UNIVERSITÉ DU QUÉBEC À MONTRÉAL

LES DÉTERMINANTS SOCIO-ÉCONOMIQUES DU PARCOURS SCOLAIRE

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN ÉCONOMIQUE

PAR

FRANÇOIS LALIBERTÉ-AUGER

FÉVRIER 2014

REMERCIEMENTS

Je tiens à remercier ma directrice, Marie Connolly, pour ses conseils et sa très grande disponibilité. Ses relectures attentives et sa patience ont contribué à la réalisation de ce mémoire. Je désire aussi remercier tout le personnel du CIQSS et plus particulièrement Élise Comoé qui malheureusement nous a quittés. Merci au FQRSC, au CRSH, au CIQSS et au Groupe d'Analyse ltée pour leur soutien financier. Merci à Vincent, Martin et David, avec qui j'ai partagé un bureau et sans qui le temps passé à travailler sur mon mémoire aurait été beaucoup moins divertissant. Je remercie mes collègues de maîtrise et le personnel du département d'économie pour l'ambiance de collégialité régnant au département.

Merci à ma famille et à mes amis d'avoir été présents lors de ces années de rédaction. Je remercie mes parents de m'avoir toujours soutenu et encouragé dans mes études. Finalement, merci infiniment à mon amoureuse, Sarah, de son soutien indéfectible, de ses nombreuses suggestions à la suite de ses nombreuses relectures et surtout d'avoir cru en moi et de m'avoir encouragé lors des périodes de doutes.

REMARQUE

Bien que la recherche et l'analyse soient fondées sur certaines des données non-publiques de Statistique Canada, les opinions exprimées ne représentent pas les vues de Statistique Canada.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AFQT	Armed Forces Qualification Test
EDTR	Enquête sur la dynamique du travail et du revenu
EJET	Enquête auprès des jeunes en transition
ELNEJ	Enquête longitudinale nationale sur les enfants et les jeunes
EM	Expectation maximization
ESG	Enquête sociale générale
GLLAMM	Generalized Linear Latent And Mixed Models
NPMLE	Nonparametric maximum likelihood estimation
OCDE	Organisation de coopération et de développement économique
PISA	Programme for International Student Assessment
WLE	Weighted maximum likelihood estimate

RÉSUMÉ

Dans ce mémoire, nous estimons l'effet de neuf déterminants socio-économiques sur le parcours scolaire des jeunes québécois. En raison du biais dans l'estimation qui résulte du modèle fréquemment utilisé, soit le modèle de transition qui utilise une séquence d'estimation logistique d'une transition à l'autre, nous utilisons un estimateur non-paramétrique de maximum de vraisemblance (NPMLE). Le principal intérêt de ce mémoire est de déterminer si l'utilisation de cette méthodologie est intéressante pour l'étude des déterminants socio-économiques du parcours scolaire. Les données utilisées sont celles de l'Enquête auprès des jeunes en transition (EJET). Nos résultats montrent que les déterminants les plus importants du parcours scolaire sont le niveau d'études des parents, les habitudes cognitives évaluées par le score au test du *Programme for International Student Assessment* (PISA) ainsi que le genre de la personne. À un moindre niveau, il y a aussi un effet des déterminants socio-économiques sur l'accès aux études collégiales ainsi que sur l'obtention d'un diplôme secondaire et sur l'accès à l'université. Dans nos estimations, le résultat le plus important est l'absence de différence entre les résultats obtenus avec l'estimateur logistique et ceux obtenus à l'aide du NPMLE. Ces résultats suggèrent que l'hétérogénéité inobservée ne joue pas un grand rôle dans le mécanisme de sélection entre les transitions. Nous avons procédé à une double estimation qui nous a permis de conclure que l'intérêt, pour le chercheur, d'utiliser l'estimateur NPMLE n'est pas suffisamment grand actuellement. Deux principales raisons ont été mentionnées : premièrement, son utilisation présente de trop grandes difficultés et, deuxièmement, il y a très peu de différences entre les résultats obtenus à l'aide de l'estimateur logistique et ceux qui résultent du NPMLE.

MOTS-CLÉS : parcours scolaire, Nonparametric maximum likelihood estimation, déterminants socio-économiques, EJET, Québec.

INTRODUCTION

L'éducation est au cœur du modèle social québécois. Lors de la Révolution tranquille, le gouvernement du Québec a fait le choix d'investir massivement dans le développement du réseau post-secondaire. Un des buts était de rattraper le retard important du taux de diplomation post-secondaire qu'accusait le Québec sur le reste du Canada et sur les États-Unis. Le choix fut fait d'offrir une accessibilité géographique, par la création des cégeps et du réseau des Universités du Québec, et une accessibilité financière par la gratuité du réseau collégial et un gel à long terme des frais de scolarité universitaires. À la même période, des chercheurs, notamment américains, ont posé les bases de l'étude des déterminants socio-économiques de l'accès à l'éducation. Malgré les nombreuses années de recherche, la question des facteurs d'accès revient fréquemment dans l'actualité, que ce soit sur la question du décrochage scolaire au niveau secondaire, sur la question des cégeps en régions qui ferment ou encore la question des frais de scolarité.

Une bonne compréhension des déterminants de l'accès à l'éducation est donc cruciale pour les décideurs publics et pour les intervenants du milieu pour mettre au point des politiques efficaces tenant compte du contexte social et des facteurs influençant l'inscription et l'obtention d'un diplôme. Ce mémoire étudie donc la relation entre les déterminants socio-économiques familiaux et le parcours scolaire des jeunes au Québec. De nombreuses études ont déjà analysé les liens entre les différents déterminants socio-économiques et l'accès aux études post-secondaires. Celles-ci se concentraient soit sur une seule transition, ou encore conjointement sur l'obtention d'un diplôme d'études secondaires et sur l'accès aux différents niveaux d'études post-secondaires (Cameron et Heckman, 2001). Cependant, la vaste majorité de ces études ne tient pas compte d'un biais lié à l'hétérogénéité des individus et à la sélection qui s'opère à chaque transition. Nous utiliserons donc un modèle proche de celui utilisé par Cameron et Heckman (1998) qui tient compte

de la sélectivité et de l'hétérogénéité.

Pour réaliser notre recherche empirique, nous utiliserons l'Enquête auprès des jeunes en transition (EJET) de Statistique Canada qui contient un ensemble de données qui répondent aux besoins de l'étude. Ceux-ci sont de deux ordres : les données doivent comporter suffisamment d'informations sur le parcours scolaire et contenir un nombre suffisamment élevé d'observations pour le Québec. Les données de l'EJET remplissent ces besoins. De plus, les données de l'EJET comprennent les résultats de l'enquête de l'Organisation de coopération et de développement économique (OCDE) sur l'acquisition de savoirs et de savoir-faire en lecture soit le *Programme for International Student Assessment* (PISA) (OCDE, sd). Ces résultats permettent de contrôler pour les aptitudes des jeunes à 15 ans, soit avant qu'une sélection ne s'effectue dans le système éducatif.

L'utilisation d'un modèle corrigeant le biais lié à l'hétérogénéité inobservée est, à notre connaissance, une première pour l'étude des déterminants socio-économiques du parcours scolaire québécois. Nous montrerons que l'utilisation d'un tel modèle ne fait pas varier de façon importante les résultats obtenus par rapport au modèle habituellement utilisé et que son utilisation se révèle ardue pour des bénéfices limités.

Ce mémoire se divise en cinq chapitres. Dans le premier chapitre, nous dressons un portrait de la littérature portant sur les différents modèles utilisés pour étudier les déterminants socio-économiques de l'accès à l'éducation et nous exposons les critiques qui sont adressées aux modèles fréquemment utilisés pour estimer l'effet des différents déterminants. Puis, nous présentons la littérature portant sur les déterminants socio-économiques de l'accès aux études post-secondaires et finalement, nous regardons les différentes données utilisées dans les études tant américaines que canadiennes. Dans le deuxième chapitre, nous présentons le modèle le plus fréquemment utilisé et exposons les limites de ce modèle en raison d'un biais lié à l'hétérogénéité inobservée. Dans le troisième chapitre, nous décrivons l'estimateur que nous utilisons pour corriger le biais lié à l'hétérogénéité inobservée. Le quatrième chapitre présente la source de données retenue, soit l'EJET, les différents choix dans la sélection de l'échantillon et les différentes

caractéristiques des variables retenues. Dans le cinquième chapitre, nous présentons les résultats, nous les discutons, et nous analysons le bien-fondé de l'utilisation de l'estimateur. En effet, nous évaluons si la méthodologie employée dans ce mémoire présente des résultats différents ou plus concluants que ceux obtenus traditionnellement. Une brève conclusion suivra.

CHAPITRE I

REVUE DE LITTÉRATURE

Dans ce chapitre, nous brossons un portrait de la littérature sur les déterminants socio-économiques du parcours scolaire et les modèles qui les étudient. Dans un premier temps, nous présentons, historiquement, quels furent les modèles empiriques utilisés dans l'étude de ces déterminants. Dans un second temps, nous nous intéressons aux travaux plus récents qui démontrent que les modèles utilisés ne tiennent pas compte d'un biais lié à l'hétérogénéité inobservée ainsi qu'à un problème de sélectivité. Dans un troisième temps, nous présentons les résultats obtenus par ces modèles, c'est-à-dire que nous décrivons les différents déterminants de l'accès à l'éducation post-secondaire. Finalement, nous nous attardons aux différentes sources de données canadiennes utilisées pour étudier la question.

1.1 Les différents modèles empiriques utilisés dans l'étude des déterminants socio-économiques

Dans cette section, nous présentons l'évolution des travaux, principalement réalisés aux États-Unis, portant sur les déterminants socio-économiques du parcours scolaire. Les premiers travaux de nature empirique portant sur la question des déterminants socio-économiques ont porté sur des mesures agrégées de l'accès à l'éducation (Duncan, 1965 cité dans Mare, 2011). Ces premières études ont fait place à des travaux utilisant des modèles linéaires visant à identifier les déterminants du nombre d'années d'études complétées (Mare, 2011; Duncan, 1967; Spady, 1967). Ces modèles qui étudient une mesure

limitée du niveau d'éducation ont été suivis par des travaux sur les différentes transitions entre les niveaux d'éducation. Les modèles de transition sont construits sur le fait que le parcours scolaire est fondé sur un processus cumulatif qui est dépendant du temps. Même si cela n'est pas complètement vrai, l'éducation peut être vue comme une « séquence unique et irréversible de transitions » [1] (Mare, 2011). Ces transitions sont modélisées à l'aide de régressions logistiques qui permettent d'obtenir de l'information sur l'effet des déterminants socio-économiques sur les différentes transitions jonchant le parcours de vie des jeunes. Les travaux fondateurs dans le domaine ont été réalisés par Mare (Mare, 1979; Mare, 1980; Mare, 1981), travaux largement cités par la suite.

De nombreux ajouts ont été apportés à ces travaux précurseurs. Nous retrouvons entre autres, selon Mare (2011), des recherches comparatives en stratification de l'éducation (Shavit et Blossfeld, 1993) et des modèles pour les systèmes qui ne peuvent être réduits à une simple séquence irréversible de transitions (Breen et Jonsson, 2000). Il y a aussi les effets liés aux transitions des parents (Mare et Chang, 2006), l'utilisation de données panel avec des variables explicatives variant dans le temps pour les modèles de transitions (Lucas, 2001; Cameron et Heckman, 2001) et finalement la critique et l'apport qui sont au centre de ce mémoire : un traitement statistique formel du biais de sélection entre autres à l'aide d'un modèle semi-paramétrique pour l'hétérogénéité inobservée (Cameron et Heckman, 1998). Nous revenons plus en détails sur ce modèle dans la prochaine section.

Suivant la critique apportée par Cameron et Heckman (1998), des auteurs ont repris la méthodologie proposée dans cet article. Nous retrouvons entre autres les travaux présentés lors du symposium sur l'hétérogénéité non mesurée dans les modèles de transition scolaire [2]. Dans ce symposium, nous retrouvons un exemple de modèle multinominal avec une correction pour un terme d'hétérogénéité (Karlson, 2011) basé sur les travaux de Cameron et Heckman. Nous retrouvons aussi un article étudiant les gains obtenus par l'utilisation de l'estimateur proposé par Cameron et Heckman, le tout à l'aide de simu-

1. Traduction libre.

2. Travaux repris dans le numéro 29 de la revue *Research in Social Stratification and Mobility*

lations afin de pouvoir quantifier les effets de l'estimateur (Buis, 2011). Malgré la sévère critique apportée par Cameron et Heckman, de nombreux auteurs continuent d'utiliser un modèle proche du modèle de transition de Mare (Corak, Lipps et Zhao, 2003; Finnie, Lascelles et Sweetman, 2005; Kamanzi *et al.*, 2009), nous pouvons penser que cela résulte de la complexité de l'application de l'estimateur corrigeant pour l'hétérogénéité inobservée, tant d'un point de vue théorique que d'un point de vue pratique.

1.2 Le biais de sélection et l'hétérogénéité inobservée

Dans la section précédente, nous avons introduit les travaux de Cameron et Heckman. Nous abordons ici plus largement la question du biais de sélection et de l'hétérogénéité inobservée. Cette hétérogénéité est liée à des caractéristiques inobservables pour le chercheur, qui varient d'un individu à l'autre, et qui ont un impact sur la probabilité de sélection entre les transitions. L'existence de ce biais était déjà connue à l'époque des travaux de Mare (1980). L'auteur mentionne cette limitation dans ses travaux de l'époque. Cependant, les outils permettant de traiter un tel biais étaient naissants et inconnus du chercheur à l'époque (Mare, 2011). Il faut attendre les travaux de Cameron et Heckman (1998) pour obtenir une critique formelle du modèle de transition. Dans leur article, les auteurs montrent qu'il existe un double biais lié aux facteurs inobservés, soit un biais lié au caractère non linéaire du modèle logistique et un biais lié à la sélection sur les facteurs inobservables (Cameron et Heckman, 1998, pp. 272-275). Ce biais de sélection apparait lorsqu'entre les transitions la sélection est influencée par les facteurs inobservables. Alors que les facteurs inobservables pouvaient être indépendants des facteurs observables lors d'une première transition, suite aux transitions, ces facteurs deviennent corrélés et ceci introduit un biais dans l'estimation des coefficients du modèle de transition. Cette corrélation mène à ce que l'estimation des effets lors des transitions subséquentes ne s'effectue plus sur la même population que dans les premières estimations, la distribution des inobservables n'étant pas la même entre les échantillons. Cameron et Heckman (1998) proposent aussi deux estimateurs permettant de corriger pour la présence d'hétérogénéité, soit un estimateur non paramétrique de maximum de vraisemblance et un modèle

de choix ordonné (Cameron et Heckman, 1998). Nous couvrirons largement la question du biais lié à l'hétérogénéité dans le chapitre 2 et de l'estimateur non-paramétrique dans le chapitre 3.

1.3 Les déterminants socio-économiques

L'étude des déterminants socio-économiques des parcours scolaires a porté sur différentes métriques de ces parcours comme nous l'avons vu précédemment. Ce qui ressort, c'est que les conclusions sont semblables lorsque l'on regarde les déterminants du nombre d'années d'études complétées et ceux de la probabilité d'effectuer les différentes transitions. De plus, l'utilisation du modèle de transition mène à l'observation d'un déclin des effets lorsque l'on progresse dans les transitions. Cependant, lorsque l'on introduit un contrôle pour l'effet d'hétérogénéité inobservée, ce déclin s'atténue. Dans cette section, nous abordons aussi le rôle complexe des compétences personnelles dans la réussite scolaire.

1.3.1 Description des déterminants socio-économiques

Dans les premiers travaux dans lesquels l'accent a été mis sur l'identification des déterminants du nombre d'années d'études complétées, les chercheurs ont conclu que le salaire familial et le niveau d'éducation des parents avaient une forte influence sur le nombre d'années d'études complétées (Duncan, 1967; Spady, 1967). Par la suite, lorsque les chercheurs se sont attardés sur les déterminants de la probabilité d'effectuer les différentes transitions possibles dans un parcours scolaire, un consensus s'est dégagé selon lequel la probabilité d'effectuer les différentes transitions varie de façon positive en fonction d'une augmentation du revenu des parents (Christofides, Hoy *et al.*, 2001; Mare, 1980; Cameron et Heckman, 1998). Cependant, les chercheurs ont trouvé un effet plus positif et plus important pour le diplôme; l'ajout de la variable du diplôme des parents, soit du père, de la mère ou des deux, réduit grandement l'effet du salaire (Finnie, Lascelles et Sweetman, 2005; Christofides, Hoy *et al.*, 2001). Les chercheurs ont aussi identifié un effet négatif marqué, sur l'accès à l'université, de la distance à celle-ci ainsi que d'ha-

biter une région rurale (Christofides, Hoy *et al.*, 2001; Finnie, Lascelles et Sweetman, 2005; Frenette, 2002). Selon les auteurs, cet effet n'est pas observé au niveau collégial probablement en raison de leur plus grande accessibilité géographique (Finnie, Lascelles et Sweetman, 2005; Frenette, 2002; Frenette, 2003). On observe aussi un déclin de l'effet lorsque l'on progresse dans les transitions (Mare, 1980). Ce déclin serait possiblement un artefact de la forme fonctionnelle utilisée et de l'absence de contrôle pour l'hétérogénéité inobservée (Cameron et Heckman, 1998). Lorsqu'un contrôle pour les facteurs inobservables est ajouté, le déclin couramment observé de l'ampleur des déterminants entre les transitions est amoindri (Cameron et Heckman, 1998; McIntosh, 2010).

1.3.2 L'évolution des déterminants socio-économiques au Canada

Au Canada, de nombreux auteurs ont tenté de quantifier les variations dans le temps des effets liés aux déterminants socio-économiques. Certains ont estimé que les inégalités d'accès liées aux revenus et aux études des parents n'avaient pas évolué au cours des années 1990 (Corak, Lipps et Zhao, 2003), alors que d'autres ont estimé qu'il y a eu une baisse des inégalités entre 1975 et 1993 (Christofides, Hoy *et al.*, 2001). Sur une plus longue période, un auteur soutient qu'il n'y a pas eu de baisse des inégalités entre 1920 et 1994 (Wanner, 1999), cependant, cette assertion est constestée en raison de la méthodologie utilisée par l'auteur. En effet, celle-ci ne tient pas compte de l'hétérogénéité inobservée et, lorsque l'on prend en compte celle-ci, nous observons une baisse des inégalités de l'accès à l'éducation sur une période de 50 ans (McIntosh, 2010).

1.3.3 Le rôle des compétences dans le parcours scolaire

Les compétences personnelles jouent un rôle important dans la réussite scolaire, mais l'interaction entre les deux est complexe. En effet, le parcours scolaire affecte les compétences, mais inversement les compétences affectent le parcours scolaire. Ces compétences sont elles-mêmes déterminées par de nombreux facteurs, outre la scolarité effectuée. Ces compétences, notament la littératie et la numératie, sont influencées par plusieurs facteurs, entre autres le niveau d'étude des parents, l'origine ethnique et le statut d'immi-

grant (Finnie et Frenette, 2003; Bonikowska, Green et Riddell, 2008). Les compétences ne se résument donc pas aux capitaux scolaires acquis, ces premières évoluant après les études (Kamanzi *et al.*, 2009). De par le rôle important des compétences, plusieurs chercheurs ont décidé d'inclure des variables quantifiant ces compétences. Plusieurs études américaines importantes utilisent le test de compétences de l'armée, soit le *Armed Forces Qualification Test* (AFQT), afin de contrôler pour les compétences (Mare, 1980; Cameron et Heckman, 1998). Au niveau des études canadiennes, nous pouvons noter que les études incluant de telles variables sont plus rares. Ceci est explicable par l'absence de résultats de tests de compétences dans les enquêtes les plus fréquemment utilisées par les chercheurs. La situation a changé avec l'arrivée de l'*Enquête sur les jeunes en transitions* qui inclut, pour une des cohortes étudiées, les résultats au test PISA conduit par l'OCDE. Peu d'études sur les déterminants socio-économiques de l'accès aux études ont utilisé, à ce jour, ces données. À notre connaissance, seul le *Projet Transitions* de la *Fondation canadienne des bourses d'études du millénaire* l'a fait (Kamanzi *et al.*, 2009). Dans leur article, les auteurs concluent que les déterminants socio-économiques ont un effet nettement plus marqué sur la fréquentation universitaire que sur la participation aux études collégiales. De plus, ils constatent que les variables d'appartenance sociale ont un effet plus grand sur l'inscription que sur la persévérence.

1.4 Les sources de données

Les différentes études mentionnées plus haut utilisent une grande variété de sources de données offrant tout un éventail d'informations. Nous présentons ici un bref survol des différentes sources de données utilisées afin de mieux cerner les caractéristiques recherchées dans une base de données sur le sujet. Nous portons notre attention sur les études canadiennes, afin de recenser les différentes sources possibles de données pour un chercheur désirant étudier les déterminants socio-économiques de l'accès à l'éducation dans un contexte canadien ou encore québécois. Ces études sont encore fréquentes dans le cadre de l'évaluation de l'évolution dans le temps de la probabilité des différentes transitions ainsi que dans l'étude de la disparité des effets entre les transitions.

L'*Enquête sociale générale* (ESG) est utilisée par de nombreux chercheurs (McIntosh, 2010; Wanner, 1999). Cette enquête présente un intérêt pour les études sur de longues périodes, car elle a été menée annullement depuis 1985 auprès de la population ayant plus de 15 ans. Lors des cycles portant sur la famille, l'ESG fournit de l'information sur la scolarité des parents, sur la langue parlée, sur le genre, sur l'origine urbaine ou rurale, sur le statut du milieu familial (monoparental ou autre) et bien d'autres données. Plusieurs informations ne sont pas disponibles, dont le revenu des parents et une mesure des compétences. De plus, l'information sur les parents est demandée aux répondants, cette information est donc moins fiable surtout lorsque les répondants sont plus âgés, puisque l'information qui leur est demandée remonte à de nombreuses années. Wanner (1999) utilise aussi, pour les données des années antérieures à l'ESG, l'*Enquête sur la mobilité canadienne* de 1973 qui comprend des informations semblables et compatibles avec l'ESG. Christofides *et al.* (2001) ainsi que Corak *et al.* (2003) ont quant à eux utilisé l'*Enquête sur les finances des consommateurs* qui offre de l'information sur les ménages. L'information sur les parents est directement fournie par ceux-ci ce qui offre des informations plus fiables. Cependant, les auteurs ont dû se limiter à utiliser les informations sur les familles ayant des jeunes de 18 à 24 ans vivant à la maison, ce qui a réduit l'échantillon de jeunes disponibles de 20 à 30 % (Corak, Lipps et Zhao, 2003). Dans cette base de données sont aussi absentes toutes les variables liées aux compétences. Drolet (2005) utilise, quant à elle, l'*Enquête sur la dynamique du travail et du revenu* (EDTR), une enquête longitudinale qui vise à quantifier pour les individus le changement dans leur bien-être économique (Drolet, 2005). Cette enquête fournit des informations sur le revenu parental et sur le type de famille. Cependant, ces informations sont seulement disponibles pour les jeunes demeurant avec un de leurs parents, réduisant du même coup la taille de l'échantillion. L'enquête contient aussi des informations sur le niveau d'études des parents pour tous les jeunes ainsi que sur l'âge, sur le genre et sur la province de résidence. De leur côté, Lefebvre et Merrigan (2010) utilisent l'*Enquête longitudinale nationale sur les enfants et les jeunes* (ELNEJ). Cette enquête contient plusieurs éléments intéressants, entre autres des résultats de tests de mathématique ainsi que des informations répétées dans le temps sur les parents dont leur revenu et

leur éducation. Comme cela est mentionné plus haut, une enquête intéressante dans le présent contexte est l'*Enquête sur les jeunes en transition* (EJET) (Finnie, Laporte et Lascelles, 2004; Kamanzi *et al.*, 2009). Cette enquête longitudinale suit deux cohortes de jeunes qui sont re-contactés aux deux ans. En plus de fournir de l'information exhaustive sur les jeunes, cette enquête interroge directement les parents d'une des deux cohortes et fournit les données relatives aux scores PISA pour cette même cohorte. Nous fournissons d'amples détails sur l'EJET dans le chapitre 4 consacré aux données utilisées dans ce mémoire.

CHAPITRE II

LES MODÈLES DE TRANSITION ET LE BIAIS D'HÉTÉROGÉNÉITÉ INOBSERVÉE

Dans ce chapitre, nous commençons par présenter, dans la première section, le modèle de transition soit le modèle le plus utilisé pour étudier les déterminants socio-économiques de l'accès aux études (Mare, 1980; Christofides, Hoy *et al.*, 2001; Finnie, Lascelles et Sweetman, 2005). Dans la deuxième section, nous regardons les problèmes induits par les facteurs d'hétérogénéité qui ne sont pas observés par le chercheur et qui rendent les résultats obtenus à l'aide du modèle de transition peu fiables. Nous terminons le chapitre en présentant les résultats de notre simulation de l'estimation du modèle de transition lorsque nous sommes en présence d'hétérogénéité inobservée.

2.1 Le modèle de transition

Les premiers modèles économétriques utilisant des micro-données se sont concentrés, à partir des années 1960, sur l'étude des déterminants du plus haut niveau d'éducation (Mare, 2011; Cameron et Heckman, 1998). Ces modèles furent suivis par les travaux fondateurs de Mare (1980) qui développent un modèle séquentiel logit que nous appelons aussi le modèle de transition. Ce modèle a comme objectif d'étudier l'effet des déterminants socio-économiques sur les différentes transitions entre des niveaux d'éducation pouvant être vécues par les jeunes. Il s'agit d'un modèle qui utilise un modèle logit pour prédire des événements dichotomiques, c'est-à-dire le fait d'effectuer ou de ne pas effectuer une transition entre deux niveaux d'éducation.

Une présentation formelle permet une meilleure compréhension du modèle et permet de jeter les bases du modèle logit qui sera utilisé plus tard. Cette présentation formelle reprend celle de Cameron et Heckman (1998). Soit s le niveau d'éducation et $D_s = 1$ si une personne a diplômé du niveau s, sinon $D_s = 0$. Soit $X_s = x_s$ les déterminants de la transition du niveau $s-1$ au niveau s. La probabilité de transition du niveau $s-1$ au niveau s est donc :

$$Pr(D_s = 1|X_s = x_s, D_{s-1} = 1) = P_{s-1,s}(x_s) \tag{2.1}$$

Bien sûr, nous ne pouvons observer $Pr(D_s = 1|X_s = x_s, D_{s-1} = 0)$. Les auteurs ont fréquemment utilisé une fonction logistique pour modeler la probabilité (Mare, 1980) :

$$P_{s-1,s}(x_s) = \frac{exp(x_s\beta_s)}{1 + exp(x_s\beta_s)} \tag{2.2}$$

Ce modèle fournit des coefficients pour chacun des déterminants, pour chacune des transitions, ce qui constitue une avancée par rapport aux modèles utilisés antérieurement, comme ceux qui étudient le plus haut niveau d'études atteint (Cameron et Heckman, 1998). Cependant, ce modèle comporte de graves lacunes soulevées dès sa première utilisation par Mare (1980) et reprises dans une critique en règle faite par Cameron et Heckman (1998).

2.2 L'hétérogénéité inobservée

Le modèle de transition, par la nature sélective des transitions, est grandement affecté par l'hétérogénéité inobservée. Cette hétérogénéité inobservée correspond aux facteurs qui différencient les individus et qui ne sont pas observés par le chercheur. Il existe deux types de facteurs inobservés, ceux qui sont statistiquement indépendants des transitions s et ceux qui varient entre les transitions. Cameron et Heckman (1980) présentent formellement le biais et fournissent les preuves. Ce qui suit est un résumé de leur présentation (Cameron et Heckman, 1998). Redéfinissons les déterminants comme $X_s = (X_{so}, X_{si})$, où « o » identifie les X_s observés par le chercheur et où « i » identifie la portion inobservée. Soit Θ_s un scalaire où $\Theta_s = X_{si}\beta_{si}$. Pour la suite de ce mémoire, $X = (X_{1o}, \ldots, X_{\bar{S}o})$ où $\bar{S}$ est le dernier niveau d'études.

Pour définir le problème, ils imposent quatre restrictions, soit la forme de l'hétérogénéité, l'indépendance entre Θ_s et X, la constance de l'hétérogénéité entre les transitions et la fonction G est connue :

(H-1) $Pr(D_s = 1|D_{s-1} = 1, X = x, \Theta_s = \theta_s) = G(x_s\beta_s + \theta_s)$

(H-2) $\Theta = (\Theta_1, \ldots, \Theta_{\bar{S}})$ est indépendant de $X = (X_1, \ldots, X_{\bar{S}})$

(H-3) $\Theta_s = \Theta$

(H-4) $G_s = G$ est connue et a un nombre fini de paramètres.

Cameron et Heckman balisent le problème avec ces hypothèses en le restreignant à un type de facteur inobservé. La première hypothèse limite l'étude à une hétérogénéité qui entre linéairement dans la fonction de probabilité. La deuxième impose une hypothèse d'indépendance entre Θ et X, cette restriction n'implique pas l'indépendance dans les périodes suivant la période initiale d'éducation. Nous verrons plus loin que malgré cette indépendance, l'omission de variables crée un biais dans l'estimation. La troisième hypothèse force les effets à être constants entre les périodes, comme le mentionnent les auteurs ces Θ peuvent être interprétés comme « l'habileté » ou encore la « motivation ». Dans la prochaine section, nous montrerons que, malgré la forme restrictive de l'hétérogénéité, celle-ci introduit un biais important dans l'estimation du modèle de transition.

2.3 Le biais lié à l'hétérogénéité inobservée

Dans cette section, nous caractérisons les biais introduits par l'hétérogénéité inobservée. Dans un premier temps, nous présentons de façon intuitive ce qu'est le biais d'hétérogénéité, par la suite, nous reprenons la présentation théorique de Caremon et Heckman (1998) et finalement nous présentons des simulations que nous avons effectuées afin d'illustrer le biais introduit par l'hétérogénéité.

Comme mentionné précédemment, le biais d'hétérogénéité est introduit par la présence de facteurs individuels qui sont inobservés par le chercheur. Ces facteurs introduisent

deux types de biais. Le premier biais provient de la sélection qui se fait entre les niveaux d'études en fonction des facteurs inobservables. Le concept d'hétérogénéité inobservée peut être vu comme provenant des différences au niveau des compétences. Par exemple, si nous divisons un groupe de jeunes en deux groupes égaux, inobservés pour le chercheur, ceux du groupe « fort » ayant des compétences élevées et ceux du groupe « faible » ayant des compétences faibles, alors l'obtention d'un diplôme d'études secondaires sera entre autres déterminée par l'appartenance à un de ces deux groupes. Par la suite, les jeunes éligibles au collège, soit ceux ayant obtenu un diplôme d'études secondaires, ne seront plus répartis également dans les deux groupes d'aptitude, ceux étant dans le groupe « fort » réussissant mieux que ceux dans le groupe « faible ». Plus nous avançons dans le parcours scolaire, plus cet effet de sélection sur les inobservables est fort, ce qui cause un biais dans la comparaison des estimations entre les transitions puisque l'estimation ne se fait pas sur les mêmes groupes, les groupes des transitions supérieures n'étant pas représentatifs de la population initiale. Le deuxième biais résulte de la corrélation qui se crée entre les observables et les inobservables lorsque l'on avance dans le parcours scolaire. Reprenons l'exemple précédent, et posons que les compétences ne sont pas corrélées avec les variables observées lors de l'obtention du diplôme d'études secondaires : en termes économétriques, les variables X ne sont donc pas corrélées avec le terme d'erreur ϵ. Cependant, lors de la sélection pour l'obtention d'un diplôme, dans le groupe « faible » seuls les jeunes ayant des variables observées « élevées » vont diplômer, alors que ceux provenant du groupe « fort » pourrons diplômer même avec des variables observées « faibles ». Lors du niveau d'étude suivant, la distribution des variables observées ne sera donc plus la même pour le groupe « fort » et pour le groupe « faible », créant donc une corrélation entre les variables observées et celles inobservées. Ceci introduit donc un biais puisqu'il y a corrélation entre les variables X et le terme d'erreur ϵ.

Cameron et Heckman fournissent une démonstration de ce biais lié l'hétérogénéité inobservée, biais induit même lorsque ces facteurs inobservés sont indépendants des variables observées. Si la véritable probabilité de transition peut s'exprimer comme suit :

$$Pr(D_1 = 1|X = x, \Theta = \theta) = \frac{exp(x\beta_1 + \theta)}{1 + exp(x\beta_1 + \theta)} \tag{2.3}$$

mais que nous estimons le modèle :

$$Pr(D_1 = 1|X = x) = \frac{exp(x\gamma_1)}{1 + exp(x\gamma_1)} \tag{2.4}$$

Alors, si nous assumons que le tirage est aléatoire et que nous sommes en grand échantillon, le biais est égal à :

$$plim(\hat{\gamma_1} - \hat{\beta_1}) = \int_0^1 \left[E_{X,\theta}\left(\frac{exp(x\beta_1 + [\theta]\lambda)}{[1 + exp(x\beta_1 + [\theta]\lambda)]^2} XX' \right) \right]^{-1} \cdot E_{X,\theta}\left[\frac{exp(x\beta_1 + [\theta]\lambda)}{[1 + exp(x\beta_1 + [\theta]\lambda)]^2} XX' \right] \mathrm{d}\lambda \tag{2.5}$$

Une démonstration du résultat est fournie par Cameron et Heckman (1998). Le terme du biais n'est pas signé, mais n'est pas égal à zéro habituellement. Il y a donc un biais même si les caractéristiques inobservées sont indépendantes de celles observées. Ceci se produit en raison de la non-linéarité du modèle qui rend inséparable Θ et X. De plus, la nature dynamique du modèle de transition introduit deux autres sources de biais dans les estimations des déterminants des transitions. Premièrement, lorsque l'on avance dans les transitions, les personnes avec de « hauts » Θ sont plus susceptible de rester, il y a donc un déplacement de distribution des Θ. De plus, parmi les familles pauvres, les jeunes ayant des Θ élevés sont les seuls à effectuer les transitions, ce qui entraîne une corrélation négative entre Θ et X.

2.4 La simulation

Précédemment, nous avons présenté les preuves démontrant qu'en présence d'hétérogénéité inobservée une estimation des déterminants socio-économiques par modèle logit était biaisée. Ce biais était introduit, premièrement, par la non-séparabilité des déterminants observables et non-observables dans un modèle non-linéaire comme le modèle logit. Deuxièment, des effets dynamiques induits par la sélection d'une étape à l'autre introduisent un second biais en raison de la sélection qui s'opère sur les données inobservables. Afin de quantifier ce biais, nous avons effectué une simulation dont nous présentons les résultats dans cette section. Celle-ci vise à mettre en lumière l'ampleur du

biais introduit par l'ajout de facteurs inobservés. Nous pouvons voir que l'introduction de tels facteurs crée un biais dès la première transition et que celui-ci s'amplifie lors des transitions supérieures.

2.4.1 La génération des données

Nous effectuons une simulation dans laquelle nous générons des données avec un terme d'erreur distribué logistiquement et nous ajoutons un terme d'hétérogénéité distribué normalement. Plus précisément, nous générons aléatoirement des valeurs observées (x_1) en tirant des valeurs d'une distribution suivant une loi normale de moyenne zéro et avec un écart-type de 1. Nous générons un terme d'erreur en pigeant des valeurs (u) d'une distribution uniforme et en générant des termes d'erreurs (ϵ) égaux à $-ln(\frac{(1-u)}{u})$. Par la suite, nous tirons des valeurs inobservées (θ) d'une distribution normale de moyenne zéro et nous faisons varier l'amplitude, c'est-à-dire la variance de la distribution de θ, durant la simulation. Nous générons ensuite une variable latente (y*) qui est égale à $1 + \beta * x_1 + 1 * \theta + \epsilon$ où $\beta = 1$. Finalement, nous générons les valeurs dépendantes observées (y) en attribuant la valeur de 1 à y si $y* > 0$ et la valeur de 0 autrement. Nous effectuons trois fois ces étapes pour générer trois transitions (y_1, y_2, y_3). Les trois transitions ont les mêmes valeurs observées, inobservées et le même terme d'erreur. La différence se situe au niveau de la sélection, les observations ayant un $y_1 = 0$ se voient attribuer automatiquement un $y_2 = 0$ et un $y_3 = 0$ et les observations ayant un $y_2 = 0$ se voient attribuer un $y_3 = 0$.

2.4.2 L'estimation et la simulation

L'estimation est effectuée à l'aide du logiciel Mata inclus dans Stata. Nous avons créé un court programme Mata dans lequel nous utilisons la commande Mata *optimize* qui permet de trouver le maximum d'une fonction définie par l'utilisateur. La fonction prend

la forme suivante :

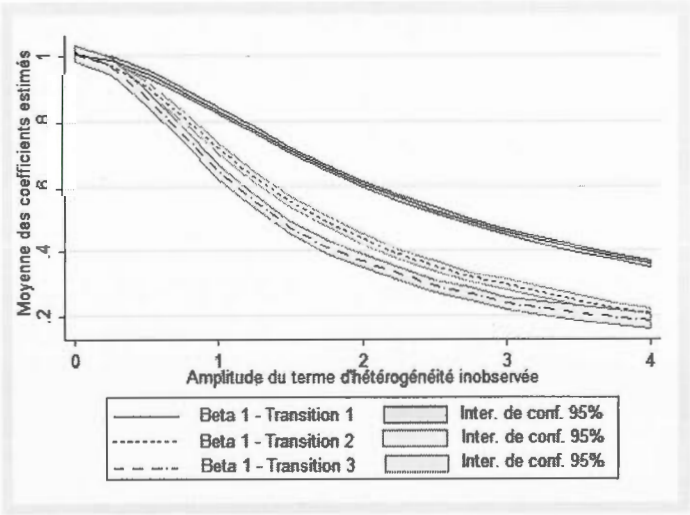
$$Qn = \sum_{i=1}^{N} \left\{ y_{1,i} * ln\left(\frac{exp[x_{1i}b_1]}{[1+x_{1i}b_1]}\right) + (1-y_{1,i}) * ln\left(\frac{1-exp[x_{1i}b_1]}{[1+x_{1i}b_1]}\right) \right.$$
$$+ y_{1,i} * ln\left(\frac{exp[x_{1i}b_2]}{[1+x_{1i}b_2]}\right) + (1-y_{2,i}) * ln\left(\frac{1-exp[x_{1i}b_2]}{[1+x_{1i}b_2]}\right)$$
$$\left. + (y_{1,i} * y_{2,i}) * ln\left(\frac{exp[x_{1i}b_3]}{[1+x_{1i}b_3]}\right) + (1-y_{3,i}) * ln\left(\frac{1-exp[x_{1i}b_3]}{[1+x_{1i}b_3]}\right) \right\} / N$$

Nous avons encapsulé le code Mata dans un programme Stata qui génère des données selon l'algorithme défini plus haut et qui appelle la commande Mata afin d'estimer les paramètres. Finalement, nous utilisons la commande *simulate* afin d'effectuer 200 piges et estimations et nous prenons la moyenne des estimations. Nous répétons cette étape en faisant varier l'amplitude du terme inobservé. Nous répétons cette étape pour des variances nulles à quatre fois celle de la variable observée.

2.4.3 Les résultats de la simulation

Les résultats de la simulation sont présentés dans la figure 2.1. Ce que nous observons

Figure 2.1 Simulation d'estimation avec biais



est concordant avec les démonstrations précédentes. Premièrement, nous observons que l'estimation s'effectue sans biais pour les trois étapes lorsqu'il n'y a pas d'hétérogénéité

inobservée (c'est-à-dire que l'amplitude est égale à 0). Deuxièmement, nous observons dès la première transition un biais important malgré qu'il n'y ait pas encore d'effet de sélection sur les valeurs inobservées. Il s'agit donc du biais induit par l'inséparabilité des x_1 et de θ. Troisièmement, nous observons que, lors des transitions, le biais augmente en raison de la sélection. Le biais devient rapidement important. Pour la troisième transition, lorsque le biais est de même ampleur que la valeur observée (c'est-à-dire que l'amplitude est égale à 1), nous obtenons un estimé de 0,6, alors que le coefficient est de 1. Dans le cas d'un biais très important (4 fois l'ampleur de x_1), nous estimons un coefficient près de 0. De nombreuses autres simulations pourraient être effectuées, en faisant varier les distributions, le nombre de variables explicatives, etc. Cependant, ceci dépasse le cadre de ce mémoire. Avec ces modestes simulations, nous montrons que la présence d'hétérogénéité inobservée introduit un biais important qui s'amplifie avec l'amplitude du biais et qui peut venir fausser les résultats. Nos résultats sont aussi concordants avec les preuves apportées par Cameron et Heckman (1998) que nous avons présentées dans la première partie de ce chapitre. En effet, ces preuves montrent qu'il existe un biais même sans sélection et que ce biais est amplifié par la sélection qui s'opère sur les valeurs inobservables. Même si les variables inobservables étaient indépendantes des variables

Tableau 2.1 Moyenne et corrélation des variables observées et inobservées

Transitions	Moyenne de X_1	Moyenne de θ	Corrélation
Première	-0,0132	0,0039	-0,0103
Deuxième	0,3666	0,3741	-0,1581
Troisième	0,5642	0,5709	-0,2556

Source : Simulation réalisée par l'auteur. Les X_1 et les θ sont tirés d'une loi normale.
Note : la variance de X_1 et θ est égale à 1.

observées avant la première sélection, elles deviennent corrélées lors de la deuxième sélection. Il s'agit d'une corrélation inverse puisque, chez les individus à faible x, seuls les individus avec de haut Θ passent à la seconde transition. C'est ce que nous observons

dans le tableau 2.1. En effet, alors que d'une transition à l'autre, la moyenne des effets observés et inobservés augmente, la corrélation entre les deux variables devient négative, augmentant le biais.

CHAPITRE III

L'ESTIMATEUR SEMI-PARAMÉTRIQUE ET LES CONDITIONS D'IDENTIFICATION DE L'ESTIMATEUR

Dans ce chapitre, nous présentons le modèle que nous utiliserons afin d'estimer l'effet des déterminants socio-écomiques sur le parcours scolaire. Il s'agit d'un modèle de transition intégrant un terme d'hétérogénéité tel que suggéré par Cameron et Heckman (1998). Dans un premier temps, nous décrivons l'estimateur non-paramétrique de maximum de vraisemblance [1] ou NPMLE pour *nonparametric maximum likelihood estimation.* Par la suite, nous esquissons un modèle économique justifiant le NPMLE ainsi que les conditions d'identification du modèle économique et du NPMLE. Puis, nous détaillons la méthode utilisée pour estimer le NPMLE et obtenir les écarts-types. Finalement, nous présentons les résultats d'une simulation que nous avons effectuée afin d'évaluer l'efficacité du NPMLE à corriger le biais lié à l'hétérogénéité inobservée.

3.1 L'estimateur non-paramétrique de maximum de vraisemblance

Au chapitre précédent, nous avons présenté et caractérisé le biais introduit par des facteurs inobservés lors des estimations par modèle logit. Dans cette section, nous présentons un modèle augmenté d'un contrôle pour l'hétérogénéité inobservée. Ce modèle se base sur des travaux d'Heckman et Singer (1984) ainsi que ceux de Follman (1985)

---

1. Bien qu'il s'agisse d'un estimateur semi-paramétrique puisqu'une partie de l'estimation est paramétrique et l'autre non, l'estimateur est nommé non-paramétrique en raison de la partie complètement non-paramétrique de l'estimation du terme d'hétérogénéité.

et de Cameron et Taber (1998). L'utilisation d'un tel estimateur dans le cas de l'étude des transitions entre les niveaux d'éducation est suggérée par Cameron et Heckman (1998). Nous résumons donc dans cette section la présentation faite de cet estimateur par Cameron et Taber (1998) et par Cameron et Heckman (1998).

Le NPMLE est un estimateur par maximum de vraisemblance avec un terme d'erreur distribué logistiquement auquel on ajoute un effet aléatoire dont la distribution est inconnue (Cameron et Taber, 1998). En se basant sur Cameron et Taber (1998) et en reprenant la notation du chapitre 2, nous avons :

$$y_{it} = \mathbf{1}(x_i'\beta + \epsilon_{it} + \alpha_t\theta_i \geq 0) \tag{3.1}$$

où $\mathbf{1}(\cdot)$ est la fonction indicatrice qui prend la valeur de 1 quand l'argument est vrai et qui prend celle de 0 sinon. Le vecteur x_i' représente les variables observables qui sont constantes d'une période à l'autre. Le même modèle peut prendre en compte des x_i variant dans le temps comme le font Cameron et Taber (1998). Cependant, nous nous en tenons à des variables constantes dans le temps en raison des limites de notre source de données. Le terme d'erreur ϵ_{it} est indépendant d'une période à l'autre et a une distribution logistique. Avec ces éléments, nous avons un simple modèle logit auquel nous ajoutons $\alpha_t\theta_i$ où θ_i est un terme inobservé qui est indépendant d'une observation à l'autre, mais qui est constant dans le temps. Dans notre contexte, il peut être interprété comme l'habileté ou les aptitudes innées (Cameron et Heckman, 1998). Sa distribution est inconnue et est estimée. Les seules hypothèses faites quant à la distribution sont que $E(\theta^2) < \infty$ et que F est indépendant de x_t et de ϵ_t pour tout t. Finalement, le terme α_t est le coefficient de l'effet de θ_i et $\alpha_1 = 1$ pour des fins d'identification. Notons, que si $\alpha_{t1} = 1$ pour tout t, alors il s'agit d'un simple modèle à effet aléatoire.

Pour estimer les β de ce modèle logit avec des effets aléatoires, nous avons donc un estimateur non-paramétrique de maximum de vraisemblance (NPMLE) pour lequel la fonction de maximum de vraisemblance pour un individu prend la forme suivante :

$$L = \int_{t=1}^{T} \left\{ F(x\beta + \alpha_t\theta)^{y_t}(1 - F(x\beta + \alpha_t\theta)) \right\} dH(\theta) \tag{3.2}$$

où $H(\theta)$ est la distribution inconnue de θ et où $F(\cdot)$ est la fonction de distribution logistique. Bien que la maximisation de cette fonction nous permette d'obtenir des estimés consistants, pratiquement, nous maximisons une version discrète dont nous prenons le logarithme népérien (Cameron et Taber, 1998). Nous obtenons une approximation de la distribution $H(\theta)$ à l'aide d'un nombre fini, mais non déterminé de points de support. Cette distribution que nous nommons $G(\theta_c)$ comporte C points. Cameron et Taber (1998) expliquent clairement pourquoi nous pouvons estimer les paramètres avec un nombre fini de points. Lors d'une estimation, nous avons n observations qui peuvent chacune avoir une valeur de θ différente. Empiriquement, la distribution de θ sera donc discrète avec un maximum de n points de support. Nous avons donc $C \leq n$, ce qui nous permet d'estimer θ_c conjointement avec (β et α). Cependant, nous verrons qu'un nombre restreint de points de support est suffisant. Avec ces éléments, nous pouvons écrire la contribution d'un individu à la fonction de log-vraisemblance que nous maximisons :

$$L = \sum_{c=1}^{C} \left\{ \prod_{t=1}^{T} F(x_i\beta + \alpha_t f_c)^{y_t}(1 - F(x_i\beta + \alpha_t f_c)) \right\} g_c \tag{3.3}$$

Nous estimons $\beta, \alpha_t, f_c, g_c$ et C. Dans cette fonction f_c est un point de support de la fonction $G(f_c)$ et g_c est la probabilité associée à chaque point. Cette probabilité est non-négative et elle somme à 1 afin que G soit une fonction de distribution. Nous verrons à la fin de ce chapitre comment nous estimons ce modèle, alors que dans la prochaine section nous présentons brièvement les conditions d'identification de ce modèle.

3.2 Un modèle économique justifiant le NPMLE et les conditions d'identification de ceux-ci

Nous résumons ici brièvement le modèle économique, qui justifie le NPMLE, développé par Cameron et Heckman (1998) ainsi que les conditions d'identification présentées par ceux-ci. Nous nous contenterons d'en esquisser les grandes lignes. Le lecteur qui désirerait approfondir la question et obtenir les preuves associées pourra se référer à l'article.

Dans cet article, les auteurs présentent un modèle micro-économique duquel il est possible de dériver le modèle statistique de transtion et le NPMLE. De leur modèle, ils obtiennent sous quelle condition un agent poursuit ses études jusqu'au niveau s. Cette condition est la suivante :

$$R(s) - c(s)\varphi(x)\epsilon v_s \geq R(s-1) - c(s-1)\varphi(x)\epsilon v_{s-1} \tag{3.4}$$

où $c(s)$ est le coût de l'éducation, $R(s)$ est le rendement actualisé de l'éducation, $\varphi(x)$ est un facteur de coût qui dépend des caractéristiques individuelles observées, alors que ϵ est un facteur individuel de coût qui est connu de l'individu, mais pas nécéssairement du chercheur et v_s est un choc de prix spécifique à la transition. L'intuition derrière cette règle est qu'un agent poursuit ses études si la valeur future de l'éducation à laquelle on soustrait les coûts s'accroit entre les niveaux d'éducation. En d'autres mots, si la différence des revenus moins les coûts en s est supérieure aux revenus moins les coûts de l'éducation en $s-1$, l'agent va compléter le niveau s d'éducation.

Cette règle de continuité ne mène pas toujours à un optimum. En effet, si les individus connaissent tous les v futurs, ils peuvent payer un v_s élevé pour se rendre en $s+1$ si v_{s+1} est bas afin d'atteindre $s+2$. Afin que cette règle de continuité fonctionne et que nous retrouvions le modèle de transition présenté au chapitre 2, ainsi que sa version avec les facteurs inobservés, nous devons postuler que les agents sont myopes (Cameron et Heckman, 1998). Cette hypothèse est moins restrictive qu'elle peut paraitre à première vue, puisque la myopie s'applique seulement aux $v_{s+1 \ldots S}$. Les agents connaissent donc les coûts de l'éducation future $c(s)$, mais pas les chocs de prix futurs. Il s'agit donc d'une

hypothèse, bien que restrictive, acceptable dans le cas présent. Nous pouvons réécrire l'équation 3.4 pour introduire les régresseurs :

$$\frac{R(s) - R(s-1)}{\varphi(x)[c(s|x) - c(s-1|x)]} \geq \epsilon v_{s-1} \tag{3.5}$$

En laissant $\varphi(x) = exp(-x\beta)$ et v_{s-1} être distribué logistiquement, nous obtenons le modèle de transition avec de l'hétérogénéité inobservée présenté au chapitre 2. Nous avons donc une probabilité d'atteindre le niveau $s+1$ quand D_{s+1} est égal à :

$$Pr(D_{s+1} = 1|D_s = 1, X = x, \omega) = \frac{exp\left[\frac{\ell(x)+x\beta+\omega}{\sigma_s}\right]}{1 + exp\left[\frac{\ell(x)+x\beta+\omega}{\sigma_s}\right]} \tag{3.6}$$

où $\omega = -log\epsilon$ et où $\ell(x)$ est le retour marginal par rapport aux coûts marginaux et où :

$$exp(\ell(x)) = \frac{R(s+1) - R(s)}{c(s+1) - c(s)}, s = 1, \ldots, \bar{S} \tag{3.7}$$

Ce modèle est identique au modèle présenté au chapitre 2 lorsque $\omega = \Theta$, $\ell(s)$ est l'intercepte et x ne comprend pas de constante. La seule différence est que dans cette version les β sont fixes entre les transitions, mais ceci n'est pas une limitation de ce modèle économique. À l'aide de ce modèle, nous rapportons les conditions d'identification présentées par Cameron et Heckman (1998).

Théorème 1

En utilisant le modèle (3.4) où $\log c(s|x) = \log c(s) - x_s\beta_s + \eta_s, s = 1, \ldots, \overline{S}$ et une distribution de $(\eta_1, \ldots, \eta_{\overline{S}})$ représentée par $F_\eta(\cdot)$ avec un support (Supp)$\prod_{i=1}^{\overline{S}}(L_i, U_i)$ où $L_i < \eta_i < U_i$ et U_i et L_i peuvent ne pas être bornés. En laissant $\ell(s)$ être l'intercepte pour la transition s. Les X ne comprennent pas de constante. En tenant compte de cela :

1. $F_\eta(\cdot)$ est absolument continue et n'est pas une fonction de X.

2. $(\eta_1, \ldots, \eta_{\overline{S}})$ est statistiquement indépendant de $(X_1, \ldots, X_{\overline{S}})$.
3. $X_S = X$ pour tout $S = 1, \ldots, \overline{S}$ est une variable aléatoire de dimension K et est de plein rang.
4. Les $m \geq 1$ premières coordonnées de X sont des variables continues ($m \leq K$).
5. $\beta_1, \ldots, \beta_m$ sont linéairement indépendants.
6. $\text{Supp}(x_S\beta_S + \ell(s) | x_{S-1}\beta_{s-1} + \ell(s-1), \ldots, x_1\beta_1 + \ell(1)) = (L_s, U_s)$ pour $s = 2, \ldots, m$, et $\text{Supp}(x_1\beta_1 + \ell(1)) = (L_1, U_1)$,

alors $F_\eta(\cdot)$ est non-paramétriquement identifiée dans ses $Min(m, \overline{s})$ premiers arguments.

Ce théorème montre que lorsque les X sont communs d'une période à l'autre le théorème est quand même identifié sous certaines conditions. Les trois conditions principales sont qu'au moins une variable de X soit continue, que les $\beta_1, ..., \beta_m$ (où m est le nombre de variables continues comprises dans X) doivent être linéairement indépendants et finalement qu'il y ait de la variation entre les $X_s\beta_s$. Les auteurs mentionnent que le manque de variation dans X entre les transitions posent de graves problèmes d'identification si le nombre de variables continues est faible comparé au nombre de variables non continues.

Les auteurs présentent le corollaire de non-identification suivant :

Corollaire du théorème 1

Sous les conditions du théorème 1, si les conditions 5 et 6 ne sont pas satisfaites, car $X_1 = X_2 = \ldots = X_{\overline{S}} = X$ et que $\beta_1 = \beta_2 = \ldots = \beta_{\overline{S}} = \beta$, mais que les autres conditions sont satisfaites, alors $F(\eta_1, \ldots, \eta_{\overline{S}})$ n'est pas identifié.

Dans notre cas, les X ne varient pas entre les transitions et l'identification de notre modèle dépend de la condition de non-colinéarité des β. Malheureusement, comme le mentionnent les auteurs, nous ne pouvons tester l'hypothèse d'indépendance linéaire quand les X ne varient pas dans le temps.

Les auteurs proposent deux autres théorèmes qui assurent l'identification du modèle de transition logistique avec des termes inobservés modélisés à l'aide d'une distribution

discrète (NPMLE).

Théorème 2

Pour le modèle de l'équation (3.5) où $\log c(s|x) = \log c(s) - x_s\beta_s + \eta_s$, nous assumons que :

1. $\eta_s = \alpha_s v + U_s, s = 1, \ldots, \overline{S}$, où α_s est un facteur de chargement pour v avec $\alpha_1 = 1, |\alpha_s| < \infty, s = 1, \ldots, \overline{S}$.
2. $(U_q, \ldots, U_{\overline{S}})$ sont mutuellement indépendants et identiquement distribués avec une médiane de 0, une variance connue et une fonction de distribution log-concave connue F.
3. v est indépendant de $(U_q, \ldots, U_{\overline{S}})$.
4. $\text{Supp}(X\beta_S) = \overline{J}_S \subseteq R^1$, est un intervalle dans R^1.
5. $X \in \overline{\overline{X}} \subseteq R^K$ ne réside pas dans un sous-espace propre.
6. v est multinominal avec un nombre inconnu, mais fini de nombre de points de masse $(= I)$ (La localisation et la probabilité des masses ne sont pas connues).

Les conditions 1,2 et 3 divisent le terme d'erreur en deux parties. Premièrement, la partie v qui est fixe entre les transitions et qui est le terme d'hétérogénéité inobservée. Cette partie a un effet différent d'une transition à l'autre qui est représentée par le terme α_S. La deuxième partie est $(U_q, \ldots, U_{\overline{S}})$ et est simplement un terme d'erreur qui, dans notre cas, a une fonction de distribution logistique et est indépendant de v. La condition 6 décrit la distribution du facteur inobservé, celle-ci étant une distribution discrète avec un nombre de points inconnu, mais fini. Ces différents éléments sont identiques à ceux du NPMLE défini dans le chapitre 2.

Avec ce théorème, si nous fixons α_s pour tout s où $\beta_s = \beta$ pour tout s, ou encore en fixant certains β_s pour des s particuliers en laissant les α_s varier, nous pouvons estimer $\alpha_s, \beta_s, l(s)$ et la distribution de v et ses paramètres de localisation. Notons que ce théorème implique que nous ne puissions connaître pour un s donné α_s et β_s (Cameron et Heckman, 1998). Cependant, le prochain théorème montre qu'il est possible de connaître

les deux lorsque X contient suffisamment de variation.

Théorème 3

Définissons $\overline{I}_s$ =Supp(η_s), le support de η. Sous les conditions du théorème avec une condition 4 revue pour lire 4' : $\overline{I}_s \subseteq$ Supp($X\beta_s$) et $X\beta_s$ suppose soit des valeurs arbitrairement grandes ou arbitrairement petites ou les deux et 6' : $Max_i|v_i| < \infty$. Alors $\alpha_s, \beta_s, l(s)$ et la distribution de v sont identifiés.

En résumé, les théorèmes précédents nous montrent que le modèle de transition avec un facteur d'hétérogénéité non-observée est identifié sous certaines hypothèses. Premièrement, sous les conditions du théorème 1, particulièrement celle de l'indépendance des β_s, le modèle est identifié non paramétriquement. Cependant, cette hypothèse d'indépendance ne peut être testée. Le théorème 2 nous montre que le modèle de transition est identifié, même en relâchant l'hypothèse de non-colinéarité des β_s, si l'on impose une forme au terme d'erreur et que l'on impose certaines restrictions aux coefficients α_s et β_s. Finalement, le théorème 3 nous montre que nous n'avons pas besoin d'imposer de restriction aux coefficients α_s et β_s si X est suffisamment grand.

3.3 L'estimation du NPMLE

L'estimation du modèle de sélection sans correction pour l'hétérogénéité se fait simplement à l'aide des fonctions standard des différents logiciels statistiques tel que Stata. Cependant, à notre connaissance, il n'existe pas de commande standard permettant d'estimer un modèle avec une correction pour l'hétérogénéité tel que décrit par Cameron et Taber (1998). Dans un premier temps, nous avons étudié les méthodes de calcul numérique permettant d'estimer une fonction de vraisemblance telle que celle du NPLME. Une des méthodes possibles est l'utilisation d'un algorithme de type *expectation maximization* (EM) comme le mentionnent Cameron et Taber (1998). Il est à noter qu'en raison de problèmes de vitesse de convergence de l'algorithme EM, Cameron et Taber (1998) ont utilisé un algorithme plus rapide et plus commun soit un algorithme quasi-Newton. Cependant, comme la programmation de tels algorithmes pour une fonction avec une

composante non paramétrique sort du champ de ce mémoire, une solution alternative est utilisée.

Nos recherches pour la revue de littérature nous ont conduits à finalement utiliser une commande Stata nommée GLLAMM pour *Generalized Linear Latent And Mixed Models* (Rabe-Hesketh, Skrondal et Pickles, 2005; Rabe-Hesketh, Skrondal et Pickles, 2004). En effet, dans son article Karlson (2011) utilise la commande GLLAMM pour estimer un modèle NPLME. De plus, il nous a été possible, en contactant l'auteur, d'obtenir le code utilisé pour cet article et de nous en inspirer. Les GLLAMM comprennent, comme cas spécial, les modèles à classe latente en plus de permettre une distribution discrète de la classe latente et aussi non paramétrique, permettant l'estimation du NPLME.

Quelques ajustements ont dû être apportés afin de faire fonctionner la commande pour estimer des β différents pour chaque période. Nous avons finalement créé une ligne par période et répliqué chaque variable en trois exemplaires (une par période). Nous avons attribué à chaque exemplaire la valeur 0 pour les deux observations qui n'étaient pas associées à la même période que la variable (Kolenikov, sd). À terme, nous avons obtenu un fichier de données ressemblant au tableau 3.1.

Tableau 3.1 Arrangement des données

Observation	Période	x1_1	x1_2	x1_3	x2_1	x2_2	x2_3
1	1	1	0	0	1	0	0
1	2	0	1	0	0	1	0
1	3	0	0	1	0	0	1
2	1	1	0	0	1	0	0
2	2	0	1	0	0	1	0
2	3	0	0	1	0	0	1

3.4 Les effets marginaux et le calcul des écarts-types

L'interprétation des coefficients estimés à l'aide du NPLME, tout comme ceux estimés avec le modèle logit, est peu intéressante. En effet, ce que nous recherchons c'est l'effet marginal d'une variable soit : $\delta E[y|x]/\delta x$. Alors que pour le modèle linéaire nous avons : $\delta E[y|x]/\delta x = \beta$, pour un modèle logit nous avons $\delta E[y|x]/\delta x = \Lambda(x\beta)/(1 - \Lambda(x\beta))\beta$. En raison de la non-linéarité, il n'existe pas d'effet marginal unique et en l'absence de consensus sur la méthode optimale de présentation, nous avons décidé de présenter les résultats sous la forme des effets marginaux moyens, soit $\overline{m}_x = 1/N \sum_i m_x(x_i)$ où i représente ici un individu et où $m_x(x_i)$ est l'effet marginal de x_i pour l'individu i.

Dans le chapitre suivant, nous présentons les données utilisées et nous verrons qu'en raison du plan complexe de sondage de Statistique Canada, l'agence responsable de l'enquête utilisée recommande l'utilisation d'écarts-types calculés à l'aide de la méthode bootstrap. Cette méthode consiste à effectuer un ré-échantillonnage aléatoire avec remplacement à même l'échantillon d'origine et ensuite de ré-estimer le modèle avec le nouvel échantillon. Nous effectuons ces deux étapes R fois. Les nouveaux échantillons sont habituellement de la même taille que l'estimateur d'origine. Par la suite, nous pouvons calculer l'écart-type des estimations ré-échantillonnées, ce qui nous fournit les écarts-types de notre modèle. Puisque la commande GLLAMM ne s'intègre pas à la commande bootstrap de Stata, nous avons dû écrire un programme permettant d'effectuer le calcul d'écart-type bootstrap. Le travail est facilité par le fait que Statistique Canada nous fournit 1000 poids ré-échantillonnés et il nous suffit donc d'effectuer l'étape d'estimation à répétition et de calculer l'écart-type des résultats.

La méthode bootstrap est très intensive en calcul puisqu'il faut ré-estimer le modèle R fois. Notre modèle est lui-même intensif et en raison de limitations de la puissance de calcul disponible, nous avons dû restreindre le calcul des écarts-types à la première spécification. Pour les analyses de sensibilité, nous nous sommes restreints à calculer les écarts-types des effets marginaux avec la méthode delta.

3.5 Simulation

La commande GLLAMM et la théorie derrière sont des outils très puissants permettant l'estimation de vastes champs de modèles. La compréhension approfondie de son fonctionnement dépasse le cadre de ce mémoire. Cependant, nous avons voulu nous assurer que les résultats obtenus étaient conformes à nos attentes. En effet, la combinaison d'un modèle complexe à estimer numériquement à l'utilisation d'une commande complexe ainsi que la nécessité de compter sur une réorganisation des données, rendait possible la génération de résultats incohérents. Afin de vérifier le fonctionnement de la commande et afin d'évaluer les possibilités de correction du biais à l'aide d'un estimateur NPMLE, nous avons effectué une simulation basée sur celles présentées au chapitre précédent.

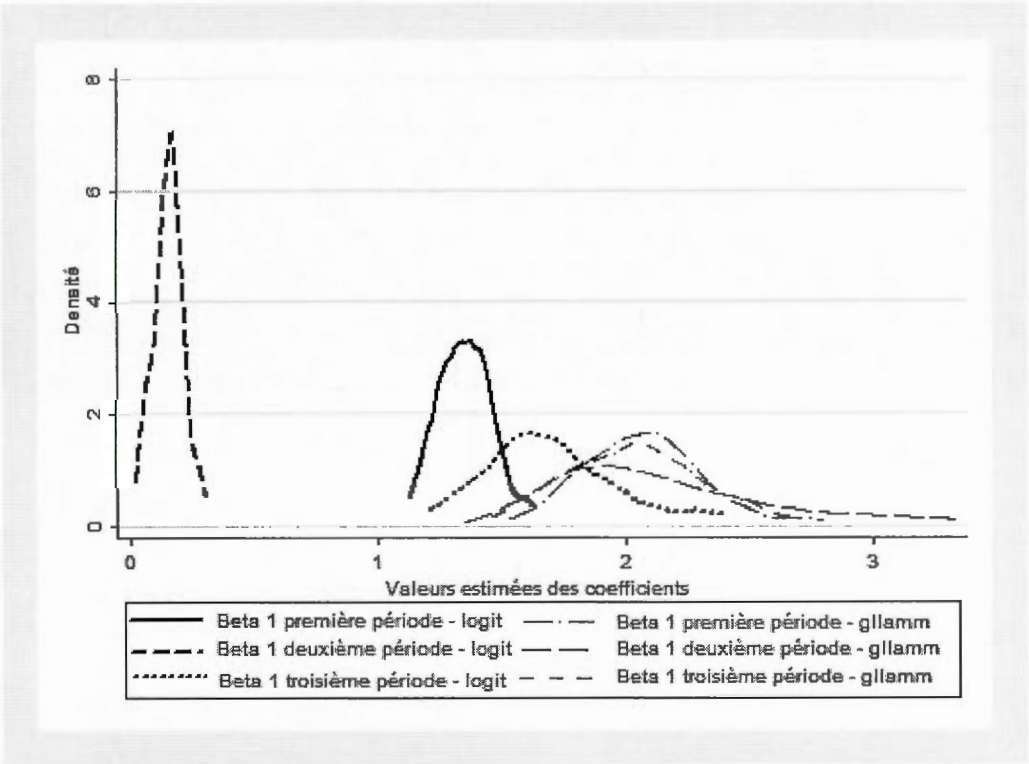
Cette simulation est construite de la même manière que celle du chapitre 2, c'est-à-dire que nous générons les variables de la même manière que précédemment, à l'exception que nous avons maintenant deux variables observées à la place d'une seule. Notons que nous avons toujours une variable inobservée. De plus, le terme inobservé a maintenant deux points de support, soit -2 et 2 et il a toujours une moyenne de zéro. Un autre changement est que les facteurs de chargements (coefficient de l'effet aléatoire) ont maintenant des valeurs différentes ($\alpha_1 = 1, \alpha_2 = 4, \alpha_3 = 2$). Nous effectuons 100 répétitions et nous gardons seulement celles pour lesquelles les commandes GLLAMM et logit ont convergé [2] (91/100). Chaque répétition comprend une estimation par logit et une estimation avec GLLAMM. La figure 3.1 présente la distribution du premier coefficient pour les 91 répétitions, alors que le tableau 3.2 présente la moyenne ainsi que l'écart-type des β en plus de la moyenne des facteurs de chargements.

---

2. La convergence est atteinte lorsque la différence entre deux intérations de la valeur de la fonction vraisemblance est inférieure au critère d'arrêt du programme (logit et GLLAMM). À ce moment, le programme considère qu'il a rencontré le maximum de la fonction et donc s'arrête. Il peut cependant s'agir d'un maximum global ou d'un maximum local. De plus, en raison du temps de calcul, nous avons limité le nombre d'itérations à 50, et si après ces itérations la différence entre les deux dernières itérations est encore supérieure au critère d'arrêt alors il n'y a pas convergence.

La figure 3.1 et le tableau 3.2 montrent qu'en présence d'hétérogénéité inobservée, l'estimation par estimateur logistique est biaisée, alors que celle effectuée par NPMLE est très proche de la vraie valeur. Visuellement, la figure 3.1 nous montre que toutes les

Figure 3.1 Simulation avec la commande GLLAMM



estimations ont une distribution ressemblant à une normale, mais que les coefficients estimés à l'aide d'un estimateur logit ne sont pas centrés sur la vraie valeur, alors que ceux estimés par NPMLE sont centrés sur une valeur très proche de la vraie valeur. Cependant, la cloche très large de l'estimateur NPMLE montre une variance beaucoup plus élevée que pour l'estimateur logit, qui lui a une cloche très mince pour les deux premières transitions. Le tableau 3.2 nous montre quant à lui que les moyennes des estimations sont très éloignées de la vraie valeur dans le cas de l'estimateur logistique, alors qu'elles sont très rapprochées de la vraie valeur dans le cas du NPMLE. En effet, pour la première transition, les moyennes des coefficients estimés par l'estimateur logistique sont de 1,35 et 3,38 et par le NPMLE leur moyenne est 2,07 et 5,16, alors que les vraies

Tableau 3.2 Simulation de la correction du biais d'hétérogénéité inobservée par NPMLE

Variables	Vraies valeurs	Logit	NPMLE
Beta 1 - transition 1	2	1,35	2,07
		(0,11)	(0,24)
Beta 1 - transition 2	2	0,16	2,15
		(0,06)	(0,43)
Beta 1 - transition 3	2	1,70	2,03
		(0,27)	(0,26)
Beta 2 - transition 1	5	3,38	5,16
		(0,18)	(0,56)
Beta 2 - transition 2	5	0,40	5,28
		(0,05)	(0,89)
Beta 2 - transition 3	8	7,05	8,23
		(0,77)	(0,81)
Facteur de chargement 1		-	1
Facteur de chargement 2		-	4,28
Facteur de chargement 3		-	1,718

Source : simulation de l'auteur. Les données sont générées à partir d'une loi normale et le terme inobservé est tiré d'une loi binomiale avec deux points de support à -2 et 2 avec une moyenne de zéro.

Note :

- Écart-type des estimations entre parenthèses
- Les coefficients présentés sont la moyenne des coefficients obtenus lors des 91 itérations convergentes de la simulation

valeurs des coefficients sont de 2 et 5. L'écart est encore plus marqué pour la deuxième transition, les coefficients estimés par l'estimateur logit étant de 0,16 et 0,40 et ceux estimés par le NPMLE étant de 2,15 et 5,28, alors que les coefficients réels sont toujours 2 et 5. Pour la dernière transition, l'écart diminue, mais reste important. L'estimation par NPMLE nous permet aussi d'estimer la valeur des facteurs de chargements, soit l'effet variant entre les périodes du facteur d'hétérogénéité inobservée. Le premier facteur de chargements est arbitrairement fixé à un et pour les autres facteurs de chargements, l'estimation donne des résultats satisfaisants avec une moyenne des valeurs estimées de 4,28 et 1,718 pour les deuxième et troisième facteurs de chargements, alors que les vraies valeurs sont de 4 et 2. Nous voyons donc avec cette simulation que l'utilisation du NPMLE calculé à l'aide de GLLAMM permet nettement de diminuer le biais de l'estimation dans

un cadre précis de simulation. Deux réserves méritent d'être mentionnées. Premièrement, les écarts-types sont beaucoup plus importants avec le NPMLE qu'avec le modèle logit. Deuxièmement, ce mémoire ne portant pas principalement sur l'évaluation des qualités du NPMLE à l'aide de simulation, nous nous sommes limités à une simulation dans un cadre bien précis et il est possible qu'avec, entre autres, d'autres distributions de facteurs inobservés, la diminution du biais ne soit pas aussi importante.

CHAPITRE IV

LA DESCRIPTION DES DONNÉES

Dans ce chapitre, nous présentons, dans un premier temps, la source de données utilisée, soit l'Enquête auprès des jeunes en transition (Statistique Canada, 2011a). Nous décrivons ses principales caractéristiques, abordons la question de l'attrition et présentons ses principaux avantages et inconvénients. Dans un deuxième temps, nous décrivons les critères utilisés dans le choix de l'échantillon. Ces critères sont multiples, d'une part géographiques, liés aux structures des systèmes d'éducation provinciale, d'une autre part, basés sur la disponibilité des données en lien avec la structure de l'enquête. Les motivations justifiant ces choix sont aussi présentées. Finalement, nous présentons les variables utilisées, le processus justifiant leur sélection et quelques statistiques les décrivant.

4.1 L'Enquête auprès des jeunes en transition

Comme mentionné dans l'introduction, nous utilisons le fichier de micro-données de l'Enquête auprès des jeunes en transition (Statistique Canada, 2011a) qui contient un ensemble de données répondant à nos besoins. Celle-ci est une enquête longitudinale élaborée par Statistique Canada. Elle suit deux cohortes, une première de jeunes nés entre 1979 et 1981 inclusivement et une deuxième comprenant des jeunes nés en 1984 qui fréquentaient tous les types d'établissements scolaires à l'exception des établissements dans les réserves autochtones et les établissements pour les étudiants avec des besoins spéciaux (Statistique Canada, 2011b). De plus, les deux cohortes se limitent aux jeunes vivant dans l'une des 10 provinces. La cohorte des jeunes ayant 15 ans en 1999 est

nommée cohorte A ou de lecture, alors que celle comprenant des jeunes ayant entre 18 et 20 ans en 1999 est appelée cohorte B. La taille initiale de la cohorte de lecture était de 38 000 étudiants, alors que la taille de la cohorte B était de 29 000 personnes. À terme, c'est 29 700 jeunes de la cohorte A et 23 000 de la cohorte B qui ont effectivement répondu au premier cycle de l'enquête. L'étude est divisée en cycles de 2 ans. Les personnes sont interrogées à la fin de chaque cycle. Seules les personnes ayant répondu au cycle précédent sont recontactées lors des cycles suivants. Pour le cycle 5, les jeunes ont été enquêtés en 2008. L'échantillon de ce cycle était de 18 762 jeunes de 23 ans pour la cohorte A et 12 360 jeunes de 26 à 28 ans pour la cohorte B. Pour diverses raisons que nous présentons plus bas, nous avons seulement utilisé les données de la cohorte A.

L'EJET est une enquête à participation volontaire et le taux de réponse varie autour de 80 % pour chacune des cohortes à chacun des cycles, ce qui explique l'attrition élevée après 5 cycles. Le taux de non-réponse est principalement dû à l'impossibilité de recontacter les personnes participantes à l'enquête en raison de la mobilité élevée de la population cible (Statistique Canada, 2011b). Dans les questionnaires remplis par les participants, certaines données sont manquantes ou aberrantes, Statistique Canada procède donc à l'imputation des données quantitatives. Il s'agit, dans un premier temps, d'une imputation à l'intérieur des réponses de la personne en dérivant les réponses à l'aide des autres réponses. Dans un deuxième temps, Statistique Canada utilise une technique appelée du « plus proche voisin » (Statistique Canada, 2011b). De plus, Statistique Canada a calculé des poids permettant d'obtenir un échantillon représentatif de la population canadienne.

Cette enquête est fort intéressante en raison du nombre important de variables mesurées. Elle fournit de l'information détaillée sur le parcours d'études ainsi que sur l'environnement socio-économique des jeunes. Une des caractéristiques importantes pour notre étude est la collecte d'informations sur les parents à l'aide d'un questionnaire directement soumis aux parents. En effet, lors du cycle 1, les parents des jeunes de la cohorte A se sont vus questionner sur différents aspects, dont leur revenu et leur niveau d'études. Cette particularité permet d'obtenir une information de première main, qui est donc

plus fiable. Cependant, ces données comportent deux désavantages. Premièrement, c'est seulement lors du premier cycle d'enquête que des questions portant sur les parents ont été posées. Ceci limite la portée des données sur le revenu des parents, celui-ci variant dans le temps et l'année de l'enquête n'étant peut-être pas représentative du revenu sur le cycle de vie. Deuxièmement, seuls les jeunes de la cohorte A ont vu leurs parents interrogés, ce qui a contribué à notre choix d'utiliser uniquement cette cohorte. Finalement, cette enquête nous fournit aussi l'information nécessaire pour établir quels sont les niveaux d'études seulement fréquentés et ceux complétés permettant d'avoir un portrait juste du parcours scolaire.

4.2 La sélection de l'échantillon

La sélection de l'échantillon est effectuée à partir de trois critères. Premièrement, les informations sur l'environnement socio-économique des jeunes doivent être les plus complètes possible. Comme mentionné plus haut, nous nous restreignons donc à la cohorte A, puisque les parents des jeunes de cette cohorte ont répondu à un questionnaire et que seule cette cohorte a répondu à l'enquête PISA. Ce choix réduit donc *de facto* la population à l'étude de 52 700 à 29 700 jeunes. De plus, l'information sur le milieu socio-économique étant au coeur de notre analyse, seuls les jeunes dont les parents, biologiques ou adoptifs, ont rempli le questionnaire sont retenus réduisant la population à 25 662 jeunes.

Deuxièmement, la population étudiée doit provenir d'un même système d'éducation afin que la série de transitions possibles soit cohérente. Le système d'éducation relevant des provinces au Canada, nous étudions donc seulement les jeunes provenant et étudiant au Québec. L'étude combinée des différentes provinces est difficile en raison des spécificités du modèle d'éducation québécois qui comprend un diplôme d'études collégiales obtenu dans un cégep. Ce diplôme, préalable à l'université, n'existe pas dans les autres provinces du Canada, ce qui rend l'étude des déterminants socio-économiques des transitions entre les niveaux d'éducation difficile lorsque l'on prend une population provenant de l'ensemble du Canada. Ce choix, de réduire l'étude aux seuls jeunes du Québec fait

passer le nombre d'observations à 4450. Nous éliminons aussi les observations n'ayant pas d'informations pour les variables utilisées par notre modèle, ces restrictions affectent sensiblement la taille de l'échantillon et le réduit à 3578 observations. La diminution est importante. Cependant, une vérification montre que ces jeunes dont l'information manque sont majoritairement des non-répondants dans les cycles suivants. Finalement, les jeunes doivent être observés suffisamment longtemps pour s'assurer d'avoir un portrait complet de leur parcours scolaire. Puisque la dernière transition étudiée est celle de l'inscription à l'université, il est important d'aller le plus loin possible, au prix de perdre plusieurs observations. Afin de choisir le nombre de cycles nécessaires, le tableau 4.1 présente le taux de personnes dans chaque transition. Nous avons donc le taux de personnes avec un diplôme d'études secondaires, le taux d'inscrits au niveau collégial et le taux d'inscrits au niveau universitaire. Alors que le taux de diplômés du secondaire et d'inscrits au cégep est stable, celui du taux d'inscrits à l'université augmente fortement jusqu'au cycle 5. Les données étant plutôt stables entre le cycle 5 et 6, nous avons gardé seulement les jeunes qui avaient répondu aux 5 premiers cycles, soit jusqu'à leurs 23 ans. Le nombre d'observations perdues au cycle 6 était trop important pour le garder. Cependant, dans le chapitre suivant, nous présentons des résultats avec le cycle 6, principalement pour montrer un scénario avec des transitions jusqu'à l'obtention d'un diplôme de l'université. Bien que nous retenons seulement les jeunes ayant répondu aux 5 premiers cycles, une exception est faite pour les jeunes qui n'ont pas complété leur 5^e secondaire au cycle 3. Pour ceux-ci, nous nous contentons des trois premiers cycles puisque nous considérons que les jeunes n'ayant pas diplômé du secondaire à 21 ans ne l'obtiendraient pas dans des proportions significatives dans le futur et que les jeunes dont le taux d'attrition est le plus élevé est celui des décrocheurs. Avec ces critères, nous obtenons un échantillon final de 2084 jeunes.

4.3 La préparation des données, le choix des variables utilisées et les statistiques descriptives

Dans cette section, nous présentons, dans un premier temps, les manipulations effectuées afin d'obtenir un fichier de données utilisable pour l'estimation à l'aide de notre modèle NPMLE. Dans un deuxième temps, nous décrivons les variables utilisées et les manipulations effectuées pour créer ces variables. Dans un troisième temps, nous présentons les statistiques descriptives des différentes variables utilisées.

4.3.1 La préparation des données

Les données concernant l'EJET sont enregistrées par cycle et par type de données. Afin de pouvoir effectuer une estimation par NPMLE, de nombreuses manipulations sont nécessaires. Nous avons besoin de regrouper l'ensemble des informations dans un seul fichier pour effectuer l'estimation. Le principal défi provient de l'identification des différents niveaux d'études fréquentés par les jeunes. En effet, un jeune peut avoir été inscrit à plusieurs programmes dans un même cycle, dans un ou plusieurs établissements, à un ou plusieurs niveaux d'études. Dans un premier temps, nous rassemblons l'information sur les programmes et les institutions fréquentées pour ensuite regrouper l'information de l'ensemble des cycles. Chaque programme pour lequel un individu a été inscrit est identifié par un numéro unique permettant le suivi entre les cycles. Lorsqu'un programme est suivi durant plus d'un cycle, seule l'information du dernier cycle est intéressante, nous informant si le programme a été complété ou non. Le numéro unique nous permet de garder l'information du dernier cycle seulement. À cette étape, quelques vérifications sont effectuées afin d'identifier des dates d'inscription incohérentes ainsi que d'autres problèmes mineurs d'incohérence. Lorsque des incohérences sont constatées, par exemple un programme terminant avant l'inscription, ce programme est écarté, mais l'individu est maintenu dans l'échantillon. Dans un deuxième temps, nous identifions dans quel niveau agrégé d'études [1] se retrouve chaque programme et nous gardons l'information sur

1. Pour le regroupement, voir la note 1 du tableau 4.1.

une seule ligne par individu. Par la suite, nous fusionnons cette information avec celle provenant du fichier de l'enquête auprès des parents et celui concernant les résultats au test de lecture PISA. Ces opérations sont répétées en s'arrêtant respectivement aux cycles 3, 4, 5 et 6 afin de pouvoir effectuer des tests de sensibilité au nombre de cycles inclus.

4.3.2 Le choix des variables

Nous utilisons un nombre restreint de variables en raison des difficultés d'estimation. En effet, puisque nous avons un modèle permettant d'évaluer des coefficients différents pour chaque transition, rapidement le nombre de coefficients devient important. Nous nous en tenons donc à un nombre réduit de variables clés largement utilisées dans la littérature. Il s'agit, en premier lieu, du niveau d'études des parents. Celui-ci est regroupé en quatre catégories : les parents sans diplôme d'études secondaires, ceux avec un diplôme d'études secondaires, ceux avec un diplôme d'études collégiales et, finalement, ceux avec un diplôme d'études universitaires. Cette variable est, dans la littérature, celle identifiée comme ayant le plus d'effet sur les différentes probabilités de transitions. Puisqu'il existe différents types de ménage et qu'il n'existe pas de mesure simple permettant de regrouper l'éducation de deux parents, nous avons retenu le niveau du plus haut diplôme par un ou l'autre des parents. Dans le cas de famille monoparentale, nous avons donc retenu le niveau d'éducation du parent responsable de la garde de l'enfant. Nous incluons aussi une variable de revenu, cette variable est le revenu total combiné du ménage, elle est la somme des neuf sources de revenus pour laquelle les répondants sont questionnés (Statistique Canada, 2011b). Comme cela est mentionné précédemment, cette mesure du revenu n'est pas idéale puisqu'elle est collectée seulement une fois, elle peut donc ne pas être représentative du revenu sur le cycle de vie. Cette mesure n'est donc peut-être pas la meilleure représentation possible des moyens financiers du ménage. Quatre autres variables sont utilisées, à savoir si le jeune est un homme ou une femme, le statut de minorité visible, la langue maternelle et si le jeune résidait dans une région rurale ou urbaine à 15 ans. Cette dernière variable sert de mandataire (« proxy ») pour l'accès

géographique aux établissements d'enseignement. Finalement, nous utilisons le score au test PISA de lecture. Les résultats de ce test sont présentés avec différentes valeurs possibles tirées d'une estimation par maximum de vraisemblance. Puisque l'inclusion de ces multiples valeurs possibles n'était pas réalisable dans le cadre de notre modèle NPMLE, nous avons seulement gardé l'estimateur WARM [2] du résultat au test PISA. Le choix de seulement prendre le résultat au test de lecture nous est dicté par l'enquête. En effet, seul le test de lecture a été fait par l'ensemble de la cohorte A, alors que les tests de mathématiques et en sciences ont été administrés à seulement un tiers de la cohorte et la perte des deux tiers de l'échantillon est bien trop importante. Il est à noter que l'utilisation du seul test de lecture pour représenter les habiletés n'est pas optimale, car ce test est bien sûr limité aux compétences de lecture. Cependant, comme nous l'avons mentionné dans le chapitre 1, peu de tests évaluant les aptitudes cognitives sont disponibles de pair avec des enquêtes de large envergure telle l'EJET. Les chercheurs doivent donc composer avec les données disponibles et dans ce cadre le test PISA est intéressant.

4.3.3 Les statistiques descriptives

Le tableau 4.2 présente les fréquences pondérées des différentes variables discrètes pour chacune des transitions. Ce que nous présentons, ce sont les fréquences des jeunes éligibles aux différentes transitions. Nous pouvons observer que le pourcentage de femmes augmente d'une transition à l'autre et que, de la même façon, le niveau de diplôme des parents augmente avec les transitions, ce qui est conséquent avec la littérature. Les autres variables sont stables et nous observons une légère baisse des jeunes provenant

---

2. Les scores proviennent de l'aggrégation des réponses aux différentes questions du test. Le score PISA est un estimé de la valeur des compétences sous-jacentes aux réponses aux différentes questions. Le score est estimé par maximum de vraisemblance et plusieurs valeurs possibles sont fournies. Alternativement, la pondération des différentes réponses permet l'estimation d'une seule valeur. Il s'agit d'un *weighted maximum likelihood estimate* (WLE) tel que suggéré par Thomas A. Warm (1989) d'où le nom d'estimateur WARM (OCDE, 2004; OCDE, 2005).

de régions rurales. Dans le tableau 4.3, nous observons une hausse du revenu total familial d'une transition à l'autre conformément à ce qui est observé dans la littérature. La variable du test de lecture PISA est présentée dans une forme centrée réduite. Nous observons, sans surprise, une forte hausse entre les transitions, résultats de la sélection des individus avec de plus grandes aptitudes. Ces résultats cohérents avec l'intuition et la littérature nous assurent une base solide pour effectuer la comparaison des analyses réalisées à l'aide du modèle de transition standard et celles réalisées à l'aide du NPMLE.

Tableau 4.1 Proportion de diplômés des études secondaires et pourcentage par cycle de jeunes ayant été inscrits au moins 1 mois - pondérés

Cycle	N	Variable[1]	Proportion[2]	Pourcentage d'inscrits[2]
3	2902	Pas de diplôme d'études secondaires	17,45%	-
		Diplôme d'études secondaires	82,55%	-
		Inscrit à des études collégiales	-	63,54%
		Inscrit à des études universitaires	-	10,71%
4	2455	Pas de diplôme d'études secondaires	12,04%	-
		Diplôme d'études secondaires	87,96%	-
		Inscrit à des études collégiales	-	62,84%
		Inscrit à des études universitaires	-	30,41%
5	2084	Pas de diplôme d'études secondaires	17,77%	-
		Diplôme d'études secondaires	82,23%	-
		Inscrit à des études collégiales	-	63,62%
		Inscrit à des études universitaires	-	37,32%
6	1658	Pas de diplôme d'études secondaires	7,59%	-
		Diplôme d'études secondaires	92,41%	-
		Inscrit à des études collégiales	-	64,64%
		Inscrit à des études universitaires	-	39,68%

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A, cycles 1 à 6.

1. Regroupé comme suit : **Pas de diplôme d'études secondaires** : études secondaires non terminées ou moins. **Diplôme d'études secondaires** : diplôme d'études secondaires ou l'équivalent ; études non terminées au collège, au Cégep ou à l'université (sans certificat, diplôme ni grade) ; certificat ou diplôme d'une école commerciale privée ou d'un institut de formation privé; **Diplôme d'études collégiales** : certificat ou diplôme d'un collège, d'un Cégep, d'une formation professionnelle ou de métiers, d'une formation d'apprentis, d'une école normale ou d'une école de sciences infirmières. **Diplômes d'études universitaires** : Certificat ou diplôme universitaire inférieur au baccalauréat; Baccalauréat universitaire, premier grade professionnel en médecine (M.D.), en dentisterie (D.D.S., D.M.D.), en médecine vétérinaire (D.M.V.), en droit (LL.B.), en optométrie (O.D.) ou en théologie (M.Th.); maîtrise; doctorat.
2. Les données utilisées sont pondérées.

Tableau 4.2 Statistiques descriptives des variables discrètes - pondérées

	Transition - obtention d'un diplôme d'études secondaires[1] (N = 2084)		Transition - inscription au niveau collégial[2] (N = 1835)		Transition - inscription au niveau universitaire[3] (N = 1454)	
Variables	**Fréquences**	**Pourcents**	**Fréquences**	**Pourcents**	**Fréquences**	**Pourcents**
Minorité visible						
Non	1961	94,22	1614	94,2	1243	93,73
Oui	120	5,78	99	5,8	83	6,27
Femme						
Non	1081	51,96	840	49,01	628	47,38
Oui	1000	48,04	874	50,99	698	52,62
Niveau d'études parent[4]						
Pas de diplôme - secondaire	249	11,94	139	8,13	92	6,96
Diplôme - secondaire	688	33,06	531	30,99	390	29,4
Diplôme - collégial	538	25,86	477	27,81	365	27,55
Diplôme - universitaire	606	29,13	567	33,07	478	36,09
Langue maternelle						
Autres langues	112	5,37	90	5,23	73	5,5
Anglais	171	8,2	154	8,98	129	9,69
Français	1799	86,43	1470	85,78	1124	84,81
Rural						
Non	1533	73,67	1275	74,37	1003	75,62
Oui	548	26,33	439	25,63	323	24,38

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 5.

1. Échantillion : l'ensemble des jeunes.
2. Échantillion : les jeunes ayant un diplôme d'études secondaires.
3. Échantillion : les jeunes inscrits au niveau collégial.
4. Même regroupement que dans le tableau 4.1.

Tableau 4.3 Statistiques descriptives des variables continues - pondérées

	Transition - obtention d'un diplôme d'études secondaires[1] (N = 2084)		Transition - inscription au niveau collégial[2] (N = 1835)		Transition - inscription au niveau universitaire[3] (N = 1454)	
	Moyennes	**Écarts-types**	**Moyennes**	**Écarts-types**	**Moyennes**	**Écarts-types**
Revenu du ménage (En milliers de $)	66,14	49,48	68,91	49,70	70,50	50,74
Score PISA au test de lecture[4]	0,00	1,00	0,23	0,85	0,33	0,82

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 5.

1. Échantillion : l'ensemble des jeunes.
2. Échantillion : les jeunes ayant un diplôme d'études secondaires.
3. Échantillion : les jeunes inscrits au niveau collégial.
4. Il s'agit de l'estimation WARM qui a été centrée réduite.

CHAPITRE V

RÉSULTATS

Nous retrouvons dans ce chapitre les résultats des estimations du NPMLE. Nous présentons six estimations afin d'évaluer la sensibilité des résultats aux choix des spécifications. Dans un premier temps, nous défendons le choix des six spécifications. Dans un deuxième temps, nous présentons les résultats obtenus en précisant les différences entre les résultats des estimations logistiques et celles effectuées à l'aide du NPMLE. Dans un dernier temps, nous analysons les résultats obtenus d'un point de vue des déterminants socio-économiques en analysant les faits saillants des estimations et leur conformité à la littérature existante et finalement nous évaluons l'utilisation du NPMLE dans un contexte d'estimation d'un modèle de transition.

5.1 Les spécifications

Dans cette section, nous présentons les six spécifications que nous avons retenues pour nos estimations. Ces différentes spécifications visent à cerner la sensibilité des résultats aux hypothèses de sélection de l'échantillon et aux différents types de sélection. Sauf indication contraire, chacune des spécifications comprend dix variables explicatives, soit le fait d'appartenir à une minorité, le genre du répondant, les trois dichotomiques pour le niveau d'études des parents, le revenu familial des parents et les deux dichotomiques pour la langue maternelle (l'anglais, les autres langues et le français comme référence) ainsi que la variable permettant d'identifier si l'individu provient d'une région rurale.

La première spécification est la plus simple du point de vue du mécanisme de sélection postulé. En effet, nous nous limitons à une sélection d'échantillons entre les transitions du diplôme d'études secondaires et l'inscription aux niveaux collégiaux, alors l'absence d'inscription au niveau collégial n'est pas considérée comme un frein à l'accès à l'université. En d'autres mots, les jeunes n'ayant pas de diplôme d'études secondaires sont exclus de l'échantillon pour l'inscription au niveau collégial, mais les jeunes n'étant pas inscrits au niveau collégial ne sont pas exclus de l'échantillon universitaire. Ce choix reflète la réalité, c'est-à-dire qu'il est possible d'accéder à l'université sans fréquenter un cégep, entre autres par une admission à 21 ans sur la base des compétences. Nous utilisons cette spécification, car elle reflète la présence de personnes inscrites à l'université n'ayant pas fréquenté le cégep. Notre échantillon comprend 87 de ces jeunes inscrits sur un total de 865 jeunes inscrits à l'université au cycle 5. Dans cette spécification, nous avons donc un échantillon de même taille à la deuxième et à la troisième transition. De plus, dans cette spécification, les écarts-types sont estimés de deux façons, premièrement, la commande GLLAMM estime automatiquement les écarts-types à l'aide de la méthode Delta et, deuxièmement, nous les estimons à l'aide de la méthode bootstrap telle que décrite à la section 3.4. Il s'agit de la seule spécification effectuée à l'aide de poids bootstrap en raison des temps de calcul très longs nécessaires à de multiples ré-estimations. Ceci permet de reconnaître l'effet du calcul des écarts-types à l'aide de la méthode bootstrap.

Dans la deuxième spécification, nous reprenons les critères de la première spécification, mais en imposant une sélection sur l'inscription au cégep : les jeunes ne s'étant pas inscrits au cégep sont considérés comme n'ayant pas accès à l'université et ne font pas partie de l'échantillon de la transition de l'inscription à l'université. Comme nous l'avons mentionné précédemment, ceci ne reflète pas complètement la réalité puisqu'il est possible d'accéder à l'université sans diplômer du cégep à certaines conditions. Il s'agit d'une hypothèse simplificatrice qui permet de cerner le cheminement le plus courant en supposant un parcours linéaire (Mare, 2011). De plus, dans notre cas, l'ajout d'une sélection reflète mieux l'esprit du NPMLE.

Dans la troisième spécification, nous reprenons le type de sélection de la deuxième spé-

cification, mais nous retirons la variable du score PISA des variables explicatives. Nous effectuons ce test de sensibilité, car comme cela est mentionné plus haut peu d'enquêtes canadiennes incluent cette variable de contrôle des aptitudes et il est donc intéressant de savoir si son absence biaise les résultats des autres déterminants. En mesurant la variation des autres déterminants en fonction de l'ajout ou du retrait du score PISA, cela permet d'informer le chercheur de l'importance ou non d'utiliser des bases de données qui incluent ce type de mesure. Dans la quatrième spécification, nous réintroduisons la variable du score PISA et nous considérons que les jeunes n'ayant pas obtenu un diplôme d'études secondaires à 21 ans, soit à la fin du cycle 3, n'en obtiennent pas par la suite. Nous appliquons cette restriction, car les jeunes ayant obtenu un diplôme d'études secondaires après leurs 21 ans sont très peu susceptibles de s'inscrire au cégep, puisque la majorité d'entre eux obtient un diplôme d'études secondaires de type professionnel. De plus, cette restriction fait diminuer le taux de diplomation du secondaire, taux qui est très élevé dans la cohorte A en comparaison des taux estimés entre autre par Statistique Canada (CIRANO, sd). Cette modification renforce l'identification du modèle en accentuant les différences de taille des échantillons entre la première et la deuxième transition et reproduit le processus de sélection que nous modélisons dans ce mémoire.

Dans la cinquième spécification, nous utilisons l'obtention d'un diplôme d'études collégiales et l'obtention d'un diplôme d'études universitaires comme variables dépendantes à la place de l'inscription à ces deux niveaux. Nous remplaçons donc les transitions liées à l'inscription au cégep et à l'université par les transitions de l'obtention d'un diplôme à ces deux niveaux. La sélection de l'échantillon se fait donc exclusivement par l'obtention d'un diplôme. Dans un premier temps, ce choix est plus intéressant que les précédents puisqu'il faut diplômer du niveau collégial et non pas simplement s'y être inscrit pour accéder à l'université. Dans un second temps, l'obtention d'un diplôme, tant d'un point de vue personnel que social, est plus intéressante que la simple inscription. Cependant, l'utilisation de la diplomation au niveau universitaire nous oblige à utiliser le cycle 6 afin de nous assurer de bien capter l'ensemble des diplômes universitaires obtenus, ce qui, comme expliqué précédemment, réduit grandement l'échantillon, le faisant passer

de 2084 à 1656 pour la première transition. Notons que cette spécification reprend la restriction sur les jeunes n'ayant pas de diplôme d'études secondaires utilisée à la quatrième spécification.

Dans la sixième spécification, nous effectuons une spécification complète en incluant cinq variables dépendantes, soit l'obtention d'un diplôme d'études secondaires ainsi que l'inscription et l'obtention d'un diplôme au cégep et à l'université. Cette spécification est celle qui représente mieux le processus de sélection qui s'effectue au cours du parcours scolaire, car l'inscription à un niveau est obligatoire pour diplômer de ce niveau et que seule l'inscription à un niveau n'est pas suffisante pour accéder au niveau supérieur, il faut en diplômer. L'ajout de transitions renforce théoriquement l'identification du modèle. Cependant, tout comme la spécification 5, l'utilisation obligatoire du cycle 6 réduit grandement l'échantillon et nuit à l'estimation.

5.2 Les résultats des estimations

Dans cette section, nous présentons les résultats des estimations des différentes spécifications exposées précédemment. Pour l'ensemble des spécifications, nous présentons côte à côte l'estimation de la spécification de l'estimateur logistique et l'estimation du NPMLE. Dans l'ensemble des cas, nous avons restreint la présentation des résultats aux effets marginaux moyens, car ceux-ci présentent plus d'intérêt que les simples coefficients. Pour la première spécification, nous présentons un premier tableau de même format que les autres spécifications et ensuite un tableau comprenant les coefficients estimés, leurs écarts-types calculés par GLLAMM ainsi que les écarts-types bootstrap.

Le premier constat à la lecture du tableau 5.1, qui présente les résultats de la première spécification, est le peu d'écart entre les effets marginaux estimés à l'aide de l'estimateur logistique et du NPMLE. En effet, la différence entre les effets marginaux est toujours incluse dans une distance d'un écart-type. Globalement, nous n'observons pas la baisse décrite dans la littérature des effets des déterminants socio-économiques lorsque l'on avance dans les transitions. Cette baisse nous ne l'observons ni pour l'estimateur logis-

Tableau 5.1 Résultats de la régression par NPMLE - Spécification 1

	Transitions					
	Obtention d'un diplome d'études secondaires (N = 2084)		Inscription au niveau collégial (N = 1954)		Inscription à l'université (N = 1954)	
Variables	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE
Minorité visible	0,048	0,010	0,061	0,057	0,124**	0,121**
	(0,051)	(0,038)	(0,058)	(0,057)	(0,052)	(0,053)
Revenu familial	0,00021	0,00033*	0,00028	0,00018	0,00103***	0,00099***
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)
Femme	0,035**	0,033***	0,040*	0,025	0,099***	0,095***
	(0,016)	(0,012)	(0,022)	(0,029)	(0,021)	(0,022)
Niveau d'études des parents[1]						
Pas de diplôme - secondaire	-0,062***	-0,043**	-0,070*	-0,038	-0,071	-0,056
	(0,018)	(0,018)	(0,037)	(0,048)	(0,044)	(0,047)
Diplôme - secondaire (réf.)	-	-	-	-	-	-
Diplôme - collégial	0,024	0,027*	0,046*	0,033	0,049*	0,046*
	(0,019)	(0,015)	(0,027)	(0,031)	(0,026)	(0,027)
Diplôme - universitaire	0,081***	0,068***	0,088***	0,059	0,203***	0,203***
	(0,029)	(0,023)	(0,028)	(0,040)	(0,026)	(0,027)
Score au test PISA	0,087***	0,093***	0,154***	0,107***	0,178***	0,170***
	(0,008)	(0,009)	(0,011)	(0,040)	(0,011)	(0,012)
Langues marternelles						
Autres langues	0,055*	0,063***	0,108***	0,080**	0,092***	0,093***
	(0,029)	(0,023)	(0,036)	(0,039)	(0,030)	(0,030)
Anglais	0,046	0,063*	0,094	0,052	0,103*	0,093*
	(0,034)	(0,035)	(0,058)	(0,077)	(0,053)	(0,054)
Français (réf.)	-	-	-	-	-	-
Rural	0,039**	0,016	0,016	0,014	-0,036	-0,038
	(0,018)	(0,015)	(0,025)	(0,025)	(0,025)	(0,026)

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 5.
Écarts-types entre parenthèses : * significatif à 10%; ** significatif à 5%; *** significatif à 1%.
Ce tableau présente les effets marginaux moyens.
1. Même regroupement que dans le tableau 4.1.

tique ni pour le NPMLE. La transition avec le plus d'effets marginaux significatifs est celle de l'inscription à l'université avec 8 effets marginaux significatifs. Nous observons un effet fort du diplôme des parents, lorsque le niveau d'études augmente, la probabilité de diplômer du niveau secondaire et de s'inscrire tant au niveau collégial qu'universitaire augmente. La vaste majorité de ces effets sont significatifs au moins à un seuil de 10 %. L'effet du score PISA est marqué et augmente avec les transitions. Avec un score PISA d'un écart-type supérieur à la moyenne, la probabilité de s'inscrire à l'université augmente de 17 %, ce qui est du même ordre de grandeur que l'effet d'avoir un parent avec un diplôme universitaire, qui est lui de 20 %. Le revenu familial n'est significatif à un seuil de 1 % que pour l'inscription à l'université et l'effet est très faible, en effet un salaire parental plus élevé de 10 000 $ n'augmente que de 1 % la probabilité d'inscription à l'université, toutes choses étant égales par ailleurs. Notons que les jeunes provenant d'une famille allophone (dont la langue maternelle est autre que le français ou l'anglais) ont une plus grande probabilité de d'obtenir un diplôme d'études secondaires et d'accéder aux niveaux collégial et universitaire, les effets marginaux étant importants, de 6 % à 9 %, et étant tous significatifs. Nous observons aussi un effet significatif d'avoir le statut de minorité visible sur la probabilité d'accéder à l'université. Sans surprise, les jeunes femmes ont une plus grande probabilité de diplômer du secondaire et d'accéder à l'université.

Nous présentons dans le tableau 5.2 les mêmes résultats que le tableau précédent. Cependant, nous y retrouvons les coefficients à la place des effets marginaux et surtout les écarts-types calculés à l'aide de la méthode bootstrap. Ces calculs d'écarts-types ont été effectués à l'aide de 184 réplications de l'estimation, chacune pondérée avec l'un des poids fournis par Statistique Canada. Ce nombre de réplications effectuées est dû au fait qu'un certain nombre d'estimations ne convergeaient pas et ont donc été écartées. Le principal changement observé lorsque l'on calcule des écarts-types à l'aide de la méthode bootstrap, c'est une augmentation des écarts-types et c'est ce que nous constatons à la lecture du tableau 5.2.

Tableau 5.2 Résultats de la régression par NPMLE - Calcul des écarts-types bootstrap - Spécification 1

	Transitions								
	Obtention d'un diplôme d'études secondaires (N = 2084)			**Inscription au niveau collégial** (N = 1954)			**Inscription au niveau universitaire** (N = 1954)		
Variables	Coefficients NPMLE	Écarts-types	Écarts-types bootstrap	Coefficients NPMLE	Écarts-types	Écarts-types bootstrap	Coefficients NPMLE	Écarts-types	Écarts-types bootstrap
Minorité visible	0,261	0,980	1,838	0,45	0,474	0,555	0,80	0,334**	0,382**
Revenu familial	0,009	0,006	0,007	0,00	0,002	0,003	0,01	0,002***	0,002***
Femme	0,867	0,404**	0,485*	0,19	0,183	0,178	0,58	0,148***	0,163***
Niveau d'études des parents[1]									
Pas de diplôme - secondaire	-1,133	0,451**	0,784	-0,30	0,338	0,418	-0,34	0,288	0,325
Diplôme - secondaire (réf.)	-	-	-	-	-	-	-	-	-
Diplôme - collégiale	0,707	0,447	0,646	0,26	0,209	0,217	0,28	0,170	0,175
Diplôme - universitaire	1,813	0,607***	0,730**	0,47	0,217**	0,279*	1,24	0,215***	0,258***
Score au test PISA	2,468	0,558***	0,726***	0,85	0,107***	0,183***	1,04	0,124***	0,163***
Langue marternelle									
Autres langues	1,68	0,633***	0,865*	0,63	0,261**	0,446	0,57	0,192***	0,187***
Anglais	1,66	1,159	1,348	0,41	0,516	0,623	0,57	0,342*	0,325*
Français (réf.)	-	-	-	-	-	-	-	-	-
Rural	0,42	0,386	0,704	0,11	0,188	0,253	-0,23	0,156	0,209
Constante	15,13	10,358	6,988**	1,21	1,979	2,831	-1,77	0,266***	0,443***

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 5.

Écarts-types: * significatif à 10%; ** significatif à 5%; *** significatif à 1%.

Ce tableau présente les coefficients de la régression.

1. Même regroupement que dans le tableau 4.1.

Cette augmentation des écarts-types n'est pas trop importante : dans la majorité des cas, les écarts-types n'ont pas augmenté de plus de 50 %. Surtout, à quelques exceptions près, les coefficients restent significatifs à un seuil de 10 % lorsque les écarts-types sont calculés à l'aide de la méthode bootstrap.

Les tableaux 5.1 et 5.2 présentaient une spécification sans sélection entre le cégep et l'université. Dans le tableau 5.3, nous présentons les résultats de la deuxième spécification pour laquelle seuls les jeunes inscrits au cégep sont sélectionnés dans l'échantillon de l'inscription à l'université. Ceci a comme premier effet de faire passer l'échantillon universitaire de 1954 à 1453 jeunes. Nous n'observons virtuellement aucune différence entre les résultats de cette spécification et les résultats de la précédente spécification pour les estimations de la première transition, l'échantillon de celle-ci restant inchangé. Pour la deuxième transition, seul l'effet de la détention d'un diplôme universitaire par un parent augmente et devient significatif. Pour la dernière transition, malgré la variation de la taille de l'échantillon, seuls les effets de l'environnement linguistique changent, ceux-ci diminuant et devenant non-significatifs. Cette diminution s'observe tant pour l'estimateur logistique que pour le NPMLE.

Les résultats de la troisième spécification présentés dans le tableau 5.4 nous montrent l'influence qu'a la variable du score PISA sur l'estimation puisque la variable du score PISA a été retirée de l'estimation. Cela fait augmenter l'effet marginal de nombreuses variables. Ainsi, pour les trois transitions, l'effet d'être une femme et les effets du niveau de diplôme parental sont amplifiés. Pour le cas extrême, l'effet de l'absence chez les parents d'un diplôme d'études secondaires sur la probabilité d'obtenir un diplôme d'études secondaires passe de -4,4 % à -31 %. Ces variations confirment l'importance de la variable du score PISA. Celle-ci capte en partie l'effet du niveau d'éducation des parents sur les aptitudes scolaires à 15 ans, âge du test. Il faut cependant noter que la variation des effets diminue fortement lorsque l'on avance dans les transitions. Cette diminution reflète probablement l'éloignement dans le temps de la mesure (le score PISA), l'éloignement diminuant son effet prédictif de l'obtention d'un diplôme. Quant à l'effet de l'environnement linguistique, celui-ci est grandement diminué par le retrait de la

Tableau 5.3 Résultats de la régression par NPMLE - Spécification 2

	Transitions					
	Obtention d'un diplôme d'études secondaires (N = 2084)		Inscription au niveau collégial (N = 1954)		Inscription au niveau universitaire (N = 1453)	
Variables	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE
Minorité visible	0,048	0,011	0,061	0,057	0,113*	0,116*
	(0,051)	(0,037)	(0,058)	(0,059)	(0,063)	(0,064)
Revenu familial	0,00021	0,00034*	0,00028	0,00021	0,00095***	0,00094***
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)
Femme	0,035**	0,034***	0,040*	0,029	0,105***	0,104***
	(0,016)	(0,013)	(0,022)	(0,024)	(0,026)	(0,026)
Niveau d'études des parents[1]						
Pas de diplôme - secondaire	-0,062***	-0,044*	-0,070*	-0,048	-0,074	-0,072
	(0,018)	(0,024)	(0,037)	(0,042)	(0,054)	(0,056)
Diplôme - secondaire (réf.)	-	-	-	-	-	-
Diplôme - collégial	0,024	0,027	0,046*	0,038	0,048	0,048
	(0,019)	(0,017)	(0,027)	(0,030)	(0,033)	(0,033)
Diplôme - universitaire	0,081***	0,068***	0,088***	0,071**	0,215***	0,216***
	(0,029)	(0,024)	(0,028)	(0,034)	(0,033)	(0,033)
Score au test PISA	0,087***	0,094***	0,154***	0,123***	0,168***	0,170***
	(0,008)	(0,009)	(0,011)	(0,031)	(0,015)	(0,017)
Langues marternelles						
Autres langues	0,055*	0,062**	0,108***	0,090**	0,044	0,047
	(0,029)	(0,024)	(0,036)	(0,041)	(0,035)	(0,037)
Anglais	0,046	0,058*	0,094	0,071	0,089	0,091
	(0,034)	(0,033)	(0,058)	(0,068)	(0,064)	(0,064)
Français (réf.)	-	-	-	-	-	-
Rural	0,039**	0,018	0,016	0,015	-0,029	-0,028
	(0,018)	(0,017)	(0,025)	(0,025)	(0,031)	(0,032)

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 5.
Écarts-types entre parenthèses : * significatif à 10%; ** significatif à 5%; *** significatif à 1%.
Ce tableau présente les effets marginaux moyens.
1. Même regroupement que dans le tableau 4.1.

Tableau 5.4 Résultats de la régression par NPMLE - Spécification 3

	Transitions					
	Obtention d'un diplôme d'études secondaires (N = 2084)		Inscription au niveau collégial (N = 1954)		Inscription au niveau universitaire (N = 1453)	
Variables	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE
Minorité visible	0,031	0,021	-0,002	-0,006	0,055	0,074
	(0,050)	(0,029)	(0,062)	(0,055)	(0,068)	(0,076)
Revenu familial	0,00015	0,00055*	0,00056	0,00042	0,00123***	0,00144***
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,001)
Femme	0,076***	0,073***	0,086***	0,063***	0,144***	0,159***
	(0,017)	(0,016)	(0,023)	(0,021)	(0,027)	(0,027)
Niveau d'études des parents[1]						
Pas de diplôme - secondaire	-0,092***	-0,310*	-0,092**	-0,059	-0,090	-0,109*
	(0,021)	(0,184)	(0,040)	(0,042)	(0,059)	(0,056)
Diplôme - secondaire (réf.)	-	-	-	-	-	-
Diplôme - collégial	0,050**	0,041**	0,084***	0,066**	0,067*	0,070**
	(0,021)	(0,019)	(0,029)	(0,028)	(0,035)	(0,035)
Diplôme - universitaire	0,144***	0,100***	0,145***	0,109***	0,251***	0,287***
	(0,035)	(0,023)	(0,030)	(0,029)	(0,033)	(0,035)
Score au test PISA	-	-	-	-	-	-
Langues marternelles						
Autres langues	-0,024	-0,012	0,083**	0,080**	0,015	-0,001
	(0,032)	(0,073)	(0,035)	(0,034)	(0,037)	(0,042)
Anglais	-0,016	0,007	0,030	0,024	0,022	0,022
	(0,039)	(0,026)	(0,059)	(0,053)	(0,068)	(0,067)
Français (réf.)	-	-	-	-	-	-
Rural	0,019	0,029*	-0,016	-0,022	-0,044	-0,027
	(0,018)	(0,016)	(0,027)	(0,025)	(0,033)	(0,033)

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 5.
Écarts-types entre parenthèses : * significatif à 10%; ** significatif à 5%; *** significatif à 1%.
Ce tableau présente les effets marginaux moyens.
1. Même regroupement que dans le tableau 4.1.

variable du score PISA. Ceci pourrait s'expliquer par un score PISA plus faible pour les jeunes de parents allophones et anglophones, mais qu'à score égal, ceux-ci sont plus portés à diplômer du secondaire et à poursuivre des études supérieures.

Le tableau 5.5 montre les résultats de la quatrième spécification où nous considérons qu'un jeune n'ayant pas obtenu de diplôme d'études secondaires à 21 ans ne l'obtiendra pas dans le futur. Cette hypothèse a pour effet de diminuer le nombre de personnes

Tableau 5.5 Résultats de la régression par NPMLE - Spécification 4

	Transitions					
	Obtention d'un diplôme d'études secondaires (N = 2084)		**Inscription au niveau collégial** (N = 1835)		**Inscription au niveau universitaire** (N = 1437)	
	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE
Minorité visible	0,025	0,008	0,039	0,040	0,142**	0,101
	(0,048)	(0,068)	(0,059)	(0,059)	(0,063)	(0,063)
Revenu familial	0,00064***	0,00059**	0,00019	0,00017	0,00090***	0,00092***
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)
Femme	0,042**	0,045***	0,030	0,027	0,106***	0,109***
	(0,017)	(0,016)	(0,022)	(0,022)	(0,026)	(0,026)
Niveau d'études des parents[1]						
Pas de diplôme - secondaire	-0,072***	-0,066***	-0,071*	-0,063	-0,069	-0,092
	(0,022)	(0,022)	(0,038)	(0,040)	(0,055)	(0,060)
Diplôme - secondaire (réf.)	-	-	-	-	-	-
Diplôme - collégial	0,063***	0,068***	0,023	0,018	0,045	0,055*
	(0,021)	(0,019)	(0,027)	(0,029)	(0,033)	(0,033)
Diplôme - universitaire	0,081***	0,069***	0,059**	0,056*	0,223***	0,214***
	(0,025)	(0,026)	(0,028)	(0,028)	(0,033)	(0,032)
Score au test PISA	0,160***	0,164***	0,114***	0,102***	0,161***	0,177***
	(0,009)	(0,009)	(0,013)	(0,022)	(0,016)	(0,018)
Langues marternelles						
Autres langues	0,119***	0,106***	0,075**	0,070**	0,044	0,048
	(0,035)	(0,037)	(0,034)	(0,034)	(0,035)	(0,035)
Anglais	0,101**	0,116**	0,100*	0,091	0,061	0,094
	(0,043)	(0,057)	(0,058)	(0,060)	(0,064)	(0,063)
Français (réf.)	-	-	-	-	-	-
Rural	0,051**	0,036*	-0,001	-0,003	-0,031	-0,025
	(0,020)	(0,019)	(0,025)	(0,024)	(0,032)	(0,031)

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 5.
Écarts-types entre parenthèses : * significatif à 10%; ** significatif à 5%; *** significatif à 1%.
Ce tableau présente les effets marginaux moyens.
1. Même regroupement que dans le tableau 4.1.

éligibles au cégep de 1954 à 1835 dans l'échantillon collégial et de diminuer l'échantillon universitaire de 1453 à 1437 personnes. Ce changement augmente sensiblement l'amplitude des résultats en comparaison aux résultats de la troisième, de la deuxième et de la première spécification. En effet, les effets marginaux de l'absence de diplôme d'études secondaires chez les parents ont un effet plus fortement négatif et sont maintenant significatifs à un seuil de 1 %. De plus, l'effet marginal moyen qu'un parent ait un diplôme d'études collégiales augmente fortement, passant de 2,7 % à 6,8 % et devenant significatif à un seuil de 1 %. L'effet du diplôme collégial d'un parent devient égal à l'effet du diplôme universitaire, ce dernier restant inchangé. De la même manière, l'effet sur la probabilité de diplômer du secondaire de venir d'une famille allophone ou anglophone est plus important et dans le cas de parents anglophones, l'effet devient significatif à un seuil de 5 %. Les effets pour les deuxième et troisième transitions varient beaucoup moins et il n'y a aucune différence notable.

Dans le tableau 5.6, nous retrouvons l'estimation de la cinquième spécification. Dans celle-ci, nous changeons les deux dernières transitions en considérant l'obtention d'un diplôme comme alternative à l'inscription. Comme cela est mentionné plus haut, nous utilisons le cycle 6 de l'EJET pour obtenir un portrait plus juste de l'obtention d'un diplôme universitaire, ceci a comme effet de diminuer l'échantillon. Pour les trois transitions, nous avons, dans l'ordre, 1656, 1585 et 883 jeunes. La diminution plus importante de la troisième transition est due à l'effet combiné de l'utilisation du cycle 6 et du fait que le critère d'inclusion pour cette transition est maintenant l'obtention d'un diplôme d'études collégiales et non la simple inscription. Nous comparons les résultats de la spécification 4, avec restriction sur les diplômes d'études secondaires, puisque nous appliquons aussi cette restriction à la présente spécification. Pour la première transition, nous observons une baisse notable de la vaste majorité des effets marginaux. Entre autres, l'effet marginal du revenu familial, d'avoir un parent avec un diplôme d'études collégiales et de vivre dans un ménage anglophone devient non-significatif. D'un autre côté, il y a une importante augmentation de l'amplitude des effets marginaux de la seconde transition. Alors que pour les autres spécifications, la deuxième transition était celle avec les

Tableau 5.6 Résultats de la régression par NPMLE - Spécification 5

	Transitions					
	Obtention d'un diplôme d'études secondaires (N = 1656)		**Obtention d'un diplôme de niveau collégial** (N = 1585)		**Obtention d'un diplôme universitaire** (N = 883)	
Variables	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE
Minorité visible	0,002	-0,016	0,139**	0,139**	0,014	0,016
	(0,046)	(0,035)	(0,068)	(0,069)	(0,076)	(0,072)
Revenu familial	-0,00002	0,00008	0,00011	0,00008	0,00054	0,00040
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)
Femme	0,022	0,030*	0,118***	0,115***	0,100***	0,093**
	(0,017)	(0,016)	(0,024)	(0,025)	(0,034)	(0,043)
Niveau d'études des parents[1]						
Pas de diplôme - secondaire	-0,054***	-0,054***	-0,126**	-0,114**	-0,102	-0,093
	(0,021)	(0,020)	(0,050)	(0,053)	(0,080)	(0,066)
Diplôme - secondaire (réf.)	-	-	-	-	-	-
Diplôme - collégial	0,021	0,022	0,002	-0,001	0,011	0,003
	(0,020)	(0,018)	(0,032)	(0,033)	(0,044)	(0,039)
Diplôme universitaire	0,108***	0,091***	0,123***	0,119***	0,151***	0,163***
	(0,035)	(0,032)	(0,033)	(0,033)	(0,043)	(0,046)
Score au test PISA	0,063***	0,064***	0,161***	0,154***	0,161***	0,161***
	(0,009)	(0,009)	(0,013)	(0,016)	(0,020)	(0,018)
Langues marternelles						
Autres langues	0,045	0,053**	0,004	-0,001	0,004	-0,004
	(0,032)	(0,027)	(0,039)	(0,039)	(0,050)	(0,049)
Anglais	0,010	0,016	0,034	0,029	0,079	0,054
	(0,040)	(0,033)	(0,072)	(0,074)	(0,080)	(0,087)
Français (réf.)	-	-	-	-	-	-
Rural	0,052**	0,037	0,035	0,031	-0,060	-0,052
	(0,020)	(0,030)	(0,029)	(0,029)	(0,041)	(0,045)

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 6.
Écarts-types entre parenthèses : * significatif à 10%; ** significatif à 5%; *** significatif à 1%.
Ce tableau présente les effets marginaux moyens.
1. Même regroupement que dans le tableau 4.1.

effets marginaux de moindre amplitude et majoritairement non-significatifs, dans cette spécification, certains effets ressortent très fort. Entre autres, être membre d'une minorité visible augmente la probabilité de diplômer de 13,9 % et celui d'être une femme de 11,5 %. Il y a aussi un effet fortement négatif de provenir d'un ménage où aucun parent n'a de diplôme d'études secondaires (-11,4 %) et inversement fortement positif de provenir d'un ménage avec un parent (au moins) avec un diplôme d'études universitaires (15,4 %) sur la probabilité d'obtenir un diplôme d'études collégiales. Pour la dernière transition, trois déterminants sont significatifs : le fait d'être une femme, le fait d'avoir un parent ayant un diplôme universitaire et le score PISA ont un effet marginal significatif, ceux-ci étant respectivement de 9,3 %, de 16,3 % et de 16,1 %. Les différences de résultats entre l'estimateur logit et NPMLE restent minimes dans cette spécification.

Finalement, dans le tableau 5.7, nous retrouvons les résultats de la sixième spécification, la plus complète, regroupant l'inscription et l'obtention d'un diplôme du collège et de l'université pour un total de cinq transitions. Comme précédemment, nous effectuons la sélection pour le niveau collégial et nous limitons l'âge d'obtention d'un diplôme d'études secondaires à 21 ans. Un des intérêts de cette spécification est, que l'estimation de l'effet des déterminants sur l'obtention d'un diplôme se fait seulement sur les inscrits, il faut donc interpréter avec précaution les estimations pour les transitions de l'obtention d'un diplôme. Dans cette spécification, nous n'observons toujours pas le déclin des effets marginaux tel qu'observé dans la littérature (Cameron et Heckman, 1998; Mare, 1979) et les effets marginaux obtenus avec l'estimateur logit et le NPMLE restent similaires. La seule variable ayant un effet significatif pour l'ensemble des transitions est le score PISA, effet variant entre 6,4 % pour l'obtention d'un diplôme d'études secondaires et 14,4 % pour l'inscription à l'université. Dans deux transitions, l'inscription au collégial et l'obtention d'un diplôme universitaire, seul l'effet marginal du score PISA est significatif à un seuil de 5 %. Pour les trois autres transitions, seule la détention d'un diplôme d'études universitaires par un des parents a un effet marqué, celui-ci atteignant 14,7 % d'augmentation pour la probabilité d'inscription à l'université. Globalement, les effets sont plus faibles dans cette spécification.

Tableau 5.7 Résultats de la régression par NPMLE - Spécification 6

	Transitions									
	Obtention d'un diplôme d'études secondaires (N = 1656)		Inscription au niveau collégial (N = 1585)		Obtention d'un diplôme de niveau collégial (N = 1191)		Inscription au niveau universitaire (N = 883)		Obtention d'un diplôme universitaire (N = 663)	
Variables	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE	Logit	NPMLE
Minorité visible	0,002	0,003	0,122*	0,113	0,098	0,066	0,098	0,098	-0,046	-0,058
	(0,046)	(0,063)	(0,072)	(0,084)	(0,075)	(0,099)	(0,081)	(0,083)	(0,078)	(0,119)
Revenu familial	-0,00002	0,00024	0,00031	0,00021	0,00001	0,00002	0,00086*	0,00084*	0,00025	0,00020
	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)	(0,000)
Femme	0,022	0,028*	0,055**	0,043	0,121***	0,114***	0,069**	0,056	0,094**	0,107
	(0,017)	(0,016)	(0,024)	(0,040)	(0,029)	(0,034)	(0,033)	(0,045)	(0,041)	(0,134)
Niveau d'études des parents[1]										
Pas de diplôme - secondaire	-0,054***	-0,047	-0,086**	-0,061	-0,083	-0,093	-0,109	-0,100	-0,044	-0,045
	(0,021)	(0,033)	(0,041)	(0,051)	(0,060)	(0,062)	(0,068)	(0,075)	(0,113)	(0,113)
Diplôme - secondaire (réf.)	-	-	-	-	-	-	-	-	-	-
Diplôme - collégial	0,021	0,023	0,041	0,030	-0,033	-0,029	-0,043	-0,040	0,058	0,055
	(0,020)	(0,017)	(0,030)	(0,046)	(0,038)	(0,038)	(0,040)	(0,043)	(0,059)	(0,068)
Diplôme - universitaire	0,108***	0,081**	0,102***	0,078	0,084**	0,081**	0,156***	0,147***	0,095*	0,092
	(0,035)	(0,032)	(0,031)	(0,063)	(0,039)	(0,040)	(0,043)	(0,053)	(0,053)	(0,081)
Score au test PISA	0,063***	0,064***	0,146***	0,123**	0,098***	0,084*	0,148***	0,144***	0,110***	0,097**
	(0,009)	(0,010)	(0,013)	(0,051)	(0,019)	(0,045)	(0,020)	(0,029)	(0,026)	(0,041)
Langues marternelles										
Autres langues	0,045	0,048**	0,071*	0,055	-0,024	-0,025	0,013	0,019	-0,015	-0,028
	(0,032)	(0,022)	(0,040)	(0,056)	(0,042)	(0,043)	(0,048)	(0,051)	(0,059)	(0,112)
Anglais	0,010	0,023	0,026	0,016	0,035	0,037	0,092	0,085	0,036	0,039
	(0,040)	(0,067)	(0,069)	(0,090)	(0,079)	(0,093)	(0,086)	(0,087)	(0,090)	(0,146)
Français (réf.)	-	-	-	-	-	-	-	-	-	-
Rural	0,052**	0,032	0,005	-0,003	0,043	0,049	-0,091**	-0,098*	0,011	0,032
	(0,020)	(0,034)	(0,027)	(0,028)	(0,035)	(0,036)	(0,038)	(0,051)	(0,053)	(0,078)

Source : Calculs de l'auteur à l'aide de données de Statistique Canada - EJET cohorte A , cycles 1 à 6.

Écarts-types entre parenthèses : * significatif à 10%; ** significatif à 5%; *** significatif à 1%.

Ce tableau présente les effets marginaux moyens.

1. Même regroupement que dans le tableau 4.1.

5.3 Analyse des résultats, constats et critiques

Dans cette section, nous analysons les résultats exposés à la section précédente. Dans un premier temps, nous tirons des constats quant aux déterminants socio-économiques du parcours scolaire et, dans un second temps, nous élaborons sur les avantages et les défis liés à l'utilisation du NPMLE comme alternative au modèle de transition standard utilisant une estimation par estimateur logit.

Des résultats précédents nous portons une attention particulière à trois spécifications. Nous retenons la quatrième spécification, celle qui restreint l'obtention du diplôme d'études secondaires aux jeunes de 21 ans et moins. Nous effectuons ce choix, car les deux hypothèses ajoutées au modèle de base permettent de mieux cerner le parcours scolaire des jeunes et que, contrairement aux deux dernières spécifications, nous ne sommes pas obligés d'utiliser les données du cycle 6 et la diminution d'échantillon attachée à cette utilisation. La cinquième spécification est aussi retenue puisqu'elle utilise les mêmes hypothèses que la précédente et qu'elle permet d'explorer l'effet des déterminants socio-économiques sur l'obtention d'un diplôme d'études collégiales et universitaires. Finalement, nous nous attardons brièvement à la sixième spécification pour sa nature englobante des deux précédentes spécifications.

Que nous apprennent les résultats précédents sur les déterminants socio-économiques du parcours scolaire ? D'abord, en se basant sur les spécifications 4 et 5, nous observons qu'un des principaux déterminants est le niveau d'éducation des parents : les jeunes dont aucun des parents n'a de diplôme d'études secondaires étant fortement pénalisés et ceux dont au moins un des parents a un diplôme d'études universitaires sont fortement avantagés. Inversement, le revenu familial n'est pas un déterminant important dans l'ensemble des transitions, l'effet le plus important étant une augmentation de la probabilité de l'inscription à l'université de 0,9 % par tranche de 10 000 $ de revenu familial dans la spécification 3. Comme pour tous les résultats, ceux-ci sont vrais toutes choses étant égales par ailleurs, il s'agit donc d'une variation du revenu familial à plus hauts diplômes parentaux égaux. Ces deux effets sont parfaitement cohérents avec les résultats recensés

dans la revue de littérature (Mare, 1980; Finnie, Laporte et Lascelles, 2004; Cameron et Heckman, 1998). De plus, les aptitudes, mesurées par le score du test PISA, ont quant à elles un effet très important dans l'ensemble des spécifications et pour l'ensemble des transitions. L'effet varie entre 6,4 % et 16,6 % par écart-type de distance à la moyenne du test dépendant de la transition et de la spécification, un effet est aussi retrouvé dans les études utilisant des mesures de l'habileté (Kamanzi *et al.*, 2009; Cameron et Heckman, 1998). En plus, nous observons une forte variation de certains effets marginaux lorsque l'on retire la variable du score PISA dans la spécification 3, confirmant l'importance d'inclure une variable de contrôle pour l'aptitude. Par la suite, nous observons que l'effet de provenir d'une famille appartenant à une minorité visible n'est pas clair. D'un côté, alors que dans la spécification 4, le fait d'appartenir à une minorité visible n'a aucun effet significatif à un seuil de 10 %, dans la spécification 5, cette appartenance a seulement un effet, très fort, pour l'obtention d'un diplôme d'études collégiales. Ces résultats sont surprenants, car Finnie *et al.* (2008), qui utilisent l'EJET dans leur étude, trouvent que les enfants d'immigrants participent plus aux études postsecondaires et ceci devrait être capté ici par la variable minorité visible. Cependant, le très faible nombre de jeunes provenant de minorités visibles (120) et les écarts-types élevés obtenus par le NPMLE peuvent expliquer cette différence. De l'autre côté, provenir d'une famille allophone augmente de façon importante la probabilité d'obtenir un diplôme d'études secondaires et de s'inscrire au niveau collégial, ce qui correspond aux résultats obtenus par Kamanzi *et al.* (2009). Puis, le genre a aussi un effet important, les jeunes femmes ayant une plus forte probabilité de diplômer des trois niveaux et de s'inscrire à l'université. Finalement, le facteur géographique n'a que très peu d'effet notamment sur l'inscription à l'université ce qui semble contraire à certaines études (Frenette, 2002) qui trouvent une probabilité plus faible d'inscription pour les jeunes provenant de régions rurales. Notre mesure de distance à l'université se limite à une variable dichotomique, soit la région rurale et la région non rurale, ce qui peut expliquer, en partie, la différence. Cependant, notre modèle est plus complet que celui de Frenette (2002) et une fois la sélection prise en compte il est possible que l'effet de la distance perde de l'importance.

Un des points forts de ces résultats est qu'ils sont cohérents d'une spécification à l'autre. Il y a bien sûr de la variation entre les effets marginaux des différentes spécifications, cependant, dans aucun des cas des effets marginaux ne changent de signe et globalement les effets marginaux restent dans le même ordre de grandeur d'une spécification à l'autre. Il apparait que l'ajout d'hypothèses ne fait pas varier de façon trop importante les résultats et, de la même manière, la diminution de l'échantillon, résultat de l'utilisation des données du cycle 6, ne cause pas de grande variation dans les effets marginaux. Le second point fort de ces résultats est, comme nous l'avons vu précédemment, leur cohérence avec les résultats obtenus dans la littérature. De plus, les effets des déterminants socio-économiques des parcours scolaires au Québec n'avaient jamais, à notre connaissance, été étudiés à l'aide d'un modèle de transition prenant en compte la sélection sur les inobservables entre les transitions. Ce que nous constatons, c'est que malgré l'utilisation de ce modèle plus complexe les résultats ressemblent à ceux observés dans la littérature. L'élément de la littérature que nous ne retrouvons pas est celui de la décroissance des effets d'une transition à l'autre (Mare, 1980; Cameron et Heckman, 1998). Cette décroissance ne s'observe pas dans les estimations par modèle logit ni lorsque nous estimons les effets à l'aide du NPMLE.

Le second point intéressant est d'évaluer l'intérêt de l'utilisation du NPMLE. Rappelons que nous utilisons un NPMLE à la place de simples estimations à l'aide d'estimateur logistique pour corriger le biais introduit par l'hétérogénéité inobservée. Ce biais est double. D'un côté, le caractère non linéaire du modèle logistique crée un premier biais et, de l'autre, la sélection en fonction des facteurs inobservables introduit un second biais (Cameron et Heckman, 1998). Nous avons montré à l'aide de simulations dans les sections 2.4.2 et 3.5 que ce biais peut être important sous certaines conditions et que le NPMLE implanté à l'aide de la commande GLLAMM peut corriger en partie ce biais. Afin d'apprécier cette correction, nous avons estimé le modèle de transition standard à l'aide d'un estimateur logistique et nous avons utilisé le NPMLE. Les résultats parlent d'eux-mêmes : les estimations sont très semblables, à quelques exceptions près, les effets marginaux sont quasi identiques. Il semble donc que malgré le biais théorique, celui-ci

ne se manifeste pas dans les données. L'absence d'observation du biais peut être due à la faible amplitude des facteurs inobservés. En effet, comme nous l'avons vu dans les simulations de la section 2.4.2, lorsque l'amplitude de l'hétérogénéité inobservée est faible le biais est faible, ce qui expliquerait le peu de différence entre nos différents résultats. De plus, il faut prendre note des limitations de notre estimation qui affaiblissent le constat de l'absence de biais. En effet, l'estimation à l'aide de la commande GLLAMM ne se fait pas sans difficulté. Premièrement, la fonction à estimer est très complexe, et il existe un nombre important de maximums locaux (Cameron et Taber, 1998). Lors des estimations, nous avons eu à plusieurs reprises des difficultés à obtenir un résultat convergent. Cette non-convergence, ou parfois une convergence à un maximum local, rend les estimations très variables. Ceci est confirmé par la simulation du chapitre 3 dans laquelle les résultats d'estimation à l'aide de la commande GLLAMM ont une variance très élevée. Nous ne pouvons dire si c'est la commande GLLAMM et le type d'algorithme utilisé par celle-ci qui pose problème ou cette difficulté de convergence est intrinsèque au NPMLE. Cette impossibilité d'identifier le problème provient de la seconde difficulté liée à l'utilisation du NPMLE, soit l'absence de commande incluse dans les principaux logiciels statistiques utilisés (Stata, SAS, R). Cette absence oblige soit à utiliser une commande comme GLLAMM, ou encore programmer soi-même les algorithmes de maximisation de la fonction de maximum de vraisemblance, ce qui représente tant un défi technique qu'un temps de travail important. De plus, l'estimation à l'aide de la commande GLLAMM du NPMLE nécessite des ressources de calcul importantes et implique des délais importants pour l'estimation. Ces délais deviennent prohibitifs lorsqu'il s'agit d'implanter un calcul des écarts-types à l'aide de la méthode bootstrap. Finalement, les difficultés de convergence et les temps de calcul importants nous obligent à restreindre de façon importante le nombre de variables dépendantes, les neuf utilisées dans le présent mémoire étant proches de la limite.

En conclusion, les résultats très semblables des estimations par estimateur logistique et par NPMLE suggèrent qu'il n'y a pas de biais important lorsque l'on estime par logit. Cependant, Cameron et Heckman (1998) obtiennent des résultats différents à l'aide du

NPMLE signe de la présence d'un biais dans l'estimation par logit. Aux résultats de Cameron et Heckman (1998) s'ajoutent les nombreuses limitations de notre estimation par NPMLE, ce qui ne nous permet pas de conclure qu'il n'y a pas de biais lors de l'estimation du modèle de transition par estimateur logistique. Nous pouvons cependant penser que malgré les limites de nos estimations s'il existait un biais important dans l'estimation des effets des déterminants socio-économiques sur le parcours scolaire des jeunes québécois par estimateur logistique, nous aurions observé des différences plus importantes entre les résultats. Nous en venons donc à la conclusion, qu'à ce jour, l'utilisation d'un tel estimateur ne représente pas un investissement en temps intéressant pour la majorité des chercheurs. Cela n'enlève en rien l'intérêt du présent exercice qui permet justement d'évaluer la sensibilité des estimations au biais d'hétérogénéité inobservée, tout en apportant un éclairage sur les déterminants socio-économiques du parcours scolaire des jeunes québécois dans un contexte de modèle de transition. En effet, bien que plusieurs travaux aient porté sur les déterminants de l'accès aux études postsecondaires au Québec et au Canada (Kamanzi *et al.*, 2009; Frenette, 2002; Finnie, Laporte et Lascelles, 2004) aucun n'a étudié, à notre connaissance, conjointement toutes les différentes transitions, soit de l'obtention d'un diplôme d'études secondaires à l'obtention d'un diplôme universitaire.

CONCLUSION

Dans ce mémoire, nous avons estimé l'effet de plusieurs déterminants socio-économiques sur le parcours scolaire des jeunes québécois. Nous avons pour ce faire utilisé un estimateur non-paramétrique de maximum de vraisemblance. Nous avons choisi d'utiliser un tel estimateur, car le modèle fréquemment utilisé, soit le modèle de transition utilisant une séquence d'estimation logistique d'une transition à l'autre, est biaisé en présence d'hétérogénéité inobservée par le chercheur. Cette critique, déjà mentionnée par Mare (1980) dans son article phare utilisant un modèle de transition standard, est reprise formellement par Cameron et Heckman (1998). Nous avons utilisé les données de l'Enquête auprès des jeunes en transition afin d'estimer le modèle de transitions à l'aide d'un estimateur logistique et à l'aide du NPMLE. Cette double estimation nous a permis de conclure que l'utilisation du NPMLE ne représente pas un grand intérêt pour les chercheurs actuellement, et ceci, pour deux principales raisons. Premièrement, en raison des difficultés liées à son utilisation et, deuxièmement, car il y a très peu de différences entre les résultats obtenus à l'aide de l'estimateur logistique et ceux résultant du NPMLE.

Dans un premier temps, nous avons brossé un portrait de la littérature où nous avons vu que l'utilisation du modèle de transition proposé pour la première fois par Mare (1980) comporte un biais d'hétérogénéité, déjà connu à l'époque et mis en évidence par Cameron et Heckman (1998). Ces auteurs suggèrent l'utilisation du NPMLE afin d'estimer sans biais les déterminants socio-économiques. Cette méthodologie a été reprise par plusieurs chercheurs, dont Karlson (2011) et Buis (2011). Dans un deuxième temps, nous avons présenté le modèle de transition qui consiste à étudier les déterminants du parcours scolaire en divisant ce dernier en transitions représentant les différentes étapes du cheminement scolaire. Dans ce modèle, il s'effectue une sélection entre les transitions, seules les personnes ayant complété la transition précédente sont incluses

dans l'échantillon de la transition. Ce modèle habituellement estimé à l'aide d'un modèle logit est biaisé, car il ne prend pas en compte l'hétérogénéité inobservée. Celle-ci induit deux biais, le premier résulte de la sélection qui s'effectue lors des transitions en fonction des facteurs inobservables et le second provient de la corrélation qui apparait entre les observables et les inobservables lorsque l'on progresse dans les transitions. À l'aide d'une simulation, nous montrons que le biais décrit théoriquement par Cameron et Heckman peut introduire une erreur d'estimation importante lorsque l'amplitude de l'hétérogénéité inobservée est égale ou supérieure à celle des variables observées.

Dans un troisième temps, nous avons introduit le NPMLE qui est un estimateur de maximum de vraisemblance dont le terme d'erreur est distribué logistiquement et qui comprend un effet aléatoire avec une distribution qui est inconnue (Cameron et Taber, 1998). Afin d'estimer l'effet des déterminants socio-économiques à l'aide du NPMLE, nous avons utilisé la commande Stata GLLAMM. Nous avons calculé les écarts-types à l'aide de la méthode bootstrap, nous avons cependant dû nous limiter à appliquer cette méthode à une seule estimation, car elle était trop intensive en calcul. Nous avons effectué une autre simulation qui nous a montré que l'utilisation du NPMLE permettait, dans la pratique, de corriger le biais d'hétérogénéité inobservée dans les conditions précises de la simulation. Dans un quatrième temps, nous avons présenté l'EJET. De cette enquête longitudinale, nous avons retenu la cohorte A qui comprend des jeunes qui avaient 15 ans au début de l'enquête et qui ont 25 ans au 6^e cycle de l'enquête. Cette enquête comprend une mesure des compétences cognitives, soit le résultat au test de lecture PISA, ainsi que des informations détaillées sur les conditions socio-économiques familiales grâce à un questionnaire administré aux parents lors du premier cycle de l'enquête. Nous avons donc, au final, retenu les jeunes de la cohorte A qui habitaient le Québec au premier cycle.

Cinquièmement, nos résultats nous montrent que globalement les déterminants les plus importants sont le niveau d'études des parents, les habitudes cognitives évaluées dans notre échantillon par le score PISA ainsi que le genre de la personne. Nous retrouvons aussi un effet moindre des déterminants socio-économiques sur l'accès aux études collé-

giales que sur l'obtention d'un diplôme secondaire ainsi que sur l'accès à l'université. Le résultat le plus marquant de nos estimations est l'absence de différence entre les résultats obtenus avec l'estimateur logistique et ceux obtenus à l'aide du NPMLE. Ces résultats suggèrent que l'hétérogénéité inobservée ne joue pas un grand rôle dans le mécanisme de sélection entre les transitions. Cependant, les nombreuses difficultés dans l'estimation à l'aide du NPMLE rendent toutes conclusions finales difficiles quant à la présence d'un biais lié à l'hétérogénéité inobservée.

Finalement, à notre avis, il n'est pas intéressant à ce jour pour un chercheur d'utiliser le NPMLE en raison des difficultés et du temps de calcul nécessaire à son implantation, le tout pour des résultats semblables à la simple utilisation d'un estimateur logistique. En effet, ce mémoire montre que l'estimation à l'aide du modèle de transition standard donne des résultats très similaires à ceux obtenus par NPMLE. De plus, nous avons montré que peu importe la méthode utilisée les résultats obtenus sont globalement semblables à ceux proposés dans la littérature sur les déterminants socio-économiques du parcours scolaire.

BIBLIOGRAPHIE

Bonikowska, A., D. A. Green, et W. C. Riddell. 2008. « Littératie et marché du travail : Les capacités cognitives et les gains des immigrants ». Statistique Canada, catalogue No 89-552-MIF2008020 au catalogue de Statistique Canada. Division de la Culture, tourisme et centre de la statistique de l'éducation. Ottawa : Statistique Canada.

Breen, R. et J. O. Jonsson. 2000. « Analyzing educational careers : A multinomial transition model », *American Sociological Review*, vol. 65, no. 5, pp. 754–772.

Buis, M. L. 2011. « The consequences of unobserved heterogeneity in a sequential logit model », *Research in Social Stratification and Mobility*, vol. 29, no. 3, pp. 247–262.

Cameron, S. V. et J. J. Heckman. 1998. « Life Cycle Schooling and Dynamic Selection Bias : Models and Evidence for Five Cohorts of American Males », *Journal of Political Economy*, vol. 106, no. 2, pp. 262–333.

———. 2001. « The Dynamics of Educational Attainment for Black, Hispanic, and White Males », *Journal of Political Economy*, vol. 109, no. 3, pp. 455–499.

Cameron, S. V. et C. R. Taber. 1998. « Evaluation and identification of semiparametric maximum likelihood models of dynamic discrete choice ». Document non publié, Département d'écononomie, Northwestern University.

Christofides, L. N., M. Hoy, et *al.* 2001. « Family income and postsecondary education in Canada », *Canadian Journal of Higher Education*, vol. 31, no. 1, pp. 177–208.

CIRANO. sd. L'éducation au Québec : L'état de la situation. en ligne : < http ://www.cirano.qc.ca/icirano/public/pdf/webevents201009_etat _de_la_situation.pdf > (Consulté le 26 août 2013), 8 p.

Corak, M. R., G. Lipps, et J. Z. Zhao. 2003. « Family income and participation in post-secondary education ». No 11F00192003210 au catalogue de Statistique Canada. Direction des études analytiques, document de recherche no 210. Ottawa : Statistique Canada.

Drolet, M. 2005. « Participation in Post-secondary Education in Canada : Has the Role Changed over the 1990s ? ». No 11F0019MIE au catalogue de Statistique Canada. Division de l'analyse des entreprises et du marché du travail, document de recherche no 243. Ottawa : Statistique Canada.

Duncan, B. 1965. « Family factors and school dropout : 1920-1960 ». Cooperative Research Project No. 2258. Ann Arbor : The University of Michigan.

———. 1967. « Education and Social Background », *American Journal of Sociology*, vol. 72, no. 4, pp. pp. 363–372.

Finnie, R. et M. Frenette. 2003. « Earning differences by major field of study : evidence from three cohorts of recent Canadian graduates », *Economics of Education Review*, vol. 22, no. 2, pp. 179–192.

Finnie, R., C. Laporte, et E. Lascelles. 2004. « Antécédents familiaux et accès aux études postsecondaires : que s'est-il passé pendant les années 1990 ? ». No 11F0019MIE au catalogue de Statistique Canada. Division de l'analyse des entreprises et du marché du travail, document de recherche no 237. Ottawa : Statistique Canada.

Finnie, R., E. Lascelles, et A. Sweetman. 2005. « Who goes ? The direct and indirect effects of family background on access to post-secondary education ». No 11F0019MIF au catalogue de Statistique Canada. Division de l'analyse des entreprises et du marché du travail, document de recherche no 226. Ottawa : Statistique Canada.

Finnie, R. et R. Mueller. 2008. « Access to post-secondary education in Canada among first and second generation Canadian immigrants : Raw differences and some of the underlying factors », *Document de travail. Montréal, Canada, Fondation canadienne des bourses d'études du millénaire.*

Follman, D. A. 1985. « Nonparametric mixtures of logistic regression models ». Thèse de Doctorat, Carnegie Mellon University.

Frenette, M. 2002. « Trop loin pour continuer ? Distance par rapport à l'établissement ». No 11F0019MIF au catalogue de Statistique Canada. Division de l'analyse des entreprises et du marché du travail, document de recherché no 191. Ottawa : Statistique Canada.

———. 2003. « Accès au collège et à l'université : est-ce que la distance importe ? », *Direction des études analytiques.* No 11F0019MIF au catalogue de Statistique Canada. Division de l'analyse des entreprises et du marché du travail, document de recherche no 201. Ottawa : Statistique Canada.

Heckman, J. et B. Singer. 1984. « A method for minimizing the impact of distributional assumptions in econometric models for duration data », *Econometrica : Journal of the Econometric Society*, vol. 52, no. 2, pp. 271–320.

Kamanzi, P. C., P. Doray, J. Murdoch, S. Moulin, E. Comoé, A. Groleau, C. Leroy, et F. Dufresne. 2009. « L'influence des déterminants sociaux et culturels sur les parcours et les transitions dans les études postsecondaires. ». (Projet Transitions, Note de recherche 6). Montréal, Fondation canadienne des Bourses du millénaire. Numéro 47.

Karlson, K. B. 2011. « Multiple paths in educational transitions : A multinomial transition model with unobserved heterogeneity », *Research in Social Stratification and Mobility*, vol. 29, no. 3, pp. 323 – 341.

Kolenikov, S. s.d. « Primer on gllamm ». en ligne : < http ://www.unc.edu/ školenik/stata/gllamm-demo.html > (Consulté le 23 août 2013).

Lefebvre, P. et P. Merrigan. 2010. « The Impact of Family Background, Cognitive and Non-Cognitive Ability in Childhood On Post-Secondary Education Attendance : Evidence from the NLSCY », *Pursuing Higher Education in Canada (Sous la direction de R.Finnie, R.Mueller, M.Frenette et A. Sweetman)*, pp. 219–242.

Lucas, S. R. 2001. « Effectively Maintained Inequality : Education Transitions, Track Mobility, and Social Background Effects », *American Journal of Sociology*, vol. 106, no. 6, pp. 1642–1690.

Mare, R. D. 1979. « Social Background Composition and Educational Growth », *Demography*, vol. 16, no. 1, pp. 55–71.

———. 1980. « Social Background and School Continuation Decisions », *Journal of the American Statistical Association*, vol. 75, no. 370, pp. 295–305.

———. 1981. « Change and Stability in Educational Stratification », *American Sociological Review*, vol. 46, no. 1, pp. 72–87.

———. 2011. « Introduction to symposium on unmeasured heterogeneity in school transition models », *Research in Social Stratification and Mobility*, vol. 29, no. 3, pp. 239–245.

Mare, R. D. et H.-C. Chang. 2006. « Family Attainment Norms and Educational Stratification in the United States and Taiwan : The Effects of Parents' School Transitions », *Mobility and inequality. Frontiers of research in sociology and economics*, pp. 195–231.

McIntosh, J. 2010. « Educational mobility in Canada : results from the 2001 general social survey », *Empirical Economics*, vol. 38, no. 2, pp. 457–470.

OCDE. 2004. *Apprendre aujourd'hui, réussir demain – Premiers résultats de PISA 2003.* 506 p.

———. 2005. *PISA 2003 Data Analysis Manual : SPSS® Users.* 10 p.

———. s.d. *Programme for International Student Assessment.* en ligne : <http ://www.oecd.org/pisa/pisaenfranais.htm > (Consulté le 23 août 2013).

Rabe-Hesketh, S., A. Skrondal, et A. Pickles. 2004. « Generalized multilevel structural equation modeling », *Psychometrika*, vol. 69, no. 2, pp. 167–190.

———. 2005. « Maximum likelihood estimation of limited and discrete dependent variable

models with nested random effects », *Journal of Econometrics*, vol. 128, no. 2, pp. 301–323.

Shavit, Y. et H.-P. Blossfeld. 1993. *Persistent Inequality : Changing Educational Attainment in Thirteen Countries. Social Inequality Series.* ERIC.

Spady, W. G. 1967. « Educational Mobility and Access : Growth and Paradoxes », *American Journal of Sociology*, vol. 73, no. 3, pp. 273–286.

Statistique-Canada. 2005. *Enquête auprès des jeunes en transition - la cohorte de lecture de l'ejet - cycle 1 - guide de l'utilisateur.*

———. 2011a. Enquête auprès des jeunes en transition. Centre interuniversitaire québécois de statistiques sociales (distributeur).

———. 2011b. Enquête auprès des jeunes en transition : Définitions, sources de données et méthodes.

Wanner, R. A. 1999. « Expansion and Ascription : Trends in Educational Opportunity in Canada, 1920-1994 », *Canadian Review of Sociology/Revue canadienne de sociologie*, vol. 36, no. 3, pp. 409–442.

Warm, T. 1989. « Weighted likelihood estimation of ability in item response theory », *Psychometrika*, vol. 54, no. 3, pp. 427–450.