces quatre variables. Cependant, une comparaison a d'abord été effectuée entre la distribution pondérée de l'ensemble des répondants au volet 2010 et la distribution de la population visée (pour ces quatre 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 la pondération a bien effectué la correction de la non-réponse jusqu'à maintenant pour diminuer le biais pour les estimations faites à partir des mesures de l'enquête. Le calage ne serait donc pas nécessaire dans ce cas et l'imputation ne le serait pas davantage.

Cette comparaison a permis de conclure que le calage ne serait effectivement pas nécessaire afin d'ajuster le poids transversal général du volet 2010. Le travail de pondération décrit à la section 3.2.3, ainsi qu'aux volets précédents, a permis de bien effectuer la réduction du biais de non-réponse quant aux mesures de réussite scolaire considérées. Bien que les valeurs manquantes aient été exclues de l'analyse, les conclusions sont les mêmes que celles énoncées dans l'analyse des données administratives du MELS faite en 2009 (Fontaine, Belleau et Courtemanche, 2009).