Pondération et analyse de la non-réponse des données du volet 2017

Luc Belleau et Karine Dion
Direction de la méthodologie et de la qualité
Institut de la statistique du Québec
16 août 2018

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 2017[1]. Dans ce rapport, le lecteur est invité à consulter les sections 1 à 4 pour connaître les étapes ayant mené au calcul de la pondération et la section 5 pour savoir comment utiliser la pondération. Plus précisément,

- la section 1 propose une description des étapes ayant mené au choix de la stratégie de pondération;
- la section 2 présente l'analyse de la non-réponse totale ayant donné lieu à la création de la pondération;
- la section 3 présente l'analyse de la non-réponse partielle;
- la section 4 détaille le taux de réponse pondéré obtenu; et
- la section 5 renseigne l'utilisateur sur le fichier de pondération ainsi que sur la façon d'utiliser les poids échantillonnaux dans les analyses statistiques. Elle comporte également quelques mises en garde concernant l'utilisation de ces poids.

1. Stratégie de pondération

1.1 Admissibilité à l'enquête au volet 2017

Parmi les 2 120 répondants au volet initial, on compte 31 familles ayant quitté définitivement le Québec et 3 familles dont l'enfant cible est décédé entre les volets 1998 et 2017. Les enfants cibles de ces familles, considérés comme inadmissibles à l'enquête, ne sont plus visés par l'enquête en ce sens qu'ils ne font plus partie de la population sur laquelle porte l'inférence. La population visée est par conséquent composée des enfants survivants qui sont demeurés au Québec entre les volets 1998 et 2017 ou qui ont quitté la province temporairement.

Les enfants cibles des familles n'ayant pu être retracés, ayant refusé de répondre ou ayant été dans l'impossibilité de le faire sont tous considérés comme admissibles à l'enquête. Bien que l'on sache que, parmi les familles n'ayant pu être retracées, certaines pourraient avoir déménagé définitivement hors du Québec, leur nombre est trop petit pour que l'on en tienne compte dans le calcul de la pondération. Sur cette base, l'échantillon admissible à l'enquête au

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

volet 2017 est composé de 2 086 jeunes[2]. Leur répartition, selon la réponse à l'enquête à chacun des volets de 1998 à 2017, est présentée au tableau I.

Tableau I Nombre de répondants[3] aux volets de 1998 à 2017.

Volets 1998 à 2002	Volets 2003 à 2015	Volet 2017	Nombre de répondants
Répondants aux 5 volets	Répondants aux 9 volets	Oui	839
		Non	78
	Répondants à 6, 7 ou 8 volets	Oui	349
		Non	105
	Répondants à 3, 4 ou 5 volets	Oui	117
		Non	127
	Répondants à 1 ou 2 volets	Oui	39
		Non	160
	Répondant à 0 volet	Oui	5
		Non	65
Répondants à 3 ou 4 volets	Répondants aux 9 volets	Oui	9
		Non	1
	Répondants à 6, 7 ou 8 volets	Oui	21
		Non	10
	Répondants à 3, 4 ou 5 volets	Oui	12
		Non	10
	Répondants à 1 ou 2 volets	Oui	9
		Non	16
	Répondant à 0 volet	Oui	0
		Non	24
Répondants à 2 volets	Répondants à 3, 4 ou 5 volets	Oui	0
		Non	1
	Répondant à 1 ou 2 volets	Oui	0
		Non	2
	Répondant à 0 volet	Oui	0
		Non	34
Répondants à 1 volet	Répondant à 0 volet	Oui	0
		Non	53
Nombre total de jeunes admissibles à l'enquête au volet 2017			2 086

Note : Il n'y a pas eu de collecte de données en 2007, 2009, 2012, 2014 et 2016.

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

[3] Aux volets 1998 à 2002, les répondants ont tous rempli le QIRI; de 2003 à 2015, les répondants ont rempli au moins un instrument de collecte au volet concerné; à partir de 2017, seul le questionnaire en ligne du jeune était à remplir.

1.2 Répondants au volet 2017

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 2017, une seule pondération principale a été créée. Celle-ci permet l'analyse des variables du volet 2017 portant sur l'ensemble des jeunes ayant répondu à ce volet d'enquête (avec peu de données manquantes pour ces variables ou des variables d'autres volets incluses dans l'analyse).

Contrairement aux volets précédents, il n'y a qu'un seul instrument de collecte, soit le questionnaire en ligne au jeune (QELJ). Il a été décidé de créer un poids qui refléterait donc le fait d'avoir complété cet instrument de collecte au volet 2017. Rappelons que pour les volets 2006 à 2015, la pondération devait refléter le fait d'avoir complété au moins un instrument de collecte, tandis que pour les volets 1998 à 2005, ce fut un poids spécifique au Questionnaire informatisé rempli par l'intervieweur (QIRI) qui a été construit. La stratégie de collecte est donc complètement axée sur le jeune au volet 2017. Le tableau II présente le nombre de répondants par volet.

Tableau II Nombre de répondants[4] à certains volets de 1998 à 2017

Nb de répondants	volet 1998	volet 1999	volet 2000	volet 2001	volet 2002	volet 2003	volet 2004	volet 2005	volet 2006	volet 2008	volet 2010	volet 2011	volet 2013	volet 2015	volet 2017
au QIRI pour un volet donné	2 120	2 045	1 997	1 950	1 944	1 759	1 492	1 528	1 451	1 334	1 396	1 290	1 400	1 252	-
au QELJ pour un volet donné	-	-	-	-	-	-	-	-	-	-	-	-	1 446	1 270	1 400
pour un volet donné	2 120	2 045	1 997	1 950	1 944	1 775	1 529	1 537	1 526	1 402	1 415	1 312	1 466	1 348	1 400
longitudinaux (pour un volet donné et ses précédents)	2 120	2 045	1 985	1 924	1 894	1 723	1 462	1 355	1 286	1 186	1 121	1 035	991	917	839

[4] Aux volets 1998 à 2002, les répondants ont tous rempli le QIRI; de 2003 à 2015, les répondants ont rempli au moins un instrument de collecte au volet concerné; à partir de 2017, seul le questionnaire en ligne du jeune était à remplir.

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

Un nouveau poids de base, soit un « poids enfant », a été créé afin de répondre aux besoins futurs de la phase 4 de l'ELDEQ. Pour ce faire, nous avons attribué un poids de base à tous les jeunes ayant répondu à au moins un volet entre 2011 et 2018, incluant le volet spécial[5] sur la santé mentale. Ainsi, nous avons un poids de base qui servira comme volet de référence pour les prochaines pondérations transversales des questionnaires sur le jeune. Cette façon de faire permet de considérer des variables mesurées directement auprès du jeune, tout en considérant également les informations stables de 2002 qui ont été utilisées pour les volets précédents dans le modèle de la participation (ou non).

1.3.1 Ajustement de la non-réponse au niveau transversal

Un poids transversal général a ainsi été créé pour les 1 400 répondants au QELJ du volet 2017. La méthode de pondération sera décrite plus en détail à la section 2.

La modélisation de la non-réponse au volet 2017 comporte cinq étapes :

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

2. Ajustement des poids QIRI du volet 1998 pour la non-réponse à l'enquête au volet 2000 parmi les répondants du volet 1998 toujours admissible à l'enquête au volet 2017 → pondération QIRI du volet 2000.

3. Ajustement des poids transversaux du volet 2000 pour la non-réponse à l'enquête au volet 2002 parmi les répondants du volet 2000 toujours admissibles à l'enquête au volet 2017 → pondération QIRI du volet 2002.

4. Ajustement des poids transversaux du volet 2002 pour la non-réponse à tous les volets entre 2011 et 2018, incluant le volet spécial sur la santé mentale, parmi les répondants du volet 2002 toujours admissibles à l'enquête au volet 2017 → pondération « enfant ».

5. Ajustement des poids « enfant » pour la non-réponse à l'enquête au volet 2017 parmi les répondants admissible à l'enquête au volet 2017→ pondération générale transversale du volet 2017.

[5] Un volet spécial est une collecte de renseignements, généralement issue d'un projet de chercheur, auprès d'une partie ou de l'ensemble de l'échantillon, non prévue dans les volets réguliers. En 2018, un volet spécial fut mené auprès des jeunes de l'échantillon et portait sur la santé mentale.

2. Analyse de la non-réponse totale

2.1 Pondération transversale des données du volet 2017

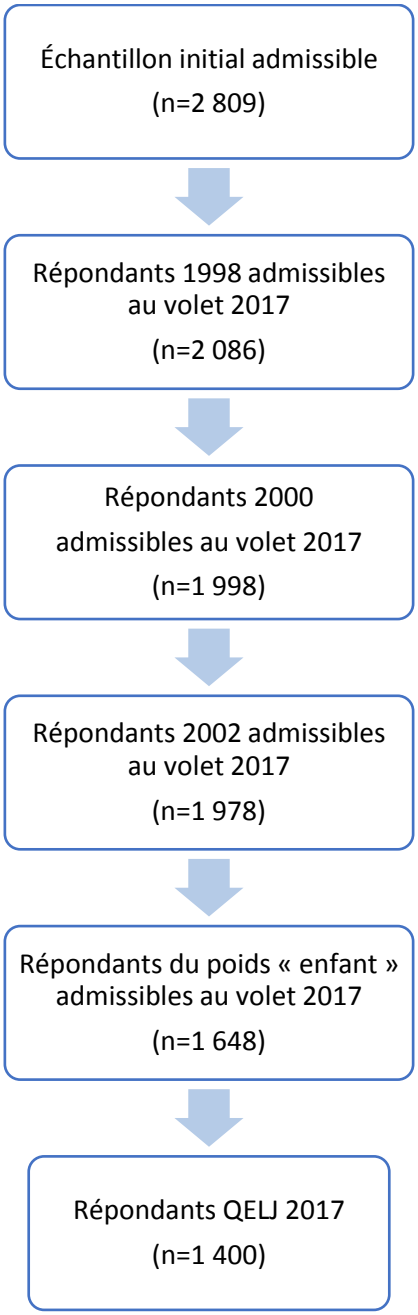
2.1.1 Démarche générale d'analyse

La création de pondérations ajustées pour la non-réponse est basée sur la formation de classes de pondération. C'est la méthode du score qui a été utilisée pour créer les classes de pondération (pour plus de détails sur la méthode, voir Haziza et Beaumont, 2007 et Eltinge et Yansaneh, 1997). Cette méthode crée des groupes homogènes selon la valeur d'un score, celui-ci étant issu d'un modèle de régression logistique. C'est la réponse au volet 2017 qui a été analysée à l'aide de ce modèle et la probabilité estimée de réponse constitue le score. Par la suite, la création des groupes s'effectue à l'aide d'une méthode de classification. Enfin, pour un jeune donné, l'ajustement de la pondération consiste à diviser le poids de référence par la proportion pondérée de jeunes répondants observée au sein du groupe auquel il appartient. Pour plus de détails concernant cette démarche d'analyse, consulter l'annexe A.

Pour tenir compte de la non-réponse au volet 2017, un ajustement a été fait à partir du poids de base « enfant » (voir la section 1.3). Cet ajustement est requis puisque les répondants au volet 2017 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 hiérarchie des répondants admissibles au volet 2017 est illustrée au schéma I.

Schéma I - Hiérarchie des répondants aux différents instruments de collecte, admissibles au volet 2017



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

2.1.2 Variables considérées et résultats

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

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

Compte tenu de la disponibilité de microdonnées administratives[6] provenant du ministère de l'Éducation et de l'Enseignement supérieur (MEES), des variables supplémentaires ont été considérées pour la pondération du volet 2017. Ces variables sont reliées au cheminement scolaire du jeune : existence d'un plan d'intervention actif, sexe, niveau scolaire, langue d'enseignement, langue maternelle, type de classe spéciale (s'il y a lieu), code de difficulté, secteur d'enseignement et diplomation. Ce fichier permet aussi de créer une variable qui exprime la mobilité du jeune entre 2015 et 2017. Cette variable a été créée en comparant les adresses de la résidence du jeune pour les années 2015 et 2017. Il est important de mentionner que la majorité des jeunes de ce volet avaient terminé leurs secondaires en 2015 ou 2016, ainsi il y avait très peu d'informations disponibles provenant du MEES. Pour certaines de ces variables, une imputation avec les microdonnées d'années antérieures était possible, telles que les variables sur la diplomation ou l'obtention du DES.

Également, des variables du questionnaire au jeune de 2013 et 2015 ont été considérées pour la pondération du volet 2017. Ces variables sont principalement liées à la motivation scolaire du jeune : aime l'école, a déjà doublé une année scolaire, plus haut niveau de scolarité qu'il désire atteindre, s'il travaille ou non, quel niveau scolaire est-il inscrit. Ces nouvelles variables permettent d'avoir une information directement mesurée auprès du jeune, et pouvant être liées ou non à la participation de celui-ci au volet 2017.

Parmi l'ensemble des variables considérées, voici celles qui ont été retenues pour le modèle

[6] Microdonnées provenant du fichier daté du mois d'avril 2018 pour les jeunes inscrits à l'école au 30 septembre 2016.

final de régression logistique :

- le plus haut niveau de scolarité de la mère/conjointe en 2002 (EEDMD01)
- le sexe du jeune (sexe)
- la langue d'enseignement en 2017 complété avec la langue d'enseignement de 2015 ou 2013, si nécessaire (lan_enseignement)
- l'obtention ou non de son diplôme d'études secondaires (DES).

Une méthode de classification hiérarchique de Ward a permis de regrouper les probabilités estimées en 5 classes de pondération. Le tableau III présente les proportions pondérées de répondants au volet 2017 parmi les répondants du poids « enfant » pour ces 5 groupes. De plus, il présente le nombre de répondants, parmi les 1 400 répondants, à qui la proportion pondérée sera appliquée en guise de correction de la non-réponse. Par exemple : il y a 159 répondants au volet 2017 dont le poids de référence sera ajusté par l'inverse de la proportion pondérée de la quatrième classe de pondération, qui est de 77,9 %.

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

Classe de pondération	Proportions pondérées de répondants au volet 2017 (en %)	Nombre de répondants
1	91,6	565
2	93,3	237
3	81,8	425
4	77,9	159
5	31,0	14

Au sein des différentes classes d'ajustement de la pondération, la proportion de répondants varie de 31 % à 93 % (relativement à une proportion globale de 84,8 %). La proportion la plus faible est observée dans une classe où l'on retrouve, en proportion, un plus grand nombre de jeunes de sexe masculin; un plus grand nombre de jeunes dont la langue d'enseignement au secondaire était le français; un moins grand nombre de jeunes ayant obtenu leur DES; et un moins grand nombre de jeunes dont la mère avait un diplôme universitaire.

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

Suite à une entente avec le ministère de l'Éducation et de l'Enseignement supérieur (MEES), l'Institut de la statistique du Québec (Institut) a pu obtenir de cet organisme des statistiques agrégées pour la population visée par l'ELDEQ (N=57 406)[7]. Ces statistiques agrégées sont en fait des totaux pour des caractéristiques choisies par l'Institut et disponibles au MEES. Par exemple : la répartition des 57 406 jeunes selon le secteur scolaire. Ces statistiques agrégées obtenues pour les données du volet 2017 permettent d'évaluer la pertinence d'effectuer un ajustement à la pondération transversale. Cet ajustement, appelé « calage », est défini comme un redressement des poids d'enquête afin que les estimations s'ajustent à des totaux connus (Lavallée et Durning, 1993). Ce redressement peut aussi être utilisé dans le but de pallier la non-réponse.

[7] L'ensemble des enfants nés au Québec entre le 1er octobre 1997 et le 20 septembre 1998 qui fréquente le système scolaire québécois au cours de l'année scolaire 2016-2017.

L'objectif du calage au volet 2017 est d'effectuer une correction supplémentaire pour diminuer le biais dû à la non-réponse, et ce, à l'aide de caractéristiques reliées aux mesures principales de l'enquête, c'est-à-dire des variables liées à la réussite scolaire. C'est dans cet esprit que les caractéristiques ont été choisies à partir de l'ensemble des variables administratives disponibles[8]. En effet, l'Institut reçoit annuellement, en plus des statistiques agrégées, un fichier de microdonnées administratives pour l'essentiel de l'échantillon de départ de la cohorte[9]. Les variables provenant des données administratives du MEES sont donc disponibles :

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

Les variables administratives considérées pour le calage sont :

- Sexe du jeune;
- Langue maternelle du jeune;
- Langue d'enseignement;
- Existence d'un plan d'intervention actif pour le jeune à l'école;
- Niveau scolaire du jeune (retard scolaire);
- Secteur scolaire du jeune;

Une comparaison a d'abord été effectuée entre la distribution pondérée de l'ensemble des répondants au volet 2017 et la distribution de la population visée (pour ces variables en excluant les valeurs manquantes). L'objectif était de vérifier si les proportions pondérées étaient près des proportions calculées pour la population. En effet, si l'écart est négligeable, cela signifie que l'ajustement de calage n'est pas nécessaire puisque le biais est faible.

C'est la variable qui exprime le secteur scolaire qui présente un écart le plus important entre les deux distributions. Cette variable identifie le(s) secteur(s) scolaire(s) auquel le jeune était inscrit au volet 2017. La proportion pondérée de jeunes dont le secteur scolaire était uniquement le collégial en 2017 est de 55,6 % dans l'échantillon comparativement à 67,2 % dans la population de l'ELDEQ. Le sexe du jeune a également été considéré dans l'ajustement des poids. En effet, il semble avoir un lien entre le sexe du jeune et le fait d'être inscrit ou non au collégial. Ainsi, l'ajustement des poids s'est effectué selon ces deux variables afin de rendre la distribution pondérée des répondants semblable à la distribution dans la population visée. Quelques variables du volet 2017 ont été choisies pour vérifier l'impact de cet ajustement apporté aux poids. Il est possible de conclure que les proportions pondérées calculées pour des caractéristiques généralement associées à une moins grande réussite scolaire ont légèrement augmenté suite à l'ajustement apporté aux poids[10]. Ce constat va dans le sens attendu, à savoir que l'échantillon de l'ELDEQ se rapproche de la population visée, cette dernière comprenant davantage d'élèves ayant des caractéristiques liées à une moins grande réussite scolaire.

[8] Pour limiter l'ampleur de la production de statistiques agrégées au MEES, un choix devait être fait.

[9] Fichier en date de février 2017 pour les jeunes inscrit à l'école au 30 septembre 2016.

[10] Par exemple, dans le QIRI du volet 2016 pour la variable qui identifie les jeunes avec un problème chronique de déficit d'attention (avec ou sans hyperactivité).

3. Analyse de la non-réponse partielle

3.1 Introduction

La pondération transversale a été produite afin de tenir compte de la non-réponse totale, mais celle-ci n'a pas été ajustée pour la non-réponse partielle à une question. Une non-réponse partielle importante peut entraîner certains biais dans les estimations, au même titre que la non-réponse totale. Cette section relève les questions (ou items) présentant un taux de non-réponse partielle supérieur à 5 %. Le taux de non-réponse partielle est défini ici comme le rapport entre le nombre pondéré de non-répondants à une question d'un instrument donnée et le nombre pondéré de répondants à cet instrument admissibles à y répondre.

Par ailleurs, la non-réponse partielle n'est considérée ici que pour chaque variable prise individuellement. Ainsi, lorsqu'une analyse implique plusieurs variables, il se peut que la non-réponse partielle cumulée soit plus élevée. Par conséquent, l'interprétation des résultats devrait tenir compte de cette non-réponse, par la création d'une pondération « sur mesure » ou en tentant d'évaluer dans quel sens irait le biais dû à la non-réponse partielle, s'il y a lieu.

À noter qu'avant de choisir une variable pour une analyse, il importe de considérer non seulement l'ampleur de la non-réponse partielle pour celle-ci, mais également la non-réponse totale à l'instrument considéré, s'il y a lieu[11].

3.2 Variables présentant une non-réponse partielle supérieure à 5%

L'analyse des caractéristiques des non-répondants partiels, c'est-à-dire les sous-populations pour lesquelles la non-réponse partielle est significativement plus élevée que celle des répondants, n'a pas été effectuée. C'est lors de la création des pondérations « sur mesure » que leurs caractéristiques seront analysées.

Le tableau IV présente les variables d'analyse pour lesquelles le taux de non-réponse partielle est supérieur à 5 %. Pour chacune des variables identifiées, le taux pondéré est présenté, ainsi que le nombre de personnes (non pondéré) qui avaient à répondre à la question correspondante[12]. Pour l'instrument de collecte du volet 2017 (QELJ), ce sont les questions qui touchent le travail et le sommeil qui sont davantage entachées de non-réponse partielle.

[11] Il est prévu que des pondérations « sur mesure » soient produites pour des analyses spécifiques lorsque la non-réponse cumulative est importante.

[12] Lorsque le nombre de personnes qui avaient à répondre à une question est inférieur à 100, seules les variables ayant un taux de non-réponse partielle supérieur à 10 % sont présentées. Par contre, lorsque le nombre de personnes qui avaient à répondre à une question est inférieur à 10, même si le taux de non-réponse partielle est supérieur à 10 %, les résultats ne sont pas présentés.

Tableau IV Questions ayant une non-réponse partielle, questionnaire en ligne du jeune (QELJ)

Fichier QELJ2001 (n=1 400)	Description	Taux de non-rép. partielle pondérée (%)
TTVNQ05M	dernier mois, nombre de minutes travaillé par semaine	57,1 (n=199)
TTVNQ13M	Emploi principal, nombre de minutes travaillé par semaine	37,9 (n=1011)
TTVNQ35BBA	Depuis août 2015, 2e emploi, depuis combien de temps (an)	7,0 (n=237)
TSONQ13H et TSONQ13M	Questions portant sur l'heure et les minutes où le jeune se couche le soir.	11,2 (n=74)
TSONQ14H et TSONQ14M	Questions portant sur l'heure et les minutes où le jeune se lève le matin	12,6 et 13,3 (n=74)
TSONQ15M	Question portant sur l'heure et les minutes où le jeune se couche le soir.	11,4 (n=74)
TSONQ16H et TSONQ16M	Questions portant sur l'heure et les minutes où le jeune se lève le matin	12,3 et 13,9 (n=74)

4. Taux de réponse

Le tableau V présente le taux de réponse pondéré transversal obtenu au volet 2017. Ce taux est obtenu en multipliant les taux obtenus aux différentes étapes de pondération, selon le cas. Mentionnons que nous avons obtenu un meilleur taux de réponse pondéré transversal au volet 2017, soit 49,5 % comparativement à 47,3 % au volet 2015.

Tableau V Taux de réponse pondéré transversal au volet 2017

Taux de réponse au volet 1998	75,3 % (n=2 809)
Proportion de répondants (incluant les pseudo-répondants) au volet 2000 parmi les répondants au volet 1998 admissibles au volet 2017	95,0 % (n=2 086)
Proportion de répondants au volet 2002 parmi les répondants au volet 2000 admissibles au volet 2017 (incluant les pseudo-répondants)	98,9 % (n=1 998)
Proportion de répondants du poids « enfant » parmi les répondants au volet 2002 admissibles au volet 2017 (incluant les pseudo-répondants)	82,4 % (n=1 978)
Proportion de répondants au volet 2017 parmi les répondants du poids « enfant » admissibles au volet 2017 (incluant les pseudo-répondants)	84,9 % (n=1 648)
Taux de réponse transversal au volet 2017	49,5 %

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

5. Utilisation de la pondération par les utilisateurs des données du volet 2017

5.1 L'importance de la pondération

Les utilisateurs des données du volet 2017 sont fortement encouragés à utiliser la pondération lors des analyses des données de l'ELDEQ. 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.

Le taux de réponse au volet 2017 est de l'ordre de 49,5 % (voir tableau V). Ce taux confirme l'importance du traitement effectué lors de la pondération.

5.2. Tests statistiques

Le fichier POIDS2001 contient la variable de pondération PEGENT20 (poids général transversal du volet 2017). C'est un poids échantillonnal, c'est-à-dire un poids qui a été multiplié par une constante de sorte que la somme des poids soit égale à la taille de l'échantillon. Ce poids doit faire partie de toute analyse des données du volet 2017, comme indiqué à la section 5.1. Des logiciels statistiques, tels que SUDAAN, SAS ou STATA, permettent l'intégration de la pondération dans les différentes procédures offertes. En plus d'intégrer la pondération afin de minimiser les biais dans les estimations, le plan de sondage peut aussi être pris en compte lors des analyses. Le logiciel SUDAAN le permet, ainsi que certaines procédures du logiciel SAS. Ainsi c'est la variance qui est correctement estimée (pour les estimations et les tests statistiques).

Si les logiciels utilisés ne tiennent pas compte du plan de sondage complexe, le poids PEGENT20 peut être utilisé pour faire des tests approximatifs.

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

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.

5.3 Choix de la pondération

Les possibilités d'analyse incluant des données du volet 2017 sont innombrables. Ainsi, en raison de la non-réponse qui varie selon les instruments de collecte et les volets considérés, le

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

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

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

6. Références bibliographiques

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

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

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

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

ANNEXE A

Les étapes de la création d'une pondération générale

Voici la description de la séquence des étapes de création de la pondération transversale pour les participants au volet 2017.

Étape 1 :

Analyses bivariées pour réduire le nombre de variables considérées pour la modélisation (environ 60 variables). Les variables ayant les seuils observés les plus faibles sont conservées.

Étape 2 :

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

Étape 3 :

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

Étape 4 :

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

Étape 5 :

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

Étape 6 :

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

Étape 7 :

Ajustement, par « calage », des poids afin que la distribution pondérée des répondants s'ajuste à celle de la population de l'ELDEQ, selon une variable déterminée. Ainsi, la pondération 2017 est constituée.