

PONDÉRATION DES DONNÉES DU VOLET 2006

Catherine Fontaine, Nathalie Plante et Robert Courtemanche
Direction de la méthodologie, de la démographie et des enquêtes spéciales
Institut de la statistique du Québec
Août 2007

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 (ELDEQ) au volet 2006. 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 ayant donné lieu à chacune des pondérations. La section 4 renseigne l'utilisateur sur le fichier de pondération ainsi que sur la façon d'utiliser les poids échantillonnaires dans les analyses statistiques. 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.**

1 Stratégie de pondération :

1.1 Admissibilité à l'enquête au volet 2006

Parmi les 2 120 répondants au volet initial, on compte 26 familles ayant quitté définitivement le Québec et deux familles dont l'enfant cible est décédé entre les volets 1998 et 2006. Ces familles ne sont plus admissibles à 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 2006 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 le calcul de la pondération. Sur cette base, l'échantillon admissible à l'enquête au volet 2006 est composé de 2 092 familles. Leur répartition, selon la participation à l'enquête à chacun des volets de 1998 à 2006, est présentée au tableau I.

Tableau I - Nombre de participants¹, volets de 1998 à 2006

Volet 1998	Volet 1999	Volet 2000	Volet 2001	Volet 2002	Volet 2003	Volet 2004	Volet 2005	Volet 2006	Nombre de participants				
Oui	Répondants 4 volets				Oui	Oui	Oui	Oui	1 286				
Oui					Oui	Oui	Oui	Non	69				
Oui					Oui	Oui	Non	Oui	33				
Oui					Oui	Oui	Non	Non	72				
Oui					Oui	Non	Oui	Oui	64				
Oui					Oui	Non	Oui	Non	11				
Oui					Oui	Non	Non	Oui	28				
Oui					Oui	Non	Non	Non	158				
Oui					Non	Oui	Oui	Oui	20				
Oui					Non	Oui	Oui	Non	3				
Oui					Non	Oui	Non	Oui	2				
Oui					Non	Oui	Non	Non	5				
Oui					Non	Non	Oui	Oui	22				
Oui					Non	Non	Oui	Non	3				
Oui					Non	Non	Non	Oui	14				
Oui					Non	Non	Non	Non	99				
Oui					Répondants à 2 ou 3 volets				Oui	Oui	Oui	Oui	24
Oui									Oui	Oui	Oui	Non	3
Oui									Oui	Oui	Non	Oui	1
Oui	Oui	Oui	Non	Non					2				
Oui	Oui	Non	Oui	Oui					8				
Oui	Oui	Non	Oui	Non					3				
Oui	Oui	Non	Non	Non					8				
Oui	Non	Oui	Oui	Oui					4				
Oui	Non	Oui	Oui	Non					1				
Oui	Non	Oui	Non	Non					2				
Oui	Non	Non	Oui	Oui					14				
Oui	Non	Non	Oui	Non					1				
Oui	Non	Non	Non	Oui					5				
Oui	Non	Non	Non	Non					37				
Oui	Répondant à 1 volet				Non	Non	Non	Non	34				
Oui					Oui	Non	Non	Non	2				
Oui					Oui	Non	Oui	Oui	1				
Oui	Répondant à 0 volet				Non	Non	Non	Non	52				
Oui					Oui	Non	Non	Non	1				
Nombre total d'enfants admissibles à l'enquête au volet 2006									2 092				

1.2 Participation au volet 2006

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,

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

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 2006, 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 2006 portant sur l'ensemble des enfants ayant participé 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 2006 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 2006 portant sur l'ensemble des enfants ayant participé à chacun des volets de l'enquête, soit de 1998 à 2006 (avec peu de données manquantes).

Concrètement, les situations 1 et 2 requièrent une pondération transversale unique, puisque la non-participation aux tests cognitifs, parmi les participants à l'enquête au volet 2006, 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 2006 mais qui étaient non participantes à l'un des volets de 1999 à 2005 est relativement important (240 enfants sur 1 526², soit 16 %), 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 2006, il a été décidé de créer un poids qui refléterait la participation à 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 années antérieures. 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 2006, l'écart entre le nombre de répondants au QIRI et le nombre de participants à ce volet est important (voir tableau II).

² Au volet 2006, le nombre de familles qui ont complété au moins un instrument était de 1 528, mais deux enfants sont exclus du calcul de pondération puisqu'ils n'ont complété qu'un ou deux tests cognitifs.

Tableau II- Nombre de répondants au QIRI et nombre de participants à au moins un instrument, volets de 1998 à 2006³

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

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

Le choix de la stratégie de pondération s'appuie sur une comparaison graphique d'estimations (Statistique Canada, 2006) selon les volets de 1998 à 2005. L'analyse de ces graphiques a permis de choisir le volet 2002 comme année de référence⁴ pour le calcul de la pondération transversale du volet 2006, plutôt que le volet 2005. 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 2005.

Par ailleurs, en 2005, une stratégie de fidélisation de l'échantillon a été mise en place afin d'en minimiser l'érosion et d'accroître le nombre de répondants relativement aux volets précédents. Si le volet 2005 avait été retenu comme année de référence, il aurait fallu faire une imputation importante pour les enfants qui avaient interrompu leur participation entre les volets 2002 et 2005 et qui participaient à nouveau en 2006. De plus, l'analyse graphique laisse croire que des répondants du volet 2005 qui ont fait un retour à l'ÉLDEQ, grâce aux efforts de fidélisation, présentent des caractéristiques pour lesquelles aucune mesure n'a été prise, ce qui peut introduire des biais dans les estimations. Le choix du volet 2002 comme année de référence à la pondération transversale du volet 2006 s'avérait donc le plus logique, d'autant plus qu'il respecte l'idée de modélisation du phénomène d'érosion de l'échantillon qui est préconisée depuis le début de l'enquête.

Un poids transversal de participation a ainsi été créé pour les 1 451 répondants au QIRI, de même que pour 75 enfants additionnels ayant répondu à au moins un autre instrument de collecte au volet 2006, soit un total de 1 526 enfants. La méthode de pondération sera décrite plus en détail à la section 3.

³ Quelques unités peuvent avoir été exclues du tableau dans le cas où très peu de données avaient été recueillies pour celles-ci lors d'un volet donné.

⁴ L'année de référence fournit la pondération de base qui fera l'objet d'un ajustement pour la non-réponse survenue ultérieurement.

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-participation à l'un ou l'autre des volets de 2003 à 2006. L'année de référence pour la pondération longitudinale a été déterminée selon la même méthode que pour la pondération transversale.

La modélisation de la non-participation au volet 2006 comporte donc, dans les deux cas, 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 les répondants du volet 1998 toujours admissibles à l'enquête au volet 2000 → 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 2002 → pondération QIRI du volet 2002
- Ajustement des poids transversaux du volet 2002 pour la non-participation à l'enquête au volet 2006 parmi les répondants du volet 2002 toujours admissibles à l'enquête au volet 2006 → pondération générale transversale du volet 2006

ou

Ajustement des poids transversaux du volet 2002 pour la non-participation à l'enquête aux volets 2003, 2004, 2005 ou 2006 parmi les répondants du volet 2002 toujours admissibles à l'enquête au volet 2006 → pondération générale longitudinale pour les volets de 1998 à 2006.

Afin d'obtenir une pondération transversale pour l'ensemble des 1 526 participants du volet 2006, les enfants qui étaient participants au volet 2006 mais non participants au volet 2002 ont été considérés répondants au volet 2002 et se sont vu attribuer un poids QIRI pour le volet 2002⁵, 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 1 526 enfants ayant participé à l'enquête au volet 2006. Cette pondération peut également être utilisée pour une analyse de variables où une petite proportion d'enfants présenterait des valeurs manquantes⁶.

Les poids longitudinaux ont, quant à eux, été obtenus par un ajustement des poids QIRI du volet 2002 et doivent être utilisés pour l'analyse des variables qui prennent une valeur pour l'ensemble des 1 286 enfants ayant participé à l'enquête à chacun des volets de 1998 à 2006. Encore une fois, la présence de quelques données manquantes pour les variables d'analyse n'invalide pas la pondération.

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

⁶ 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

Aucune pondération pour un instrument particulier n'a été effectuée au volet 2006. 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 totale 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 participants est illustrée au schéma I.

Tableau III - Nombre de répondants par instrument au volet 2006

	Nb répondants	Proportion pondérée de répondants parmi les participants au volet 2006 (%)
QIRI	1 451	94,8
Mini-QPRI ⁷	45	n/a
QAAM ⁸	1 270	81,6
QPAE ⁹	1 473	96,6
QAAENS ¹⁰	1 291	84,0
Tests cognitifs ¹¹	1 479	97,0
ÉCP ¹²	1 393	90,6
Test de motricité	1 203	77,0

⁷Court questionnaire papier rempli par l'intervieweur pour les familles ayant refusé de compléter le QIRI intégral

⁸ Questionnaire autoadministré de la mère/conjointe

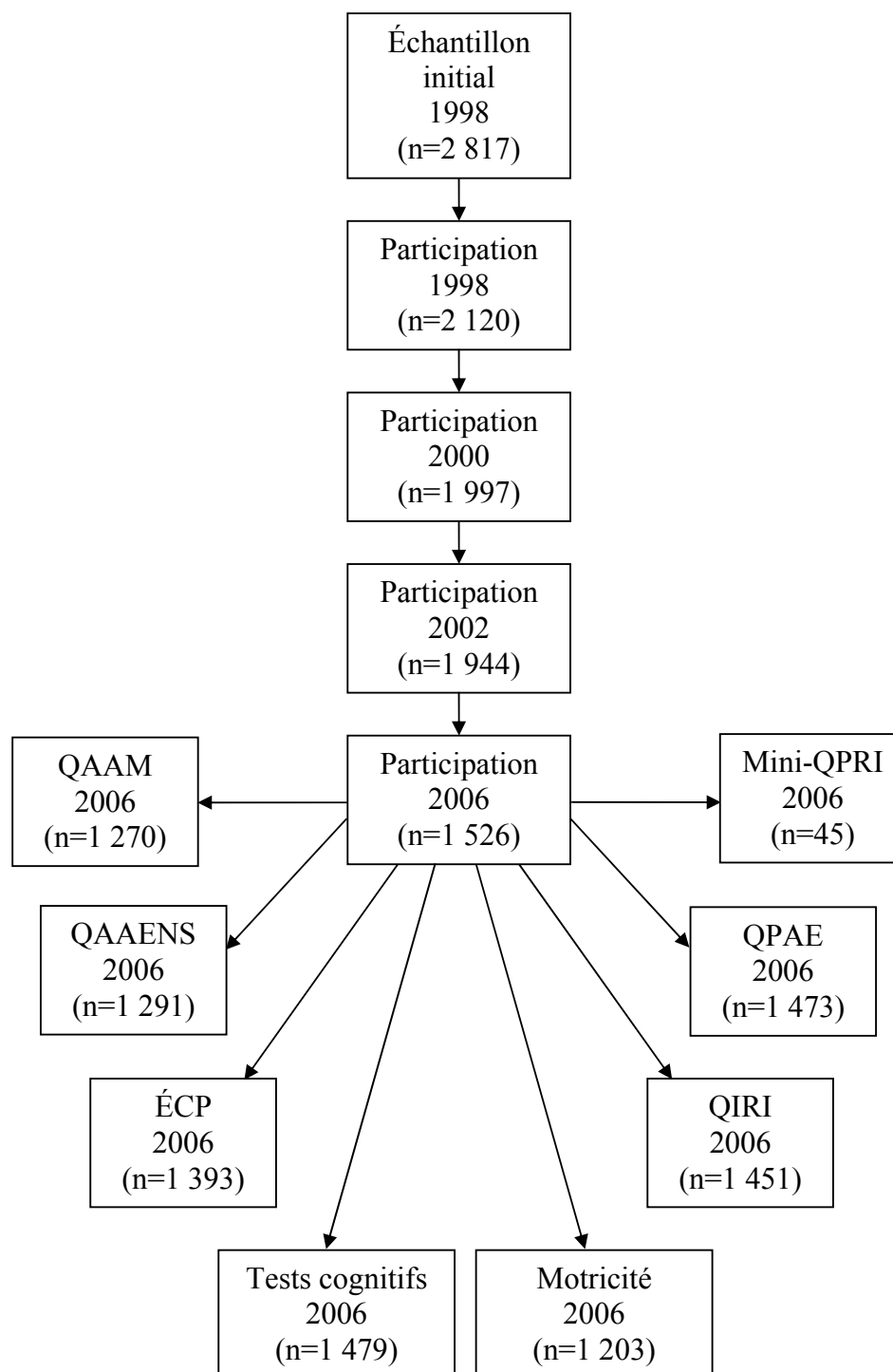
⁹ Questionnaire papier administré à l'enfant

¹⁰ Questionnaire autoadministré de l'enseignant

¹¹ Enfants ayant participé à au moins un des tests cognitifs (K-ABC, NEPSY, RAN, CAT, orthographe).

¹² Évaluation de la condition physique

Schéma I. Hiérarchie des participants aux différents instruments de collecte



Note: Ce tableau n'inclut pas les 2 enfants qui ont effectué seulement un ou deux tests cognitifs au volet 2006, et pour lesquels aucun autre instrument n'a été complété.

2 Taux de réponse :

Le tableau IV présente le taux de réponse pondéré transversal obtenu au volet 2006. 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 chacun des instruments sont présentés à la section 3.

Tableau IV - Taux de réponse pondéré transversal pour la participation au volet 2006

Taux de réponse au volet 1998	75,3% (n=2 817)
Proportion de participants au volet 2000 parmi les participants au volet 1998 admissibles au volet 2006 (incluant les nouveaux répondants ¹³)	95,0 % (n=2 092)
Proportion de participants au volet 2002 parmi les participants au volet 2000 (incluant les nouveaux répondants)	98,6 % (n=2 005)
Proportion de participants au volet 2006 parmi les participants au volet 2002 (incluant les nouveaux répondants)	75,8 % (n=1 978)
Taux de réponse transversal pour la participation au volet 2006	53,5 %

Le tableau V présente le taux de réponse pondéré longitudinal pour la participation au volet 2006. Ces taux ont également été calculés en concordance avec la démarche de pondération.

Tableau V - Taux de réponse pondéré longitudinal pour la participation aux volets de 1998 à 2006

Taux de réponse au volet 1998	75,3% (n=2 817)
Proportion de participants aux volets 1998 à 2000 parmi les participants au volet 1998 admissibles au volet 2006	93,6 % (n=2 092)
Proportion de participants aux volets 1998 à 2002 parmi les participants aux volets 1998 à 2000, admissibles au volet 2006	95,0 % (n=1 977)
Proportion de participants aux volets 1998 à 2006 parmi les participants aux volets 1998 à 2002, admissibles au volet 2006	66,0 % (n=1 890)
Taux de réponse longitudinal pour la participation aux volets 1998 à 2006	44,2 %

¹³ 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 2006, on redéfinit ici certains enfants répondants au volet 2000 alors qu'ils ne l'étaient pas en réalité (voir section 1.3).

3 Analyse de la non-réponse :

3.1 Démarche générale d'analyse

La création de pondérations ajustées pour la non-réponse aux différents instruments est basée sur la formation de classes de pondération. Comme aux volets précédents, la technique employée est la modélisation par segmentation fondée sur l'algorithme CHAID (« Chi-square automatic interaction detection ») mis au point par Kass (1980). 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 observée au sein de la classe à laquelle elle appartient.

3.2 Pondération transversale des données du volet 2006

3.2.1 Enfants participants au volet 2006 mais non participants à au moins un volet de 2002 à 2005

Aux 1 940 répondants du volet 2002 (admissibles au volet 2006) était déjà associé un poids QIRI du volet 2002. Cependant, 32 enfants dont la famille a participé au volet 2006 n'étaient pas participants au volet 2002 et n'ont de ce fait aucun poids de référence. Aux fins de la pondération transversale du volet 2006, ces enfants ont été considérés participants au volet 2002, de manière à recalculer un nouveau poids pour l'ensemble des participants au volet 2002, incluant ces derniers¹⁴.

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 2006, 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 2006 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 (PEGENT9) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 526 enfants ayant participé à l'enquête au volet 2006.

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

¹⁴ Six enfants qui n'ont participé ni au volet 2002, ni au volet 2006, mais qui ont participé à au moins un volet de 2003 à 2005 ont également été considérés participants au volet 2002, de manière à obtenir un poids transversal au volet 2006 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 participer à l'enquête au volet 2002.

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 (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 (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).

Parmi l'ensemble des variables considérées pour la modélisation, voici celles qui ont été retenues:

- le statut d'immigrant de la mère (ESDMD1A)
- le plus haut niveau de scolarité de la mère/conjointe (EEDMD01)
- le nombre de personnes dans le ménage (EREFD01)
- le statut de l'union actuel (EDEFD4A)
- le nombre de frères/sœurs de l'enfant cible (EREED01)
- le revenu du ménage (variable longitudinale - seuil de 10 000 \$)

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondants varie de 47 % à 92 %, comparativement à 76 % globalement. La proportion la plus faible est observée dans la classe des mères immigrantes dont l'enfant cible n'a aucun frère ou sœur.

3.3 Pondération longitudinale des données des volets de 1998 à 2006

La non-réponse aux volets de 2003 à 2006¹⁵, parmi les répondants longitudinaux du volet 2002¹⁶, a été modélisée à l'aide des mêmes variables que celles étudiées pour la pondération transversale. Les variables suivantes ont été retenues pour la création des classes de pondération :

- le plus haut niveau de scolarité de la mère/conjointe (EEDMD01)
- le statut d'immigrant de la mère/conjointe (ESDMD1A)
- le revenu du ménage (variable longitudinale - seuil de 10 000 \$)
- le nombre de frères/sœurs de l'enfant cible (EREED01)

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondants varie de 29 % à 86 %, comparativement à 66 % globalement. La proportion la plus faible est observée dans la classe formée des mères immigrantes dont le plus haut niveau de scolarité est un diplôme d'études secondaires ou moins.

La nouvelle variable de pondération longitudinale (PEGENL9) est appropriée pour l'analyse des variables qui prennent une valeur pour la presque totalité des 1 286 enfants ayant participé à l'enquête à chacun des volets de 1998 à 2006.

¹⁵ Cette non-réponse est ici définie comme la non-participation à au moins un volet de 2003 à 2006.

¹⁶ Seuls sont ici considérés les enfants répondants du volet 2002 qui étaient toujours admissibles à l'enquête au volet 2006.

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

4.1 Fichier de pondération

Le fichier SAS POIDS901 contient les variables de pondération suivantes: PEGENT9 (poids général transversal 2006) et PEGENL9 (poids général longitudinal 1998-2006). 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.

Le fichier de pondération (POIDS901) contient également une variable (INAPTE_E9) permettant d'identifier les enfants définis comme étant « inaptes », soit ceux souffrant de problèmes de santé chroniques diagnostiqués tels que l'autisme ou l'incapacité mentale (sur la base des troubles déclarés par la PCM à l'un ou l'autre des volets de 2001 à 2006). Les enfants « inaptes » peuvent être considérés répondants aux fins de l'analyse des données des tests cognitifs et de motricité, qu'ils aient complété une tâche ou non; il est alors suggéré d'imputer les scores les plus faibles (théoriques ou rencontrés) aux essais pour lesquels les données sont manquantes. Autrement, ces enfants pourraient également être exclus de certaines analyses portant sur les tests cognitifs ou de motricité et pour lesquelles l'inférence est faite à la sous-population des enfants qui ne présentent pas de telle incapacité chronique.

4.2 Tests statistiques

Les poids contenus dans le fichier POIDS901 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 2006 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 totale, 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 en annexe.

5 Références bibliographiques :

KASS, G. V. (1980). « An exploratory technique for investigating large quantities of categorical data », *Applied Statistics*, 29, p.119-127.

STATISTIQUE CANADA. (2006). The National Longitudinal Survey of Children : *10-Year Review of the Methodology*. Sarah Franklin, ed. Household Survey Methods Division. Document interne à Statistique Canada.

ANNEXE

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 chacun des scores aux tests cognitifs (K-ABC, NEPSY, RAN, CAT2/mod et orthographe) et les troubles d'hyperactivité ou d'inattention de l'enfant, tels qu'évalués par la personne qui connaît le mieux l'enfant au volet 2006 (variable dérivée IBEET01O). Le nombre d'enfants pour lesquels on a une telle mesure d'hyperactivité ou d'inattention, en plus d'un score, varie, selon le test cognitif considéré, de 1 238 à 1 395.

Puisque l'analyse considérée porte sur des variables du volet 2006 uniquement, il s'agit d'évaluer si la pondération générale transversale du volet 2006 PEGENT9 convient. Cette pondération prend une valeur pour les 1 526 enfants ayant participé au volet 2006; toutes les unités d'analyse considérées ont donc un tel poids transversal. Il n'y aurait donc aucune perte d'unités liée à l'utilisation de cette pondération. Toutefois, il demeure une certaine part de la non-réponse pour laquelle aucun ajustement n'a été fait, soit de 9 % $((1526-1395)/1526)$ à 19 % $((1526-1238) / 1526)$ selon le test¹⁷. Plus précisément, pour tous les tests sauf le « NEPSY - pseudo-mots », cette proportion ne surpasse pas 10 %.

Ainsi, pour l'analyse des données du test des « NEPSY - pseudo-mots », il serait recommandé de faire un ajustement sommaire pour la non-réponse à partir de la pondération PEGENT9. L'ISQ pourrait créer un poids sur mesure pour une telle analyse.

Les scores aux autres tests cognitifs du volet 2006 pourraient être analysés à l'aide de la pondération PEGENT9, en s'assurant au préalable que parmi les enfants ayant un tel poids, les répondants ne diffèrent pas trop des non-répondants. Pour les 1 526 enfants en cause, il s'agit dans ce cas de faire une étude de biais en croisant, par exemple, le fait d'être répondant ou non (c'est-à-dire de prendre une valeur pour les variables d'analyse) avec certaines variables socioéconomiques importantes qui pourraient être

¹⁷ Les proportions pondérées correspondantes vont également de 9 % à 19 %.

liées au sujet à l'étude, comme la langue parlée à la maison par la mère, le niveau de suffisance de revenu du ménage et le type de famille au volet 2006.

L'analyse de biais a dans ce cas-ci révélé que la non-réponse est significativement plus importante chez les enfants des familles dont le revenu est très insuffisant ou dont la mère n'a pas travaillé au cours des 12 mois ayant précédé l'enquête. Les écarts observés demeurent modestes. Aucun écart significatif n'a pu être détecté entre les enfants selon le type de famille, le statut d'immigrant ou le niveau de scolarité de la mère, ou encore la langue parlée à la maison par celle-ci. Pour cette analyse, il serait acceptable d'avoir recours à la pondération PEGENT9 existante, tout en discutant des possibilités d'un léger biais. En supposant, par exemple, qu'un niveau de suffisance de revenu plus faible et une mère ne travaillant pas sont davantage associés à un score cognitif plus faible et que les troubles d'hyperactivité et d'inattention ne sont pas liées à ces variables socioéconomiques, il est possible que l'utilisation des poids PEGENT9 entraîne une légère sous-estimation de la force de la relation entre les scores et les troubles d'hyperactivité ou d'inattention.

Exemple 2

Supposons qu'on veuille mettre en relation la réussite scolaire de l'enfant de deuxième année (volet 2006), telle qu'évaluée par l'enseignant (IAEIQ04), et les résultats au test de préparation à l'école « LOLLIPOP » (scores GLOEJEU1 à GLOEJEU4) administré à l'enfant à la maternelle (volet 2004). On compte 959 enfants ayant réalisé le test du « LOLLIPOP » (4 scores) et pour lesquels la réussite scolaire a été évaluée.

Sachant qu'il y a eu une légère érosion de l'échantillon des familles participantes entre les volets 2004 et 2006 et qu'une proportion non négligeable d'enfants des familles participantes à l'enquête au volet 2004 n'ont pas complété les tests cognitifs, les pondérations potentielles pour cette analyse seraient les suivantes :

1. PEGENT9 - pondération générale transversale du volet 2006 (n=1 526)
2. PEGENL9 - pondération générale longitudinale du volet 2006 (n=1 286)
3. PEJEUX7 - pondération des tests cognitifs du volet 2004 (n= 1 209)

Les pondérations 1 et 3 présentent un poids pour chacun des 959 enfants, alors que la pondération 2 vaut seulement pour 914 des 959 enfants. L'utilisation de la pondération 2 entraînerait par conséquent une perte d'environ 5 % des unités d'analyse. Par ailleurs, la proportion de la non-réponse pour laquelle aucun ajustement n'a été fait est, respectivement, de 37 % $((1526-959) / 1526)$, 29 % $((1286-914) / 1286)$ et 21 % $((1209-959) / 1209)$ pour les pondérations 1 à 3¹⁸. C'est donc dire qu'aucune des pondérations disponibles ne peut être utilisée telle quelle pour faire inférence à la population visée sans risque de biais dû à la non-réponse. Dans ce cas-ci, il serait recommandé de faire un ajustement sommaire pour la non-réponse à partir de la pondération PEJEUX7 pour laquelle la proportion de non-réponse non prise en compte est la plus faible, sans perte d'unités d'analyse. L'ISQ pourrait créer un poids sur mesure pour une telle analyse.

¹⁸ Ces proportions pondérées sont de 39 %, 30 % et 22 %.