Essays in Applied Microeconometrics with Applications to Risk-Taking and Savings Decisions

Thèse

Steeve Marchand

Doctorat en économique
Philosophiæ doctor (Ph. D.)

Québec, Canada

Essays in Applied Microeconometrics with Applications to Risk-Taking and Savings Decisions

Thèse

Steeve Marchand

Sous la direction de:

Bernard Fortin, directeur de recherche
Vincent Boucher, codirecteur de recherche

Résumé

Cette thèse présente trois chapitres qui utilisent et développent des méthodes microéconométriques pour l'analyse de microdonnées en économique. Le premier chapitre étudie comment les interactions sociales entre entrepreneurs affectent la prise de décisions en face de risque. Pour ce faire, nous menons deux expériences permettant de mesurer le niveau d'aversion au risque avec de jeunes entrepreneurs ougandais. Entre les deux expériences, les entrepreneurs participent à une activité sociale dans laquelle ils peuvent partager leur connaissance et discuter entre eux. Nous recueillons des données sur la formation du réseau de pairs résultant de cette activité et sur les choix des participants avant et après l'activité. Nous trouvons que les participants ont tendance à faire des choix plus (moins) risqués dans la seconde expérience si les pairs avec qui ils ont discuté font en moyenne des choix plus (moins) risqués dans la première expérience. Ceci suggère que même les interactions sociales à court terme peuvent affecter la prise de décisions en face de risque. Nous constatons également que les participants qui font des choix (in)cohérents dans les expériences ont tendance à développer des relations avec des individus qui font des choix (in)cohérents, même en conditionnant sur des variables observables comme l'éducation et le genre, suggérant que les réseaux de pairs sont formés en fonction de caractéristiques difficilement observables liées à la capacité cognitive.

Le deuxième chapitre étudie si les politiques de comptes d'épargne à avantages fiscaux au Canada conviennent à tous les individus étant donné l'évolution de leur revenu et les différences dans la fiscalité entre les provinces. Les deux principales formes de comptes d'épargne à avantages fiscaux, les TEE et les EET, imposent l'épargne à l'année de cotisation et de retrait respectivement. Ainsi, les rendements relatifs des deux véhicules d'épargne dépendent des taux d'imposition marginaux effectifs au cours de ces deux années, qui dépendent à leur tour de la dynamique des revenus. J'estime un modèle de dynamique des revenus à l'aide d'une base de données administrative longitudinale canadienne contenant des millions d'individus, ce qui permet une hétérogénéité substantielle dans l'évolution des revenus entre différents groupes. Le modèle est ensuite utilisé, conjointement avec un calculateur d'impôt et de transferts gouvernementaux, pour prédire comment les rendements des EET et des TEE varient entre ces groupes. Les résultats suggèrent que les comptes de type TEE génèrent en général des rendements plus élevés, en particulier pour les groupes à faible revenu. La comparaison des choix d'épargne optimaux prédits par le modèle avec les choix d'épargne observés dans les

données suggère que les EET sont en général trop favorisées dans la population, surtout au Québec. Ces résultats ont d'importantes implications sur les politiques de « nudge »qui sont actuellement mises en œuvre au Québec, obligeant les employeurs à inscrire automatiquement leurs employés dans des comptes d'épargne de type EET. Ceux-ci pourraient produire des rendements très faibles pour les personnes à faible revenu, qui sont connues pour être les plus sensibles au « nudge ».

Enfin, le troisième chapitre étudie les problèmes méthodologiques qui surviennent fréquemment dans les modèles de régression par discontinuité (RD). Il considère plus précisément le problème des erreurs d'arrondissement dans la variable déterminant le traitement, ce qui rend souvent la variable de traitement inobservable pour certaines observations autour du seuil. Alors que les chercheurs rejettent généralement ces observations, je montre qu'ils contiennent des informations importantes, car la distribution des résultats se divise en deux en fonction de l'effet du traitement. L'intégration de cette information dans des critères standard de sélection de modèles améliore la performance et permet d'éviter les biais de spécification. Cette méthode est prometteuse, en particulier pour améliorer les estimations des effets causaux dans les très grandes bases de données, où le nombre d'observations rejetées peut être très important, comme le LAD utilisé au chapitre 2.

Abstract

This thesis presents three chapters that use and develop microeconometric methods for microdata analysis in economics. The first chapter studies how social interactions influence entrepreneurs' risk-taking decisions. We conduct two risk-taking experiments with young Ugandan entrepreneurs. Between the two experiments, the entrepreneurs participate in a networking activity where they build relationships and discuss with each other. We collect data on peer network formation and on participants' choices before and after the networking activity. We find that participants tend to make more (less) risky choices in the second experiment if the peers they discuss with make on average more (less) risky choices in the first experiment. This suggests that even short term social interactions may affect risk-taking decisions. We also find that participants who make (in)consistent choices in the experiments tend to develop relationships with individuals who also make (in)consistent choices, even when controlling for observable variables such as education and gender, suggesting that peer networks are formed according to unobservable characteristics linked to cognitive ability.

The second chapter studies whether tax-preferred saving accounts policies in Canada are suited to all individuals given they different income path and given differences in tax codes across provinces. The two main forms of tax-preferred saving accounts – TEE and EET – tax savings at the contribution and withdrawal years respectively. Thus the relative returns of the two saving vehicles depend on the effective marginal tax rates in these two years, which in turn depend on earning dynamics. This chapter estimates a model of earning dynamics on a Canadian longitudinal administrative database containing millions of individuals, allowing for substantial heterogeneity in the evolution of income across income groups. The model is then used, together with a tax and credit calculator, to predict how the returns of EET and TEE vary across these groups. The results suggest that TEE accounts yield in general higher returns, especially for low-income groups. Comparing optimal saving choices predicted by the model with observed saving choices in the data suggests that EET are over-chosen, especially in the province of Quebec. These results have important implications for "nudging" policies that are currently being implemented in Quebec, forcing employers to automatically enrol their employees in savings accounts similar to EET. These could yield very low returns for low-income individuals, which are known to be the most sensitive to nudging.

Finally, the third chapter is concerned with methodological problems often arising in regression discontinuity designs (RDD). It considers the problem of rounding errors in the running variable of RDD, which often make the treatment variable unobservable for some observations around the threshold. While researchers usually discard these observations, I show that they contain valuable information because the outcome's distribution splits in two as a function of the treatment effect. Integrating this information in standard data driven criteria helps in choosing the best model specification and avoid specification biases. This method is promising, especially for improving estimates of causal effects in very large database (where the number of observations discarded can be very large), such as the LAD used in Chapter 2.

Table des matières

Liste des tableaux

Liste des figures

À Laure, pour tout le bonheur de vivre cette aventure ensemble.

. . . [T]hrough pressure of conformity, there is freedom of choice, but nothing to choose from.

Peter Ustinov

Remerciements

J'ai passé la majeure partie de ma vie adulte à l'Université Laval, où j'ai fait toutes mes études universitaires et plus encore. J'y ai rencontré plusieurs personnes qui ont positivement bouleversé mes aspirations et ma vie.

Je remercie tout d'abord mon directeur de thèse Bernard Fortin de m'avoir guidé tout au long de mon doctorat. J'ai énormément bénéficié de sa volonté extraordinaire de transmettre son savoir et sa passion pour la recherche. Il a toujours su me donner envie de continuer et d'aller plus loin. Ses conseils furent précieux pour mener cette thèse à terme et ont fait de moi un meilleur chercheur. Je remercie également Vincent Boucher, mon codirecteur de thèse, pour son dévouement à aider ses étudiants. Ses conseils ont à plusieurs reprises redirigé ma recherche vers la bonne voie quand j'en avais le plus besoin.

Je tiens à remercier particulièrement Charles Bellemare, qui fut mon directeur de maitrise et avec qui j'ai continué à travailler sur plusieurs projets par la suite. Il a compris plus rapidement que moi que j'étais fait pour le monde de la recherche et a joué un rôle important dans ma décision d'entreprendre un doctorat. J'ai profité pendant tout mon parcours de sa porte toujours ouverte pour apprendre de lui et pour prendre conseil. Je remercie aussi Guy Lacroix qui, par son dévouement à son travail, a su trouver le temps nécessaire pour m'appuyer dans certains moments les plus décisifs de mon cheminement.

Je garde un remerciement spécial pour Luc Bissonnette. J'ai grandement profité de sa volonté de transmettre son savoir en économétrie et en programmation. Merci à Jean-Yves Duclos, avec qui j'ai travaillé avant et pendant le début de mon doctorat. Il m'a intégré à plusieurs projets grâce auxquels j'ai beaucoup appris et grandi. Je remercie aussi Pierre-Carl Michaud. Ce fut extrêmement formateur pour moi de travailler avec lui et de m'inspirer de son travail au cours des multiples projets auxquels nous avons collaboré. Je tiens également à remercier Thomas Lemieux d'avoir accepté d'évaluer cette thèse, ainsi que pour ses précieux commentaires.

Je remercie le *Conseil de recherches en sciences humaines du Canada* (CRSH) et le *Fonds de Recherche du Québec - Société et culture* (FRQSC) pour mes bourses de doctorat. Je remercie également la *Chaire de recherche Industrielle Alliance sur les enjeux économiques des changements démographiques* pour le soutien financier.

Merci à toute ma famille, qui m'a appuyé sans réserve dans mon parcours. Finalement, merci à Laure pour tout ce temps à m'appuyer, à m'encourager, à me supporter dans les moments de stress et à célébrer les réussites. Je ne peux exprimer à quel point son soutien a compté dans cette aventure.

Avant-propos

Cette thèse comprend trois chapitres. Le chapitre 1 est écrit avec conjointement avec Maria Adelaida Lopera, ancienne étudiante de doctorat au département d'économique de l'Université Laval et actuellement affiliée au « Partnership for Economic Policy »(PEP). Ce chapitre a été publié en juin 2018 dans la revue *Journal of Economic Behavior and Organization*. Elsevier, l'éditeur de cette revue, permet aux auteurs d'ajouter leur article publié dans leur thèse (voir https ://www.elsevier.com/about/policies/copyright/permissions) et la diffusion de cette thèse. Comme cela est la norme en économique, les deux auteurs ont contribué à parts égales à toutes les étapes de l'article. L'ordre des auteurs figurant dans l'article publié est alphabétique. Les chapitres 2 et 3 sont écrits sans coauteur(e)s et ne sont pas encore publiés.

Introduction

Applied microeconometrics – or the application of econometric methods to microdata – has been a very active field in the last decades and continues to be. Compared to other fields, it continuously benefits from the never-ending creation of new, larger and better databases. What is more, economists increasingly combine microeconometrics methods to other methods, such as theoretical models, field experiments or laboratory experiments to provide innovative research designs that allow to answer new questions. This thesis is an example of how these innovative methods and larger than usual databases can provide new insights to important questions, applying different methodologies to analyze risk-taking and savings decisions. It also explores new econometrics methodologies that seek to improve the analysis of microdata in general.

In the first chapter, I combine a theoretical model of social conformity, laboratory experiments that allow to measure risk aversion, and data on a peer network created during a field experiment conducted with young entrepreneurs in Uganda. The field experiment involved among other things creating a network of entrepreneurs so that they may develop new relationships, share knowledge and learn from each other. I show how this innovative research design allows to answer questions usually difficult to tackle with more standard designs. The design allows to estimate separately identify homophily effects (i.e. the tendency of individuals to develop relationships with peers similar to themselves) from social conformity effects on entrepreneurs' risk-taking decisions in Uganda. I can also identify social learning effects, and thus measure whether the entrepreneurs tend to learn from each other to make more sound decisions when facing risk. This chapter shows how combining multiple modern methodologies allows to answer new questions on populations that are the most relevant.

The second chapter is a good example how very rich data, combined with microeconometric analysis, can improve our understanding of important questions and eventually guide public policies. I use a very rich Canadian longitudinal database in which individuals are observed up to 31 years in a row. The data comprises millions of individuals per year. The richness of this database allows me to estimate a model of income dynamics that allows for substantial heterogeneity in evolutions of careers. I combine the predictions from this model with a calculator of taxes and government transfers, which allows me to predict optimal savings choices in

tax-preferred savings accounts, and how these optimal choices vary across provinces, gender, income groups or family status. The results provide insights that should prove useful to guide policy. For example, I find that Tax-Free Savings Accounts (TFSA) tend to provide better returns than Registered Retirement Savings Plans (RRSP) for a large share of the population (especially for low-income individuals), but that TFSAs are underused in Quebec compared to other provinces. Still, the government of Quebec encourages individuals to save more in new savings vehicles similar to RRSPs, which will likely lead to poor savings choices in the lowest income groups.

The third chapter seeks to improve the methodology used in regression discontinuity designs (RDD) – one of the most widely used microeconometric methods that economists use to estimate causal effects. In those designs, a problem that often arises in practice is that the variable that determines whether an individual receives the treatment of interest is rounded. This usually leads researchers to discard many observations. I show that these observations are nevertheless useful and can help correcting problems caused by rounding errors. This chapter adds to the applied microeconometric's toolbox and should prove useful especially with very large databases, which are increasingly used in economics.

Chapitre 1

Peer Effects and Risk-Taking Among Entrepreneurs : Lab-in-the-Field Evidence[1]

1.1 Résumé

Nous étudions comment les interactions sociales influencent les décisions prises par les entrepreneurs en matière de risque. Nous menons deux expériences de prise de risque avec de jeunes entrepreneurs ougandais. Entre les deux expériences, les entrepreneurs participent à une activité de réseautage où ils établissent des relations et discutent entre eux. Nous recueillons des données sur la formation de réseaux de pairs et sur les choix des participants avant et après l'activité de réseautage. Nous constatons que les participants ont tendance à faire plus (moins) de choix risqués dans la seconde expérience si les pairs avec lesquels ils discutent font en moyenne plus (moins) de choix risqués dans la première expérience. Cela suggère que même les interactions sociales à court terme peuvent affecter les décisions de prise de risque. Nous constatons également que les participants qui font des choix (in)cohérents dans les expériences ont tendance à développer des relations avec des individus qui font aussi des choix

1. This chapter is co-written with Maria Adelaida Lopera and was published at the Journal of Economic Behavior and Organization (see Lopera and Marchand (2018)). We thank Charles Bellemare, Luc Bissonnette, Vincent Boucher, Bernard Fortin, and two anonymous referees for useful comments that greatly improved the quality of this paper. This study was carried out with financial and scientific support from the Partnership for Economic Policy (PEP) (www.pep-net.org) and funding from the Department for international Development of the UK Aid and the government of Canada through the International Development Research Centre. We are especially grateful to the team of PEP-researchers (PIERI-12451) led by Juliet Ssekandi, who allowed us to join her evaluation project to collect experimental data. We also thank Benjamin Kachero and Samuel Galiwango for their extraordinary assistance in the field. We are grateful for the support provided by the PEP Research Director of Experimental Impact Evaluations Maria Laura Alzua, and by the PEP Scientific Advisor John Cockburn. This research benefited from collaborations with the Department for Children and Youth at the Ministry of Gender, Labor and Social Development (MGLSD), UNICEF-Uganda and Enterprise Uganda. We thank the Fonds de recherche du Québec - Société et culture and the Social Sciences and Humanities Research Council for our scholarships.

(in)cohérents, même en conditionnant sur des variables observables telles que l'éducation et le sexe, suggérant la formation de réseaux selon des caractéristiques inobservables liées à la capacité cognitive.

1.2 Abstract

We study how social interactions influence entrepreneurs' risk-taking decisions. We conduct two risk-taking experiments with young Ugandan entrepreneurs. Between the two experiments, the entrepreneurs participate in a networking activity where they build relationships and discuss with each other. We collect data on peer network formation and on participants' choices before and after the networking activity. We find that participants tend to make more (less) risky choices in the second experiment if the peers they discuss with make on average more (less) risky choices in the first experiment. This suggests that even short term social interactions may affect risk-taking decisions. We also find that participants who make (in)consistent choices in the experiments tend to develop relationships with individuals who also make (in)consistent choices, even when controlling for observable variables such as education and gender, suggesting that peer networks are formed according to unobservable characteristics linked to cognitive ability.

1.3 Introduction

Risk plays a fundamental role in economic decision-making. For instance, evidence suggests that entrepreneurship is associated with a higher than average tolerance toward risk (Cramer et al., 2002 ; Ekelund et al., 2005 ; Ahn, 2010). Risk preferences may also affect businesses' success rates conditional on entry (Caliendo et al., 2010). But do individuals make risk-taking decisions solely according to their own risk preferences, or are there other important determinants of these choices ? In this paper, we study the role of social interactions on risk-taking among groups of entrepreneurs. Using an original experimental design, we find a significant impact of conformity on risk-taking. Our findings suggest that even short-term social interactions are sufficient to affect entrepreneurs' risk-taking behaviors.

Entrepreneurs face more risk-taking decisions than paid employees in their daily life, which makes them a particularly interesting population to study the determinant of risk-taking. To focus on this population, we conducted lab-in-the-field experiments on risk-taking within workshops organized for young entrepreneurs in Uganda. Conducting these experiment in a developing country allows to incentivize participants with large amounts relatively to their income. [2] The workshops included a networking activity where entrepreneurs develop new relationships and converse with each other. We collected detailed information on who participants

2. For example, as we state latter in the paper, the highest possible payoff in one of our experiment is $10,000$ Ugandan shillings, which represents more than 16 hours of work at Uganda's 2012-13 median wage.

conversed with during this activity. The entrepreneurs also participated in two risk-taking experiments : one before and one after the networking activity. These two experiments are adaptations of the well-known Holt and Laury (2002) multiple choice lotteries designed to measure risk aversion. The two experiments, combined with data on the peer network formation, provide an innovative experimental design that allows us to capture the causal effect of social interactions on entrepreneurs' choices with respect to risk.

We find significant social conformity effects : Participants tend to make more (less) risky choices in the second experiment if their peers made on average more (less) risky choices in the first experiment. This suggests that social interactions may counterbalance individual risk preferences. Given some risk preferences, an entrepreneur could become more (less) inclined to take risk following a relatively short discussion with an entrepreneur who is more (less) risk tolerant. In the second experiment, part of the participants were assigned to an experiment that included an ambiguity component (i.e. uncertainty on the exact probabilities linked to the lotteries' outcomes). As pointed out by Klibanoff et al. (2005), the uncertainty on the probabilities in the lotteries gives more room for subjective expectations to affect decisions. It is possible that social influence affects these subjective believes differently than attitude toward pure risk. [3] We also distinguish between preferences to conform with *successful* peers (who made the choice that led to the highest payoff given the lotteries' results) from preferences to conform with *unsuccessful* peers (who made the choice that led to the lowest payoff given the lotteries' results). Under pure risk, we find that participants tend to conform with successful peers, but not with unsuccessful ones. However, when the experiment includes an ambiguity component, we find that participants tend to conform with their peers regardless of the outcome.

Our design allows us to control for homophily, which is commonly a challenge in the estimation of peer effects. Homophily is the tendency of individuals to develop relationships with people similar to themselves. This behavior creates a correlation between one's peer variable (e.g. peers' average outcome) and his own choice even in the absence of peer effects, leading to identification issues. Attanasio et al. (2012) present evidence that individuals form social networks according to similarities in risk attitudes. However, in their context, as opposed to ours, individuals form networks with the objective of pooling risk. Thus, it is not necessarily the case that this behavior will also occur in our context. Nevertheless, individuals could still develop relationships according to some factors that also affect risk preferences. In other words, the peer network formation may be endogenous. There is a large and expanding literature that seeks to control for endogenous networks (for example, see Goldsmith-Pinkham and Imbens, 2013 ; Arduini et al., 2015 ; Qu and Lee, 2015 ; Boucher, 2016 ; Hsieh and Lee, 2016). However,

3. A paper investigating how risk attitudes may change with and without ambiguity is Cohn et al. (2015). They find that ambiguity causes no differences in how their treatment (showing participants a graph of stock market boom or crash) affects risk attitude. They interpret this finding as evidence that their treatment affects pure risk preferences, and not subjective expectations.

controlling for endogeneity necessarily requires strong assumptions. [4] Our design allows us to identify peer effects in the presence of homophily under weaker assumptions. We use choices made in the two experiments to control for time-invariant individual characteristics through a first-difference approach. Assuming that individuals develop relationships based on these time-invariant characteristics is sufficient to rule out that the relationship between one's choice and those of her peers is caused by homophily. Furthermore, we can directly test for homophily effects. The choices made in our first experiment cannot possibly result from peer effects, because this experiment takes place before the networking activity. Therefore, the observed similarities between individuals' choices and those of the future peers they have not yet met can be used to identify homophily effects. We find no evidence of homophily according to characteristics that affect risk choices.

We also study the impact of social interactions on the consistency of individuals' choices. Indeed, in multiple choice lotteries experiments, some combinations of choices are inconsistent with standard risk preferences. We therefore test for homophily effects according to characteristics that affect the consistency of choices. We find that participants who make (in)consistent choices tend to develop relationships with individuals who also make (in)consistent choices. We finally test for social learning peer effects that would cause individuals to make more consistent choices if the peers they met made more consistent choices. We find no evidence of such social learning effects.

We contribute to the literature on the determinants of risk-taking, as well as the literature on peer effects and risk-taking. Firstly, there is a growing literature that suggests risk attitude vary across contexts (Barseghyan et al., 2011) and over time (Baucells and Villasís, 2010). [5] Understanding the factors that drive these variations is of particular importance to understand decisions about becoming an entrepreneur. Evidence suggests that family dynamics are important in shaping individuals' preferences toward entrepreneurship. Dunn and Holtz-Eakin (2000) find that parental entrepreneurial experience is a stronger predictor of entrepreneurship than individual or parental wealth. This correlation may result from both *nature* and *nurture* factors, but evidence suggests nurture factors play a larger role (Lindquist et al., 2015). The social context outside of the family can also shape individuals' attitudes toward risk and entrepreneurship, or their beliefs or confidence about the expected returns of starting a business. For instance, having entrepreneurial peers could create non-monetary benefits of running a business (Giannetti and Simonov, 2009). Nanda and Sørensen (2010) find that individuals are more likely to become entrepreneurs if they work with peers who have previously been entrepreneurs. They argue that past workers' experience may spill over to their coworkers

4. For example, Goldsmith-Pinkham and Imbens (2013) assume that there exist two unobserved types of individuals and that those of the same type have a greater probability to become peers. Together with other distributional assumptions, this allows them to write the joint likelihood of the observed outcomes and peer network.

5. Risk attitude may also be affected by emotional states such as joviality, sadness, fear and anger (Conte et al., forthcoming), or by stress (Cahlíková and Cingl, 2017).

by influencing their entrepreneurial skills, knowledge or motivation. Our paper explores the complementary idea that entrepreneurs' risk attitude may also spill over to others through peer effects. Secondly, our paper contributes to the expanding literature on peer effects on decisions made under risk. Bursztyn et al. (2014) study peer effects on the purchase of financial assets in a field experiment conducted at a financial brokerage. They find evidence of peer effects driven by both social learning (i.e. learning from peers) and social utility (i.e. utility that results directly from a peer's possession of an asset). Ahern et al. (2014) conduct an experiment about peer effects on risk aversion among MBA students and find significant peer effects. Gioia (2016) conducts a lab experiment and finds that the intensity of peer effects on risk-taking is determined in part by group identity : when peers are matched according to interest, the influence they exert on each other is greater. This suggests that peer effects might be important in our context, as our participants all share a common entrepreneurial identity. Our paper adds to these literatures by being the first (to our knowledge) to isolate the causal effect of interactions with peers on risk-taking decisions within a sample of entrepreneurs. Our paper further distinguishes itself in that it suggests that risk-taking can be influenced by peers in the very short run, following a networking activity a few hours only.

Another related paper is Lahno and Serra-Garcia (2015), who conduct a laboratory experiment to investigate whether participants' decisions about risk are influenced by their peers. They find that peer effects on risk-taking seem to be driven by a desire to conform with peers' choices. They argue that this implies that policymakers who seek to influence behaviors related to risk-taking (e.g. decisions to purchase insurance or acquire or repay debt) could publicly inform others about choices made by the population. This implication is particularly relevant for our paper, as we study real entrepreneurs. Our participants are people who need to finance their business projects with loans (this is discussed in detail in the next section). A policymaker could easily inform entrepreneurs about borrowing or insurance choices made by other entrepreneurs (for example, in an activity organized for them such as our workshops). He could also decide to make certain choices public in order to encourage specific behaviors (e.g. posting only the names of entrepreneurs who choose to insure their business). The policymaker could finally create networking activities aimed at discussing risk-taking decisions. These activities may generate social conformity effects that would push behaviors toward the average behavior, reducing excessive risk-taking and increasing risk tolerance for excessively risk averse individuals.

The next section describes our experimental design and data. Section 1.5 models participants' risk choices and presents the estimation of the social conformity effects. Section 1.6 models participants' consistency of choices and estimates social learning peer effects. Section 1.7 tests for homophily effect, and Section 1.8 concludes.

1.4 Experimental Design and Data

1.4.1 The Workshops

We contributed to the organization of six two-day workshops, along with the Partnership for Economic Policy,[6] a group of local researchers and UNICEF Uganda. The workshops took place in early 2014 in several locations in Uganda.[7] Their primary aim was to evaluate and improve financial literacy among young Ugandan entrepreneurs. The workshops included training in finance and business planning, as well as a networking activity where entrepreneurs could share their knowledge with each other. Within each workshop, we ran two experiments on risk-taking : one before and one after the networking activity.

Entrepreneurs were recruited using U-report, a free Short Message Service (SMS) platform created and managed by UNICEF to engage Ugandan youth into policymaking and governance. In 2014 the platform counted around 200,000 subscribers across Uganda.[8] The first contact was an SMS message asking, "Are you an entrepreneur below 35 years old ?"[9] If the answer was affirmative, a second SMS message was sent : "Would you be interested in obtaining a credit loan from the Youth Venture Capital Fund ?" This question aimed at selecting only entrepreneurs who were considering a business loan. If the answer was affirmative again, the potential participant received a phone call from a recruiter. The recruiter asked whether the potential participant was available for a two-day workshop near his/her home. Interested individuals were invited to the workshop, and the potential participant either accepted or rejected the invitation.

In total, 540 entrepreneurs participated in one of the workshops. Upon arrival, participants completed a survey about their sociodemographic characteristics, registered using their full name and were attributed an identification number. All subjects then participated in an initial risk-taking experiment, which we describe in the next subsection. After this experiment, subjects proceeded to the networking activity, which included a lunch and a discussion time. All participants in a given workshop were in the same room for both the lunch and the discussion time, which together lasted three to four hours. We provided them no information on what would happen after the networking activity. They did not know they would play a second experiment at this time, so they had no incentive to seek information from their peers that would guide them in their choices for the second experiment. This strategy allows us to observe interactions occurring naturally without guidance from the experimenter. Throughout the activity, participants wore a tag indicating their full name and identification number. They had to write the name and identification number of at most seven participants with whom they

6. www.pep-net.org.

7. Four workshops took place in the districts of Wakiso, M'bale, Gulu and M'barara. The other two workshops took place in the capital city of Kampala.

8. The average age was 24 years old and 23% were female. Interested readers can visit www.unicef.org/uganda/voy.html for more information about the U-report platform.

9. We sent a total of 2,278 text messages in large cities.

had spent the most time chatting, thus allowing us to record their peer network. They also had to identify each relationship as either an extended family member, a friend from before the workshop, or a person they met at the workshop. Once all participants had registered this information, a random sample of half the participants in each workshop (258 in total) was chosen to participate in a second risk-taking experiment, also described in the next subsection. [10] The procedure of the random selection was to hide a label - either blue or red- inside each participant's tag prior to the workshop. After the networking activity, participants were asked to look at their attributed color, and those with a given color had to play the second experiment. The first day of the workshop then ended, participants were paid the amount they had won in the experiments, and then returned home. The second day of the workshop included training in finance and business planning, which are outside the scope of this paper.

Two points are important to note. First, we provided no indication regarding what participants should discuss during the activity. They were completely free to discuss, or not to discuss, the experiment they had played. This makes the social interaction effects that we estimate latter in this paper more authentic : they occur naturally in a setting that resembles the real world. It is of course likely that some participants have not revealed any information that may affect their peers' choices for the second experiment. Thus, the peer effects we will present may understate the peer effects that would arise in a full information setting in which participants would precisely know choices made by their peers the first time. Second, although the targeted participants declared being interested in a credit loan, we believe it is unlikely that participants were concerned about potential effect of their choices on the loan. We, as well as the other organizers of the workshops, were not offering loans ourselves, and there was no link between us and the institutions that could grant this loan. Participants who would decide to get a loan would have to contact the institution of their choices by themselves.

1.4.2 The Risk-Taking Experiments

All subjects participate in the first risk-taking experiment, which takes place before the networking activity. The experiment is an adapted version of the well-known Holt and Laury (2002) experiment designed to measure risk preferences. It consists of nine games in which participants must choose between two lotteries : a safe lottery or a risky lottery, with the risky lottery having more variability between the potential payoffs. Each game is presented to participants in the form of a big transparent box containing 40 large white and black balls. The white balls represent low payoffs and black balls represent high payoffs. The proportion of black balls is low in the first game and increases in each subsequent game. Participants also receive a paper questionnaire that provides them with the exact proportion of the two colors in each box. Participants are told that there are no good or wrong answers so that they do not feel that the experimenter is monitoring them. They are finally told that after all decisions

10. Participants who were not selected for the second experiment received training in finance and business planning that was also part of the workshop, but which we do not address in this paper.

are made, only one box (only one of the nine games) will be selected at random, with one ball selected at random from inside that box. They will then be paid according to this ball's color and the choice they made in the corresponding game. Each within-workshop experiment is split in three or four sessions (depending on the number of participants in the workshop), and the lottery and ball that are selected are specific to each session. This is done to create variation in the amount won across participants conditional on choices made. Decisions are made individually and participants are not allowed to consult each other. Appendix 1.B provides additional details about how the experiment is presented to participants.

Table 1.1 presents the two possible payoffs for each lottery. The amounts are substantial. For example, 10,000 Ugandan shillings (UGX), the highest possible payoff, represents more than 16 hours of work at Uganda's 2012-13 median wage. [11]

Table 1.2 shows the probability that the high payoff ball is picked for each game. It is low in the first game and increases for each game, so that the incentive to choose the risky lottery increases in each game. The last column shows the difference in expected payoffs between choosing the safe lottery and choosing the risky lottery. The combination of choices made by an individual is informative of his preferences. For example, a risk-neutral individual should choose the safe lottery in games 1 to 4, and then switch to the risky lottery in games 5 to 9. Our main variable of interest — the number of safe choices – is the number of games in which the individual chooses the safe lottery. It ranges from 0 (all risky choices) to 9 (no risky choices). A risk-neutral individual should therefore make four safe choices, because he would choose the safe lottery from games 1 to 4.

In theory, a participant should not switch his choice more than once. That is, if a participant chooses the safe lottery in game k and the risky lottery in game $k+1$, it would be inconsistent to switch back to the safe lottery in game $k+2$. In practice, in our experiment as in other studies, some participants do switch more than once. [12] This could be the result of a participant misunderstanding the experiment or having difficulty calculating the expected outcomes of each lottery. As pointed out by Andersen et al. (2006), it could also result from participants being indifferent between choices of lotteries, which requires preferences to be weakly convex rather than strictly convex. Still, in the following sections, we will refer to a second outcome of interest : the consistency of choices (i.e. consistent with strictly convex preferences), a dummy variable that equals one if the participant switches no more than once, and zero otherwise.

In the second experiment (after the networking activity), within each workshop, each participant is randomly assigned to one of two subgroups. This creates 12 subgroups in total. Some

11. The median monthly earnings in Uganda was about 110,000 UGX in 2012-13 for a paid employee, with the average work week comprised of approximately 41 hours. Because a month comprises 4.35 weeks on average, the average hourly earningsare about 617 UGX per hour (see page 12 of the Uganda National Household Survey of 2012-13 (UBOS, 2014).

12. For example, see Holt and Laury (2002) and Jacobson and Petrie (2009).

TABLE 1.1 – Game payoffs (in UGX)

	Return	
	Low	High
Safe lottery	4,000	6,000
Risky lottery	1,000	10,000

TABLE 1.2 – Probability of high payoff in each game

Game	Probability of high payoff	Expected payoff difference : safe - risky (in UGX)
1	1/10	2,300
2	2/10	1,600
3	3/10	900
4	4/10	200
5	5/10	-500
6	6/10	-1,200
7	7/10	-1,900
8	8/10	-2,600
9	9/10	-3,300

subgroups replay the original experiment. The other subgroups play three different versions of the experiment, where we introduce an ambiguity component. For these groups, in the second experiment, a small proportion of the balls are wrapped in opaque bags so that participants cannot see whether they are black or white. The proportion of balls of unknown color in the low, medium and high ambiguity groups are 5%, 10% and 15% respectively and remain fixed in all nine games. Participants are not provided any information about the distribution of the colors of the hidden balls. As for the balls that are not hidden, the proportions of white and black balls remain as described in Table 1.2. Following Klibanoff et al. (2005), the uncertainty on the exact share of high payoff balls leaves more room for subjective expectations to affect decisions, so that social influence may affect more (or less) strongly these subjective beliefs than attitude toward pure risk. As we will see in Section 1.5, we will test whether there are any difference in peer effects when individuals face ambiguity. The lottery and ball that are selected for this second experiment are specific to each subgroup. Appendix 1.B provides details on all the experiments.

1.4.3 Data

Table 1.3 summarizes the data collected from the sociodemographic questionnaire, peer network questionnaire and the two risk-taking experiments' results. The average number of safe choices in the first experiment is 4.61 and slightly increases to 4.81 in the second experiment. The standard deviation of the differences in participants' number of safe choices in the two experiments is 1.81. This indicates that the number of safe choices varies upward and downward between the two experiments, even though the aggregate change is relatively small. The

proportion of participants who make consistent choices in the first experiment is 54% and increases to 69% in the second second experiment. This increase could, among other things, be the result of playing the game a second time or of social learning effects.

On average, participants identify 4.52 peers who they met at the workshop and 1.76 peers who they knew before the workshop. Although we do not distinguish between these two types of peers in our main results, Appendix 1.A shows that the significance of the peer effects we estimate in Section 1.5 mainly results from interactions between peers who have met at the workshop, ruling out the concern of social interactions that could have occurred before the networking activity (this is discussed in Section 1.5.3).

Table 1.4 decomposes the averages of our two outcomes of interest, for both experiments, for each type of second experiment played. The first column presents the first experiment's outcomes for those who did not play a second experiment, while the other four columns present the outcomes for the two experiments for those who played a second experiment with no, low, medium or high level of ambiguity.[13] The average number of safe choices increases in the second experiment for all experiments with ambiguity, although there is no obvious trend relating to the level of ambiguity. As for the proportion of participants who made consistent choices, it increases in all experiments, with no clear trend regarding the level of ambiguity.

Table 1.5 presents the bounds of risk aversion parameters that are implied by the observed choices assuming a constant relative risk aversion (CRRA) utility (i.e. $U(x) = x^{1-r}/(1-r)$). This allows us to compare our results with Holt and Laury (2002) and the literature that followed. Note that CRRA utility is consistent with an individual's observed choices only if he made consistent choices (i.e. switched no more than once). Nevertheless, the table presents the proportion of each number of safe choices for all participants (both those who made consistent and inconsistent choices) and relate this number to risk aversion assuming the x safe choices are made for the x first games. The third column shows the proportion for all participants in the first experiment, while the fourth column shows this proportion only for participants in the second experiment who played a game without ambiguity. This is also done by Holt and Laury (2002), who argue that inconsistent choices may simply result from errors around the "true" switching point of the individual. We find that a high concentration of choices in the $r \in (-0.1, 0.56)$ range (50% in the fist experiment and 60% in the second). Figure 1.1 compares our cumulative distribution of CRRA risk aversion measures to those of Holt and Laury (2002). Since we offer substantial payoffs, we compare our results with their high payoff experiment. Note that it is not possible to bound upward the value of r for individuals who only made risky choices, so the cumulative distribution does not reach 100%. Our participants are clearly less risk averse than those of Holt and Laury (2002). This could be due to a sorting

13. Average outcome values may systematically differ even in the first experiment, since the ambiguity level varied across workshops. As the workshops were held in different cities, participants may tend to differ in their risk attitude across workshops.

TABLE 1.3 – Summary statistics

	Mean	SD	Min.	Max.	Obs.
Number of safe choices (0 to 9)					
1st experiment	4.61	1.86	0	9	540
2nd experiment	4.83	1.91	0	9	258
Difference between 2nd and 1st	0.26	1.81	-6	7	258
Consistence of choices (0 or 1)					
1st experiment	0.54	0.50	0	1	540
2nd experiment	0.69	0.46	0	1	258
Difference between 2nd and 1st	0.16	0.56	-1	1	258
Experiments' payoffs (in UGX)					
1st experiment	5,025	3,193	1,000	10,000	540
2nd experiment	4,852	3,184	1,000	10,000	244
Number of peers					
Met at the workshop	4.52	2.35	0	7	540
Family, friends, other	1.76	2.15	0	7	540
Age	26.63	4.41	17	50	540
Male	0.82	0.38	0	1	540
Education level					
Primary	0.14	0.34	0	1	540
Secondary	0.30	0.46	0	1	540
Technical	0.30	0.46	0	1	540
University	0.26	0.44	0	1	540
City					
Kampala 1	0.17	0.37	0	1	540
Kampala 2	0.14	0.35	0	1	540
Wakiso	0.17	0.37	0	1	540
M'bale	0.19	0.39	0	1	540
Gulu	0.19	0.39	0	1	540
M'barara	0.15	0.35	0	1	540
Ambiguity level in 2nd exp.					
None	0.19	0.40	0	1	258
Low	0.33	0.47	0	1	258
Medium	0.30	0.46	0	1	258
High	0.17	0.38	0	1	258

effect : risk tolerant individuals can be more prone to becoming an entrepreneur. We therefore also compare our results to those of the high payoff treatment of Bellemare and Shearer (2010), who conduct similar experiments on workers who face substantial income risk. Consistently with sorting, our results are closer to theirs. The cumulative distributions are near equal below a risk aversion of 0.2. Among individuals with higher risk aversion, our participants are slightly more risk averse than theirs.

After the second experiment, we asked participants to identify the main reason why they changed their choices between the two experiments (if they did change their choices). Table 1.6 presents the frequency of each possible answer among participants who reported having changed their choices. Almost 42% answered that the discussions they had with their peers during the networking activity had changed their mind. This suggests that participants discussed the experiment and choice strategies during the networking activity, even though we did not instruct them to. It also suggests that they influenced each others in these discussions.

TABLE 1.4 – Summary statistics by type of 2nd experiment played

	None	**No amb.**	**Low**	**Med.**	**High**
		Mean value of outcome			
Number of safe choices (0 to 9)					
1st experiment	4.65	4.36	4.76	4.64	4.29
2nd experiment	-	4.38	5.06	5.00	4.60
Difference between 2nd and 1st	-	0.02	0.29	0.36	0.31
Consistence of choices (0 or 1)					
1st experiment	0.55	0.48	0.54	0.49	0.62
2nd experiment	-	0.74	0.65	0.71	0.69
Difference between 2nd and 1st	-	0.26	0.11	0.22	0.07
Number of observations	282	50	85	78	45

TABLE 1.5 – Implied risk aversion parameter from a CRRA utility function (only experiments without ambiguity)

Number of safe choices	Range of relative risk aversion for $U(x) = x^{1-r}/(1-r)$	Proportion of choices First exp.	2nd exp. (no amb.)
0	$r < -1.68$	0.02	0.04
1	$-1.68 < r < -0.94$	0.03	0.02
2	$-0.94 < r < -0.47$	0.05	0.08
3	$-0.47 < r < -0.1$	0.14	0.10
4	$-0.10 < r < 0.23$	0.27	0.30
5	$0.23 < r < 0.56$	0.23	0.30
6	$0.56 < r < 0.89$	0.10	0.04
7	$0.89 < r < 1.29$	0.08	0.04
8	$1.29 < r < 1.85$	0.05	0.04
9	$1.85 < r$	0.03	0.04
Number of obs.		540	50

FIGURE 1.1 – Comparison of CRRA risk aversion measures

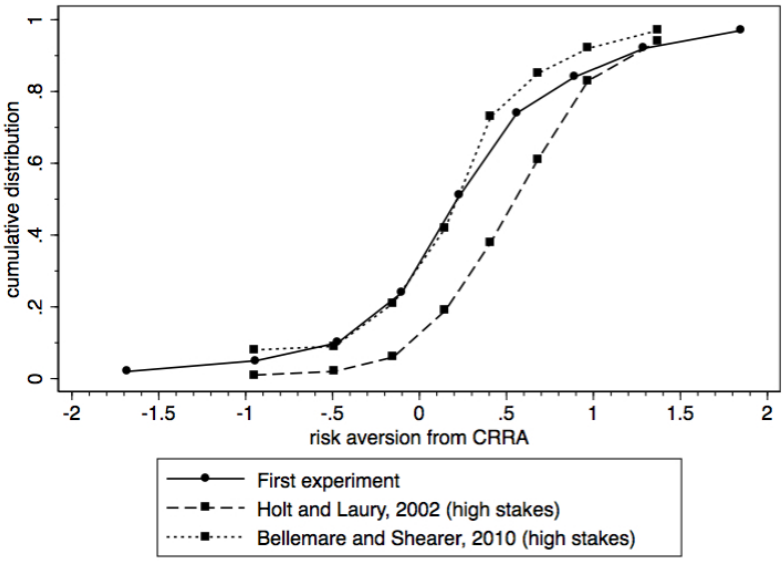


TABLE 1.6 – Self-reported reasons for changing choices in the 2nd experiment

Why did you change any of your choices ?	**Freq.**	**Percent**
I did not understand the first time	18	12.08
The game was different	49	32.89
Discussions with others changed my mind	62	41.61
I lost the first time	20	13.42
Total	149	100

1.5 Social Conformity and Risk-Taking Decisions

1.5.1 The Empirical Models

Participants' choices involve choosing between safe and risky lotteries. We therefore let the choice variable be y_{ir}, the number of safe choices individual i made in experiment $r \in \{1, 2\}$, where $r = 1$ is the first experiment (before the networking activity) and $r = 2$ is the second (after the networking activity). We model utility as a trade-off an individual faces : making choices according to his own characteristics and underlying preferences, or according to his or her peers' choices. As in Brock and Durlauf (2001b), Brock and Durlauf (2001a), Bisin et al. (2006) and Boucher (2016), we use a quadratic utility function to model this trade-off. The utility function penalizes the individual more if he chooses a value of y_{i1} that is further from his characteristics, as well as if he chooses a value that is further from average choices of his peers. In the first experiment, however, individuals do not face the trade-off because they do not know their peers' choices, so participants simply choose y_{i1} according to their own

characteristics :

$$U_{i1}(y_{i1}) = -\frac{1}{2}(y_{i1} - \alpha_1 - \mathbf{x}_i\boldsymbol{\beta} - \eta_i - \epsilon_{i1})^2, \tag{1.1}$$

where $\mathbf{x}_i$ is a vector of individual i's observed characteristics. The parameters $\boldsymbol{\beta}$ and η_i are respectively the effect of the individual's observed and unobserved characteristics and may therefore capture the individual's idiosyncratic risk preferences. Thus, we allow for these preferences to be specific to the individual and to be a function of individual characteristics. This is consistent with the literature, which finds differences in risk preferences across individuals (for example, see Croson and Gneezy (2009), who find gender-based differences in risk preferences). Both $\mathbf{x}_i$ and η_i are constant over time (i.e. $\forall r \in \{1, 2\}$). The error term ϵ_{i1} is specific to i and to the first experiment. It allows for shocks, such as stress or other emotions, which might temporally affect choices under risk (see Cahlíková and Cingl (2017) and Conte et al. (forthcoming). The error term also acknowledges that we do not directly observe risk preferences, but rather an imperfect measure of it. [14] Thus, in the spirit of Baucells and Villasís (2010), the number of safe choices y_{i1} could be the result of both risk preferences and a random error component. Note that this utility function does not intend to measure individuals' values of a structural parameter of risk preferences. [15] The first-order condition is :

$$y_{i1} = \alpha_1 + \mathbf{x}_i\boldsymbol{\beta} + \eta_i + \epsilon_{i1}. \tag{1.2}$$

In the second experiment ($r = 2$) after the networking activity, participants face a trade-off between staying true to their own characteristics and conforming with their peers' choices. We model social conformity using two specifications : homogeneous peer effects, where participants partly conform with the average behavior of their peers, and heterogeneous peer effects, where participants may conform differently with different peers according to the first experiment's results.

Before we present our modelization of peer effect, it is important to clarify what we mean by social conformity. Our model is designed to capture a tendency to make choices that are closer to peers' average choices. Although we refer to these peer effects as social conformity effects, we cannot completely rule out that they capture other types of peer effects than a pure preference to conform. For example, participants could have no preferences to conform, but still be influenced by peers' average choices through learning effects. Peer effects could also possibly arise from a taste for competition among peers. Still, throughout this section, we refer to the peer effects we find as social conformity effects because we believe this is the most convincing

14. Preference elicitation methods other than Holt and Laury (2002) lotteries could lead to different measures (Anderson and Mellor, 2009).

15. As seen in Section 1.4.3, using utility functions integrating these parameters such as a CRRA only allows to bound the parameters. This lack of point identification would greatly complexify the identification of additional social interaction parameters. In our view, our utility function is the simplest estimable empirical model that acknowledges that utility is a trade-off between the individuals' own characteristics and their peers' choices.

mechanism in our setting. In Section 1.6, we explore a separate social learning effect affecting the consistency of choices and find no significant effect. In our view, this makes social learning effects on risk preferences less convincing as well. Regarding potential competition effects, we estimate in Section 1.5.3 separate peer effects depending on the first experiment's results and argue that these estimations provide no evidence of competition effects. Nevertheless, it is worth keeping in mind that the effects we present in this section could possibly also capture a tendency to conform to peers' average choices resulting from social learning or competition.

Homogeneous peer effects specification

We assume that individual i in the second experiment maximizes the following utility function :

$$U_{i2}(y_{i2}) = -\frac{1}{2}(y_{i2} - \alpha_1 - \alpha_2 - \alpha_2^g - \mathbf{x}_i\boldsymbol{\beta} - \delta W_i - \eta_i - \epsilon_{i2})^2 - \frac{\theta}{2}\left(y_{i2} - \frac{1}{n_i}\sum_{j\in N_i} y_{j1}\right)^2 \quad (1.3)$$

where n_i is i's number of peers and N_i is his set of peers. The first part on the right-hand side is the private component of the utility function and the second is its social component. Utility is decreasing with the distance between the individual's choice and the average choice of his peers. We allow for the possibility that playing the experiment a second time affects risk choices in some way through the parameter α_2. We also include α_2^g, a dummy variable specific to the ambiguity-level fixed effect $g \in \{none, low, medium, high\}$ (recall from the last section that participants in the second experiment are randomly assigned to games with different ambiguity levels). We thus allow for each of these four games to have a different effect on the utility that results from choices. We set the reference category to $g = none$ so that $\alpha_2^{none} = 0$. W_i is the individual's payoff from the first experiment (divided by 1,000), so that δ may capture wealth effects.[16] The parameter θ is the social conformity effect, modeled as a preference to conform with peers' average behavior. A value of θ of zero would imply that individuals are not affected by their peers' choices. A negative value would mean that utility increases with the distance between the individual's choice and the average choice of his peers (implying anti-conformity preferences). A value of $\theta = 1$ would mean that the individual attributes the same weight to the private component than to the social component of the utility function. Finally, a value of θ that would tend toward infinity would mean that the individual only cares about imitating his peers. Since any value of θ is theoretically plausible, we do not constraint its value in our estimations below. We allow this parameter to differ depending on whether the participant faces ambiguity or not, so that we have :

$$\theta = \begin{cases} \theta_{na} & \text{if} \quad g = none, \\ \theta_a & \text{otherwise.} \end{cases} \quad (1.4)$$

16. The results we will present are robust to using the logarithm of the payoff instead, or to not controlling for the payoff.

Therefore, θ_{na} is the social conformity effect of participants who participate in the exact same experiment the second time, whereas θ_a is the social conformity effect for those who participate in one of the games that includes ambiguity.[17] Note that the time-invariant effects α_1, $\mathbf{x}_i\boldsymbol{\beta}$ and η_i enter the private component of the utility function in the same manner that in the first experiment. This implies we can use choices from the first experiment to control for these time-invariant effects. We do this by substituting equation (1.2) into equation (1.3), which yields :

$$U_{i2}(y_{i2}) = -\frac{1}{2}(y_{i2} - y_{i1} - \alpha_2 - \alpha_2^g - \delta W_i - \epsilon_i)^2 - \frac{\theta}{2}\left(y_{i2} - \frac{1}{n_i}\sum_{j\in N_i} y_{j1}\right)^2 \tag{1.5}$$

where $\epsilon_i \equiv \epsilon_{i2} - \epsilon_{i1}$. Note that this strategy of substituting the first experiment's first order condition in the above equation writes off $\mathbf{x}_i\boldsymbol{\beta}$ and η_i. This corresponds to using a first-difference approach in the private part of the utility function to control for the individual's fixed unobserved effect η_i. Thus, y_{i1} may capture the effect of the individual's risk preferences on choices.[18] Following this strategy, taking the first-order condition leads to the following estimable empirical model, in which the error term does not include the individual's fixed effect η_i :[19]

$$y_{i2} = \frac{1}{1+\theta}\left(\alpha_2 + \alpha_2^g + y_{i1} + \delta W_i + \frac{\theta}{n_i}\sum_{j\in N_i} y_{j1} + \epsilon_i\right). \tag{1.6}$$

Equation (1.6) provides an empirical model we can estimate. Note that the model allows to separately identify the effect of playing a second time from conformity effects. This is because the conformity effects are attributed to variations in peer choices, which are specific to each individual, while the effect of playing a second time is common to all individual and is thus captured by the constant.[20] Importantly, the model also allows us to bypass usual empirical challenges in the estimation of peer effects. First, the peer variable ($\frac{1}{n_i}\sum_{j\in N_i} y_{j1}$) is predetermined, ruling out endogeneity issues and the reflection problem described by Manski (1993), which arises when the dependent variable and the peer variable are simultaneously determined. Second, the model implicitly controls for homophily (i.e. the tendency individuals have

17. Separate peer effect estimates for all levels of ambiguity ($low, medium, high$) are available upon request. We do not find a systematic link between the magnitude of the peer effect and the ambiguity level, possibly because separate estimates are not enough precise.

18. While the second experiment introduces ambiguity, it is in large part similar to the first one given the low fraction of the balls with unknown color (see Section 1.4.2). Participants' choices should therefore still be in large part linked to the choices they made in the first experiment.

19. If the individual has no peers ($n_i = 0$), the utility function simplifies to $U_{i2}(y_{i2}) = -\frac{1}{2}(y_{i2} - y_{i1} - \alpha_2 - \alpha_2^g - \delta W_i - \epsilon_i)^2$ and the first-order condition becomes $y_{i2} = \alpha_2 + \alpha_2^g + y_{i1} + \delta W_i + \epsilon_i$. Only one individual in our sample did not report having any peers. As we will see below, we estimate the model using nonlinear least squares, which allows to estimate this individual's first-order condition jointly with those of other individuals. Furthermore, all the results we present are robust to removing this individual.

20. The constant α_2 could in principle also capture some form of peer effect : the effect of having met peers, regardless of these peers' choices. Our estimate of the social conformity peer effect θ is meant to exclude this effect, as social conformity is driven by comparison with peers' choices.

to develop relationships with people similar to themselves). Homophily is usually a concern in the estimation of peer effects. Individuals may match according to observable variables (e.g. gender, age, education), which is generally not a problem because these variables' effects can be controlled for. A more important concern is the possibility of homophily according to unobserved characteristics that might affect the variable of interest. In our model, this would mean that individuals with similar values of η_i would tend to become peers. This would imply a correlation between y_{ir} and the average outcome of i's peers even in the absence of peer effects. Fortunately, our first-difference approach in the private component of the utility function cancels out η_i in equation (1.5). Our identification strategy relies on the assumption that individuals do not choose peers based on their values on ϵ_{i1} and ϵ_{i2}. There may be homophily based on unobserved characteristics that affect risk choices in both experiments (η_i), but we assume the remaining error term ϵ_i is independent of peers' average outcome. Finally, note that the model implies a coefficient restriction because θ appears twice. It is possible to test this restriction by allowing the two parameters to differ and by testing their equality, which we do in Section 1.5.3.

Heterogeneous peer effects specification

We now allow for heterogeneous peer effects between *successful* peers and *unsuccessful* peers. We define being successful (unsuccessful) as having made the choice that led to the highest (lowest) payoff given the game and the ball that were picked at random in the first experiment. Let N_i^s be the set of peers of i who were successful in the first experiment and N_i^u be the set of peers who were unsuccessful. Additionally, let n_i^s and n_i^u be the respective numbers of i's peers in these two groups (so $n_i = n_i^s + n_i^u$). Our model with heterogeneous peer effects becomes :

$$\begin{aligned} U_{i2}(y_{i2}) = &-\frac{1}{2}(y_{i2} - y_{i1} - \alpha_2 - \alpha_2^g - \delta W_i - \epsilon_i)^2 \\ &- \frac{\theta^s n_i^s}{2n_i}\left(y_{i2} - \frac{1}{n_i^s}\sum_{j\in N_i^s} y_{j1}\right)^2 - \frac{\theta^u n_i^u}{2n_i}\left(y_{i2} - \frac{1}{n_i^u}\sum_{j\in N_i^u} y_{j1}\right)^2, \end{aligned} \tag{1.7}$$

where θ^k is the social conformity effect for the peer group $k \in \{s, u\}$, modeled as a preference to conform with this group's average behavior. The relative importance of each group is weighted by the proportion of peers in each category n_i^k/n_i. The first-order condition is :

$$y_{i2} = \frac{n_i}{n_i + \theta^s n_i^s + \theta^u n_i^u}\left(\alpha_2 + \alpha_2^g + y_{i1} + \delta W_i + \frac{\theta^s}{n_i}\sum_{j\in N_i^s} y_{j1} + \frac{\theta^u}{n_i}\sum_{j\in N_i^u} y_{j1} + \epsilon_i\right), \tag{1.8}$$

which we use as an empirical model for estimation. The marginal effect of peers' average number of safe choices in the group k (i.e. $\frac{1}{n_i^k}\sum_{j\in N_i^k} y_{j1}$) is given by $\theta^k n_i^k/(n_i + \theta^s n_i^s + \theta^u n_i^u)$. [21]

21. To see this, first add (n_i^k/n_i^k) in front of the term $\frac{1}{n_i}\sum_{j\in N_i^k} y_{j1}$ in equation 1.8.

Notice that the marginal effect of *successful* and *unsuccessful* peers' average number of safe choices (which we label as ME^s and ME^u) can be rewritten respectively as :

$$ME^s(p_i^s) = \frac{p_i^s \theta^s}{1 + \theta^s p_i^s + \theta^u (1 - p_i^s)} \qquad \text{and} \qquad ME^u(p_i^s) = \frac{(1 - p_i^s)\theta^u}{1 + \theta^s p_i^s + \theta^u (1 - p_i^s)}, \tag{1.9}$$

where p_i^s is the proportion of i's peers who where *successful* in the first experiment. Therefore, these marginal effects vary across individuals according to their proportion of peers belonging to each group. Also, the marginal effect of peers' average number of safe choices in one group decreases with the size of the peer effect of the opposite group. Equation 1.9 shows that the size and statistical significance of each of our estimates of ME^s and ME^u, which we present in the next subsection, will depend on the size and statistical significance of both peer effect estimates ($\hat{\theta}^s$ and $\hat{\theta}^u$).

Finally, note that the proportion of peers in each group could potentially be endogeneous. This would for instance be the case if participants were concerned with selecting the "right" peers in their peer group, for example to please the experimenter. We therefore test whether participants tend to systematically favour making successful peers. We compare (1) the average proportion of successful peers declared by participants who played the second experiment (i.e. $1/N \sum_{i=1}^{N} n_i^s/n_i$) to (2) the proportion of participants who were indeed successful in the first experiment. A t-test reveals that the two proportions are close (0.74 and 0.69, respectively) and not significantly different from each other. The test does not reject the null hypothesis that they are the same at a 10% significance level. Thus, we do not find convincing evidence that participants systematically choose successful peers in their network.

1.5.2 Estimation and Results

We estimate our two specifications (equations 1.6 and 1.8) using nonlinear least squares (NLS). NLS relies on the assumption that the expected value of the error term, conditional on explanatory and predetermined variables, is zero. Thus, it relies on weaker assumptions than other nonlinear methods, such as maximum likelihood estimation, that rely on distributional assumptions. [22] NLS minimizes the sum of the squares of the residuals, assuming the predicted value of y_{i2} is given by a function $g(\boldsymbol{\omega}, \mathbf{y}_{i1})$ where $\boldsymbol{\omega}$ is the vector of all parameters entering the model and $\mathbf{y}_{i1}$ is the vector of choices made by i's peers in the first experiment (i.e. containing all y_{j1} for which $j \in N_i$). NLS therefore choses the value of $\boldsymbol{\omega}$ that minimizes the objective function $\sum_{i=1}^{N}(y_i - g(\boldsymbol{\omega}, \mathbf{y}_{i1}))^2$, where

$$g(\boldsymbol{\omega}, \mathbf{y}_{i1}) = \frac{1}{1 + \theta} \left(\alpha_2 + \alpha_2^g + y_{i1} + \delta W_i + \frac{\theta}{n_i} \sum_{j \in N_i} y_{j1} \right) \tag{1.10}$$

22. See chapter 5 of Cameron and Trivedi (2005) for explanations on nonlinear estimators.

for our homogeneous specification and

$$g(\boldsymbol{\omega}, \mathbf{y}_{i1}) = \frac{n_i}{n_i + \theta^s n_i^s + \theta^u n_i^u} \left(\alpha_2 + \alpha_2^g + y_{i1} + \delta W_i + \frac{\theta^s}{n_i} \sum_{j \in N_i^s} y_{j1} + \frac{\theta^u}{n_i} \sum_{j \in N_i^u} y_{j1} \right) \quad (1.11)$$

for our heterogeneous specification. Table 1.7 presents the results for both specifications. We use the sandwich estimator of variance to calculate standard errors. Column (a) shows the estimates for the homogeneous peer effects specification. The peer effect θ_{na} (for those who participated in the same experiment the second time) is 0.783 and is significant at the 10 percent level. As discussed in Section 1.4.2, if peer effects affect more (or less) strongly participants' subjective expectations than attitudes toward pure risk, we would expect to find different estimates depending on whether or not participants played a game with ambiguity. Among those who played an experiment with ambiguity, we find a lower social conformity effect ($\hat{\theta}_a = 0.627$). This effect is more precisely estimated and significant, possibly because of the higher number of participants who played an experiment with ambiguity. A Wald test (not shown) does not reject the null hypothesis that the peer effects with and without ambiguity are the same for any reasonable level of significance. Thus, our results provide no evidence that peer effects arise more strongly by impacting subjective expectations. Column (a) of Table 1.8 presents the estimate of the marginal effect of peers' average number of safe choices, which equals $\hat{\theta}/(1+\hat{\theta})$. It equals 0.439 for individuals who played the same experiment the second time and 0.385 for those who played a experiment with ambiguity. We use the delta method to calculate their standard errors. [23] Both marginal effects are significant at a 1 percent level.

Column (b) of Table 1.7 presents the estimates of the peer effects from the heterogeneous specification. For those who participated in the same experiment (without ambiguity) the second time, we find that participants tend to conform with their peers who were successful the first time. Conversely, we find a negative but not statistically significant conformity effect from peers who were unsuccessful, and reject the null hypothesis that social conformity effects from successful and unsuccessful peers are equal. On the contrary, for participants who played a different game with ambiguity in the second experiment, we find positive social conformity effects from the two peer groups and do not reject that the two are equal. Furthermore, a Wald test (not shown) rejects that the peer effects from *unsuccessful peer* with and without ambiguity are equal with a p-value of 0.012, suggesting that peer effects may arise differently in ambiguous environment. In the presence of ambiguity, individuals may simply conform with their peers' choices regardless of the outcome.

The marginal effects from both peer groups' average number of safe choices, which are given by equation 1.9 vary across individuals since they depend on the proportion of peers belonging to each group. Column (b) of Table 1.8 present the estimates of the average marginal effects from

23. The estimate of the standard error of the marginal effect in the homogeneous specification equals $1/(1+\hat{\theta})^2 \hat{\sigma}_\theta$, where $\hat{\sigma}_\theta$ is the estimate of the standard error of $\hat{\theta}$.

the average number of safe choices made by *successful* peers (AME^s) and by *unsuccessful* peers (AME^u). The standard errors are again calculated using the delta method. [24] For individuals who played a game without ambiguity, the average marginal effect from *successful* peers' average choices is 0.539 and is significant at a 5 percent level. The negative average marginal effect from *unsuccessful* peers' average choices is less important, in part because it pushed down by the positive and important peer effect from *successful* peers. For individuals who played an experiment with ambiguity, marginal effects are both positive and statistically significant. Overall, our findings suggest a significant impact of conformism on risk-taking decisions. We also find that having won a higher payoff in the first experiment tends to make individuals more willing to take risks.

TABLE 1.7 – Peer effects on the number of safe choices - Nonlinear least squares estimation

	Hom. effects (a)	Het. effects (b)
peer effect -	0.783*	
no ambiguity θ_{na}	(0.459)	
peer effect -	0.627***	
ambiguity θ_a	(0.184)	
peer effect (successful peers) -		1.207**
no ambiguity θ^s_{na}		(0.594)
peer effect (unsuccessful peers) -		-0.935
no ambiguity θ^u_{na}		(0.734)
peer effect (successful peers) -		0.387**
ambiguity θ^s_a		(0.164)
peer effect (unsuccessful peers) -		1.261**
ambiguity θ^u_a		(0.496)
second exp. effect α_2	1.122*	1.439**
	(0.594)	(0.602)
1st exp payoff effect δ	-0.200***	-0.223***
(in thousands of UGX)	(0.069)	(0.073)
p-value H_0 :$\theta^s_{na} = \theta^u_{na}$		0.05
p-value H_0 :$\theta^s_a = \theta^u_a$		0.09
Number of observations	258	258
Ambiguity fixed effects α^g_2	Yes	Yes

*** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$

1.5.3 Additional Tests and Estimations

As mentioned in Section 1.4.3, some participants already knew each other before the workshop. Thus, even though most peers met each other at the workshop for the first time (participants

24. From the delta method, the variance of $[\hat{AME}^s \quad \hat{AME}^u]'$ equals JVJ', where V is the variance-covariance matrix of $[\hat{\theta}^s \quad \hat{\theta}^u]'$, and J is the Jacobian matrix of $[\hat{AME}^s \quad \hat{AME}^u]'$.

TABLE 1.8 – Average marginal effects from the nonlinear least squares estimations

	Hom. effects (a)	Het. effects (b)
Avg. y_{j1} - no ambiguity	0.439***	
	(0.144)	
Avg. y_{j1} - ambiguity	0.385***	
	(0.070)	
Avg. y_{j1} (successful peers) - no ambiguity		0.539**
		(0.168)
Avg. y_{j1} (unsuccessful peers) - no ambiguity		-0.209
		(1.146)
Avg. y_{j1} (successful peers) - ambiguity		0.154**
		(0.061)
Avg. y_{j1} (unsuccessful peers) - ambiguity		0.250***
		(0.058)
number of observations	258	258

Standard errors are calculated using the delta method
*** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$

have on average 4.54 peers they met at the workshop and 1.76 peers they knew before), one may be worried that our results are largely driven by these few individuals, and that the peer effects we find might only occur among these. We test for this possibility by estimating an empirical model similar to our heterogeneous specification from equation (1.8), except that "successful" and "unsuccessful" types of peers are replaced by "pre-existing" and "new" types of peers. Table 1.11 from Appendix 1.A presents separate peer effect estimates from these two types of peers. Column (a) is the homogeneous peer effects specification – exactly the same as column (a) from our main results presented in Table 1.7, whereas column (b) shows heterogeneous peer effects from "pre-existing" and "new" peers, where "pre-existing" peers refer to those the individual already knew before the workshop and "new" peers refers to those met at the workshop. The results show that the significance of our peer effect estimates is mostly driven by interactions that occurred among peers who met the first time at the workshops.

In order to explore the idea that some subsamples of participants can be more influenced than others, Table 1.12 in Appendix 1.A provides additional estimations of our homogeneous peer effect specification (equation 1.6) made on subsamples of our participants. Columns (a) and (b) estimate the model only for unsuccessful and successful participants respectively. Potentially, participants who were unsuccessful could seek more information from their peers with the objective of being successful the second time. Also, if peer effects capture competition effects beside conformity, we would expect them to differ according to the results of the first experiment. Column (c) estimates the model only for those who stated that the discussion with other had changed their mind, and column (d) estimates it only on those who made

consistent choices. Unfortunately, the smaller number of observations in these subsamples leads to imprecise estimates, especially for the effect of those who played a game without ambiguity. The estimate of the peer effects for those who played an experiment with ambiguity are still significant and in line with the estimate from our main specification. Note that, while the estimate of peer effect without ambiguity might seem high, it is not inconsistent with reasonable values. The resulting estimated marginal effect is 0.732, which is larger than the marginal effect from our main results (0.439) but not inconsistent. Overall, our results present no evidence that the peer effects we estimate are driven solely by some distinguishable subsample of our participants, or that they arise from competition effects.

Finally, as mentioned previously, the social conformity effect θ in our empirical model appears twice. It is therefore possible to test the implied coefficient restrictions. Allowing the two coefficients to differ changes our homogeneous and heterogeneous specifications respectively to

$$y_{i2} = \frac{1}{1+\theta^A}\left(\alpha_2 + \alpha_2^g + y_{i1} + \delta W_i + \frac{\theta^B}{n_i}\sum_{j\in N_i} y_{j1} + \epsilon_i\right) \tag{1.12}$$

and

$$y_{i2} = \frac{n_i}{n_i + \theta^{sA}n_i^s + \theta^{uA}n_i^u}\left(\alpha_2 + \alpha_2^g + y_{i1} + \delta W_i + \frac{\theta^{sB}}{n_i}\sum_{j\in N_i^s} y_{j1} + \frac{\theta^{uB}}{n_i}\sum_{j\in N_i^u} y_{j1} + \epsilon_i\right). \tag{1.13}$$

Table 1.13 in Appendix 1.A presents the two NLS estimations, as well as the tests of coefficient restrictions implied by our main specifications (equations 1.6 and 1.8). While some restrictions are not rejected at any reasonable level of confidence, others are rejected. In the homogeneous specification, the restriction for the parameters with ambiguity is rejected, as is the restriction on the peer effects from successful peers with ambiguity in the heterogeneous specification. However, the estimate of the coefficients themselves are mostly consistent with our main results regarding their sign and amplitude. The estimates of the effects from successful peers without ambiguity are positive and high, while those from unsuccessful peers are negative and high. The sign and amplitude of the coefficients with ambiguity are also consistent for the peer effects from unsuccessful peers. The exception is the estimates of the coefficients of the peer effects from successful peers with ambiguity : one is negative and the other positive. The negative coefficient is however very imprecise relative to its size. Overall, while these tests formally reject some of our model's coefficient restrictions, we believe that the structure we impose helps uncovering more precise estimates of the peer effects.

1.6 Social Learning and Consistency of Choices

1.6.1 The Empirical Models

We now investigate whether participants learn from their peers who made consistent choices. We assume participants make some effort to understand how to make good choices. This implies a different model underlying participants' choices than the one described in Section 1.5. Let the latent variable e^*_{ir} be the effort that an individual i puts into understanding experiment $r \in \{1, 2\}$. Participants have to reach some minimal level of understanding, normalized to 0, to make consistent choices. This leads to the standard latent variable framework :

$$e_{ir} = \begin{cases} 1 & \text{if} \quad e^*_{ir} \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{1.14}$$

where e_{ir} is the consistency of choices that results from putting enough effort into understanding the experiment. In the first experiment ($r = 1$), they choose the effort that maximizes the following utility :

$$V_{i1}(e^*_{i1}) = (c_1 + \mathbf{x}_i\boldsymbol{\gamma} + \mu_i + \psi_{i1})e^*_{i1} - \frac{{e^*_{i1}}^2}{2}, \tag{1.15}$$

where $\mathbf{x}_i$ and μ_i are the individual's fixed observed and unobserved characteristics, respectively, and ψ_{i1} is an error term. The first portion of the right-hand side represents the individual's perceived benefit from exerting effort, while the second portion represents the increasing cost of effort. The perceived benefit of effort depends on individual characteristics. For example, a low-skill person (low $\mathbf{x}_i$ or μ_i) may not see why he should try to calculate anything, and instead prefer to pick lotteries at random. [25] The first-order condition is :

$$e^*_{i1} = c_1 + \mathbf{x}_i\boldsymbol{\gamma} + \mu_i + \psi_{i1}. \tag{1.16}$$

In the second experiment - after the networking activity - participants may now have learned from their peers who made consistent choices the first time. Let m_i be the number of i's peers who made consistent choices in the first experiment. In the second experiment, individual i chooses effort e^*_{i2} in order to maximize :

$$V_{i2}(e^*_{i2}) = (c_1 + c_2 + c_2^g + \mathbf{x}_i\boldsymbol{\gamma} + \mu_i + \epsilon_{i2})e^*_{i2} - \frac{{e^*_{i2}}^2}{2} + \lambda m_i e^*_{i2}, \tag{1.17}$$

where c_2 is a constant that adds to the first experiment's constant. It might (among other things) capture a learning effect of doing the experiment a second time or a fatigue effect. We again add ambiguity dummies c_2^g specific to the level of ambiguity $g \in \{none, low, medium, high\}$ in the second experiment. The reference category is set to $g = none$ so that $c_2^{none} = 0$. The

25. Conversely, individual characteristics could be seen as affecting the cost of effort instead of its perceived benefit : a high-skill person may find it less costly to provide sufficient effort to understand the experiment.

individual's perceived utility is affected by his peers through social learning effects. The m_i peers who understood the experiment the first time may make it easier for i to understand the experiment because he can learn from them. We can see this as a reduction in the cost of effort needed to understand the experiment. As in the last section, we let the peer effect λ differ for those who participated in a treatment that included ambiguity the second time, so that :

$$\lambda = \begin{cases} \lambda_{na} & \text{if} \quad g = none, \\ \lambda_a & \text{otherwise.} \end{cases} \tag{1.18}$$

The first-order condition is :

$$e^*_{i2} = c_1 + c_2 + c^g_2 + \mathbf{x}_i\boldsymbol{\gamma} + \lambda m_i + \mu_i + \epsilon_{i2}, \tag{1.19}$$

which provides an empirical model we can estimate. Once again, the peer variable m_i is predetermined, which rules out the reflection problem of Manski (1993). It also rules out the multiple equilibriums problem that arises in binary outcome models where the dependent variable and the peer variables are simultaneously determined (Brock and Durlauf, 2001a).

Naive Specification

Contrary to Section 1.5, the latent variable framework implies we cannot use the first-difference approach to remove equation (1.19)'s time-invariant observed or unobserved variables. Thus, if there is homophily according to μ_i, m_i should be correlated with the error term. Nevertheless, as a benchmark, we first ignore homophily concerns and use equation (1.19) as our empirical model assuming $E(\mu_i + \epsilon_i | m_i, \mathbf{x}_i) = 0$.

Difference-in-Differences Specification

Homophily and peer effects may both create similarities in peers' choices in the second experiment. However, in the first experiment, only homophily may create these similarities. We can therefore use the choices in the first experiment to separately identify the two effects.

We use a specification analogous to a difference-in-differences (DID) estimation. In a standard DID setting, a control group and a treatment group are observed both before and after a treatment occurs. The variation in the outcome of interest that occurs between the two periods for reasons other than the treatment can be controlled for using the variation in this outcome among the control group. The additional variation that is specific to the treatment group is then attributed to the treatment effect.

In our setting, the number of peers who made consistent choices (m_i) is analogous to the DID treatment variable. As in a standard DID estimation, individuals with different values of m_i may on average have different levels of understanding about the experiment, even before social

interactions occur, because of homophily. The variation in the outcome that occurs between our two experiments for reasons other than social interactions can also be controlled for using a dummy variable that equals 1 if $r = 2$ and 0 otherwise. The additional variation that arises in the the second experiment as a function of m_i can then be used to identify peer effects. Specifically, we estimate the following model :

$$e^*_{ir} = c_1 + \mathbf{x}_i\boldsymbol{\gamma} + \tilde{\lambda}m_i + \mathbb{1}(r = 2)\left[c_2 + c_2^g + \lambda m_i\right] + \mu_i + \epsilon_{ir}, \tag{1.20}$$

where $\mathbb{1}(r = 2)$ equals 1 if $r = 2$ and 0 otherwise. The correlation between m_i and μ_i that comes from homophily is present in the two experiments and is thus captured by $\tilde{\lambda}$. Besides homophily effects, the estimate of $\tilde{\lambda}$ captures any relationship between μ_i and m_i that arises for reasons other than the social interactions occurring after the first experiment. Thus, λ excludes the effect of homophily and captures the peer effects, which only arise in the second experiment.

1.6.2 Estimation and Results

We estimate our two specifications using probit estimations. Table 1.9 presents the estimated average marginal effects. We include in $\mathbf{x}_i$ age, sex and education, as well as fixed effects for the six locations in which the experiments took place. Column (a) presents the naive specification (equation 1.19) and column (b) presents the DID specification (equation 1.20). The number of observations in column (b) is greater because we use the choices from the first experiment to control for homophily. The standard errors are clustered by individual, but the results are robust to using the sandwich estimator of variance without clustering. [26]

The naive peer effect estimates show a significant relationship between an individual's consistency of choices and his number of peers who made consistent choices in the first experiment. However, this relationship is significant only for participants who participated in an experiment without ambiguity the second time. The relationship may, however, include both a peer effects and a homophily effect.

Our DID estimation yields a significant homophily effect. An individual's probability of making consistent choices in the first experiment is 3.9 percentage points greater, on average, for each peer who made consistent choices, even if participants have not yet discussed with each other. The additional effect of the number of peers who made consistent choices in the second experiment — the social learning effect of having met and discussed with these peers — is not significant. Therefore, we can see that neglecting the role of homophily would have led us to interpret the relationship between one's consistency of choices and those of her peers as peer effects.

26. We avoid clustering by the six locations (on top of the locations' fixed effects), because clustering with too few clusters leads to a downward-biased variance matrix estimate, and thus to over-rejection. However, small cluster sizes may also lead to a biased estimate of the variance matrix. See Cameron and Miller (2015) for a discussion on problems that arise with few clusters or with small clusters.

TABLE 1.9 – Peer effects on consistency of choices - Average marginal effects of a probit estimation

	Naive (a)	DID (b)
peer effect - no ambiguity λ_{na}	0.113**	0.044
	(0.049)	(0.049)
peer effect - ambiguity λ_a	0.025	-0.020
	(0.023)	(0.026)
homophily effect $\tilde{\lambda}$		0.039***
		(0.014)
2nd exp. effect c_2		0.011
		(0.204)
observable characteristics		
age	-0.001	-0.004
	(0.006)	(0.004)
male	0.066	0.149***
	(0.076)	(0.048)
education : secondary	0.337***	0.173***
	(0.104)	(0.062)
education : technical	0.233**	0.170***
	(0.110)	(0.061)
education : university	0.324***	0.233***
	(0.106)	(0.062)
Number of observations	258	798
Number of individuals	258	258
Ambiguity fixed effects c_2^g	Yes	Yes
District fixed effects	Yes	Yes

*** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$

One drawback of this estimation is that our estimate of the homophily effect cannot easily be interpreted as an effect on the probability of developing a relationship. The next section explores homophily in greater details.

1.7 Testing for homophily

As mentioned previously, our models from the last two sections both control for homophily. Nevertheless, because homophily is interesting in itself, we now explore it in greater details. Homophily according to observable characteristics can be tested for by looking at whether individuals tend to be peers with others who share these observable characteristics. Furthermore, because we observe behaviors before social interactions occur, we can also test for homophily on unobservable characteristics that affect the outcome. We do so by testing for correlations in outcomes between future peers who have not yet met. This correlation cannot possibly come from peer effects and should therefore be attributable to homophily.

Let the network tie d_{ij} be equal to 1 if individual i states that individual j is his new friend and 0 otherwise. We allow the network to be directed, meaning that d_{ij} is not necessarily equal to d_{ji}. As suggested by Bramoullé and Fortin (2010), we let the probability that $d_{ij} = 1$ depend on the absolute distance between i and j's variables (which capture homophily effects) and on both i and j's variables. We model individual i's decision to state that j is one of his friends by the following rule :

$$d_{ij}^{*} = \delta_0 + \mathbf{x}_i\boldsymbol{\delta}_1 + \mathbf{x}_j\boldsymbol{\delta}_2 + y_{i1}\delta_3 + y_{j1}\delta_4 + |\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\rho}_x + |y_{i1} - y_{j1}|\rho_y + \upsilon_{ij} \tag{1.21}$$

$$d_{ij} = \begin{cases} 1 & \text{if} \quad d_{ij}^{*} > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{1.22}$$

We call $\boldsymbol{\rho}_x$ the vector of homophily according to observable characteristics effects and ρ_y the effect of homophily on unobservable characteristics (that affect y_{i1}). Importantly, the outcome variables (y_{ir} and y_{jr}) are those of the first experiment ($r = 1$) before social interactions occur, so that ρ_y may not capture peer effects. Depending on the specification, we let the outcome variable be our risk aversion measure or the consistency of choices. We also estimate a model that includes both variables.

We estimate this model using a probit estimation. Because this is a model of peer network formation, we remove observations where peers stated that they already knew each other before the workshop. It is important to note that this model has some limitations in explaining some features of the network formation. It assumes that the probability that i and j become peers is independent of other links formed in the network. Thus, this model may not explain clustering (i.e. the stylized fact that two individuals who share a peer in common have a higher probability of becoming peers with each other). One should consult Chandrasekhar (2016) for a review of econometric models that are more consistent with stylized facts. Nevertheless, this simple model allows us to test for the existence of homophily effects. Table 1.10 presents the average marginal effects for the three specifications : column (a) uses the consistency of choices as the outcome, column (b) rather uses our measure of risk aversion, and column (c) uses both. Regardless of the specification used, we find no evidence of homophily according to observable variables. We also do not find evidence of homophily according to unobserved characteristics that affect the number of safe choices, as shown in columns (b) and (c). However, consistently with our results from Section 1.6, we do find significant homophily effects according to unobserved characteristics that affect the consistency of choices, in both columns (a) and (c). Specifically, we find that the probability that individual i becomes peer with individual j is lower by 0.46 percentage points if one of them made consistent choices while the other did not. These findings suggest that participants develop relationships according to some characteristics linked to cognitive skills that are not easily observable.

TABLE 1.10 – Average marginal effects of a probit estimate - dependent variable : friendship (friends who already knew each other before the workshop are excluded)

	(a)	(b)	(c)
Absolute value of the difference between individual variables			
Consistency of choices	-0.0046 **		-0.0046 **
	(0.0020)		(0.0020)
Number of safe choices		-0.0000	-0.0001
		(0.0006)	(0.0006)
Age	-0.0003	-0.0003	-0.0003
	(0.0003)	(0.0003)	(0.0003)
Male	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)
Education	-0.0000	-0.0000	-0.0000
	(0.0006)	(0.0006)	(0.0006)
Individual's variable			
Consistency of choices	0.0021		0.0022
	(0.0020)		(0.0021)
Number of safe choices		0.0008	0.0008
		(0.0005)	(0.0005)
Age	0.0002	0.0002	0.0002
	(0.0002)	(0.0002)	(0.0002)
Male	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)
Education	-0.0000	0.0000	-0.0000
	(0.0006)	(0.0006)	(0.0006)
Potential peer's variable			
Consistency of choices	-0.0020		-0.0020
	(0.0020)		(0.0021)
Number of safe choices		-0.0005	-0.0005
		(0.0006)	(0.0006)
Age	0.0002	0.0002	0.0002
	(0.0002)	(0.0002)	(0.0002)
Male	0.0000	0.0000	0.0000
	(0.0000)	(0.0000)	(0.0000)
Education	-0.0000	-0.0000	-0.0000
	(0.0006)	(0.0006)	(0.0006)
Number of obs.	47,664	47,664	47,664

Notes :
1 - Dummy variables for the district in which the experiment took place are also included in the regression but are not shown.
2 - Standard errors are clustered by "two potential peers" identifiers.
*** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$.

1.8 Conclusion

In this paper, we combine information on the formation of a network of entrepreneurs with observations from a field experiment on choices under risk before and after social interactions occur. This design allows us to estimate social conformity effects while controlling for homophily. We find that entrepreneurs tend to conform with their peers' choices, which suggests that social interactions play a role in risk-taking decisions.

Interestingly, Herbst and Mas (2015) compare peer effects on workers' output estimated in the lab to those estimated in the field in a meta-analysis. They find that peer effects estimates in the lab generalize quantitatively. If their results also apply in the context of peer effects on risk-taking, our results imply that a policymaker could influence entrepreneurs' real life risk-related choices, such as decisions about loans or insurance, by making other entrepreneurs' choices public. He could also influence risk-taking behaviors by organizing networking activities aimed at discussing risk-taking decisions. Social conformity effects may push behaviors toward the average behavior, reducing excessive risk-taking behaviors and increasing risk tolerance for excessively risk averse individuals.

We also find that participants who make (in)consistent choices in the experiments tend to develop relationships with individuals who also make (in)consistent choices, even when controlling for observable variables such as education or gender, suggesting that peer networks are formed according to characteristics linked to cognitive ability, but n easily observable. This has implications for researchers seeking to estimate peer effects when the network is potentially endogeneous : if the outcome of interest relates to cognitive skills (e.g. educational achievement), estimated peer effects on this outcome may capture homophily also, as does our naive specification of peer effects on the consistency of choices.

The social interactions and homophily behaviors captured in our experiment are authentic ; we do not influence the network formation or the discussions participants have. Furthermore, the peer effects we estimate result from a three to four hour-long networking activity. Our finding that these few hours of free discussion time are enough to influence one's choices, at least in the short run, complement other findings in the literature that suggest that long-lasting social relationships play a role in shaping individuals' risk attitudes in the long run (e.g. Dohmen et al., 2012). Our results also raise the issue of the direction of the causal relationship between risk preferences and the decision to start a business. If individuals who start a business enter a social world of entrepreneurs who tend to have higher risk tolerances, entry into entrepreneurship might cause more risk-taking. Cramer et al. (2002) raise the possibility of reverse causality, finding a negative effect of risk aversion on entry into entrepreneurship but questioning the causality of the relationship. Brachert and Hyll (2014) find that entry into entrepreneurship is associated with an increased willingness to take risks and argue that this entry may cause a change in risk attitudes for several reasons ; our evidence suggests that

social interactions with other entrepreneurs could be one of these reasons.

A limitation of our study is that the results shed no light on how our estimated peer effects may perpetuate in the long run. While it is conceivable that the effect of a one-time-only social interaction may disappear in the long run, real life social interactions are often repeated daily, so the repeating peer effects may possibly shape long run risk-taking decisions. This is however far from obvious : the effect of social interactions could move individuals' choices away from their preferences only temporarily and may dissipate in the long run, as is the case with other behavioral effects (see Erev and Haruvy, 2013). We leave the question of whether or not repeated social interactions generate peer effects that perpetuate in the long run for future research.

1.9 Bibliography for Chapter 1

Ahern, K. R., R. Duchin, and T. Shumway (2014). Peer effects in risk aversion and trust. *Review of Financial Studies 27*(11), 3213–3240.

Ahn, T. (2010). Attitudes toward risk and self-employment of young workers. *Labour Economics 17*(2), 434–442.

Andersen, S., G. W. Harrison, M. I. Lau, and E. E. Rutström (2006). Elicitation using multiple price list formats. *Experimental Economics 9*(4), 383–405.

Anderson, L. R. and J. M. Mellor (2009). Are risk preferences stable ? comparing an experimental measure with a validated survey-based measure. *Journal of Risk and Uncertainty 39*(2), 137–160.

Arduini, T., E. Patacchini, E. Rainone, et al. (2015). Parametric and semiparametric iv estimation of network models with selectivity. Technical Report 9, Einaudi Institute for Economics and Finance (EIEF).

Attanasio, O., A. Barr, J. C. Cardenas, G. Genicot, and C. Meghir (2012). Risk pooling, risk preferences, and social networks. *American Economic Journal : Applied Economics 4*(2), 134–167.

Barseghyan, L., J. Prince, and J. C. Teitelbaum (2011). Are risk preferences stable across contexts ? evidence from insurance data. *The American Economic Review 101*(2), 591–631.

Baucells, M. and A. Villasís (2010). Stability of risk preferences and the reflection effect of prospect theory. *Theory and Decision 68*(1-2), 193–211.

Bellemare, C. and B. Shearer (2010). Sorting, incentives and risk preferences : Evidence from a field experiment. *Economics Letters 108*(3), 345–348.

Bisin, A., U. Horst, and O. Özgür (2006). Rational expectations equilibria of economies with local interactions. *Journal of Economic Theory 127*(1), 74–116.

Boucher, V. (2016). Conformism and self-selection in social networks. *Journal of Public Economics 136*, 30–44.

Brachert, M. and W. Hyll (2014). On the stability of preferences : Repercussions of entrepreneurship on risk attitudes. Technical Report 667, SOEP papers on Multidisciplinary Panel Data Research.

Bramoullé, Y. and B. Fortin (2010). Social networks : econometrics. *The New Palgrave Dictionary of Economics*.

Brock, W. A. and S. N. Durlauf (2001a). Discrete choice with social interactions. *The Review of Economic Studies 68*(2), 235–260.

Brock, W. A. and S. N. Durlauf (2001b). Interactions-based models. In *Handbook of econometrics*, Volume 5, pp. 3297–3380. Elsevier.

Bursztyn, L., F. Ederer, B. Ferman, and N. Yuchtman (2014). Understanding mechanisms underlying peer effects : Evidence from a field experiment on financial decisions. *Econometrica 82*(4), 1273–1301.

Cahlíková, J. and L. Cingl (2017). Risk preferences under acute stress. *Experimental Economics 20*(1), 209–236.

Caliendo, M., F. Fossen, and A. Kritikos (2010). The impact of risk attitudes on entrepreneurial survival. *Journal of Economic Behavior & Organization 76*(1), 45–63.

Cameron, A. C. and D. L. Miller (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources 50*(2), 317–372.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics : methods and applications*. Cambridge University Press.

Chandrasekhar, A. (2016). Econometrics of network formation. In Y. Bramoullé, A. Galeotti, and B. Rogers (Eds.), *The Oxford Handbook of the Economics of Networks*, Chapter 13, pp. 303–357.

Cohn, A., J. Engelmann, E. Fehr, and M. A. Maréchal (2015). Evidence for countercyclical risk aversion : an experiment with financial professionals. *The American Economic Review 105*(2), 860–885.

Conte, A., M. V. Levati, and C. Nardi (forthcoming). Risk preferences and the role of emotions. *Economica*.

Cramer, J. S., J. Hartog, N. Jonker, and C. M. Van Praag (2002). Low risk aversion encourages the choice for entrepreneurship : An empirical test of a truism. *Journal of Economic Behavior & Organization 48*(1), 29–36.

Croson, R. and U. Gneezy (2009). Gender differences in preferences. *Journal of Economic literature 47*(2), 448–474.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2012). The intergenerational transmission of risk and trust attitudes. *The Review of Economic Studies 79*(2), 645–677.

Dunn, T. and D. Holtz-Eakin (2000). Financial capital, human capital, and the transition to self-employment : Evidence from intergenerational links. *Journal of Labor Economics 18*(2), 282–305.

Ekelund, J., E. Johansson, M.-R. Järvelin, and D. Lichtermann (2005). Self-employment and risk aversion—evidence from psychological test data. *Labour Economics 12*(5), 649–659.

Erev, I. and E. Haruvy (2013). Learning and the economics of small decisions. In J. H. Kagel and A. E. Roth (Eds.), *The Handbook of Experimental Economics*, Volume 2, pp. 1–123.

Giannetti, M. and A. Simonov (2009). Social interactions and entrepreneurial activity. *Journal of Economics & Management Strategy 18*(3), 665–709.

Gioia, F. (2016). Peer effects on risk behaviour : the importance of group identity. *Experimental Economics 20*(1), 100–129.

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*(3), 253–264.

Herbst, D. and A. Mas (2015). Peer effects on worker output in the laboratory generalize to the field. *Science 350*(6260), 545–549.

Holt, C. A. and S. K. Laury (2002). Risk aversion and incentive effects. *American Economic Review 92*(5), 1644–1655.

Hsieh, C.-S. and L. F. Lee (2016). A social interactions model with endogenous friendship formation and selectivity. *Journal of Applied Econometrics 31*(2), 301–319.

Jacobson, S. and R. Petrie (2009). Learning from mistakes : What do inconsistent choices over risk tell us ? *Journal of Risk and Uncertainty 38*(2), 143–158.

Klibanoff, P., M. Marinacci, and S. Mukerji (2005). A smooth model of decision making under ambiguity. *Econometrica 73*(6), 1849–1892.

Lahno, A. M. and M. Serra-Garcia (2015). Peer effects in risk taking : Envy or conformity ? *Journal of Risk and Uncertainty 50*(1), 73–95.

Lindquist, M. J., J. Sol, and M. Van Praag (2015). Why do entrepreneurial parents have entrepreneurial children ? *Journal of Labor Economics 33*(2), 269–296.

Lopera, M. A. and S. Marchand (2018). Peer effects and risk-taking among entrepreneurs : Lab-in-the-field evidence. *Journal of Economic Behavior & Organization 150*, 182 – 201.

Manski, C. F. (1993). Identification of endogenous social effects : The reflection problem. *The Review of Economic Studies 60*(3), 531–542.

Nanda, R. and J. B. Sørensen (2010). Workplace peers and entrepreneurship. *Management Science 56*(7), 1116–1126.

Qu, X. and L.-f. Lee (2015). Estimating a spatial autoregressive model with an endogenous spatial weight matrix. *Journal of Econometrics 184*(2), 209–232.

Uganda Bureau of Statistics [UBOS] (2014). Uganda National Household Survey 2012/2013. Kampala, Uganda.

1.A Additional estimations

TABLE 1.11 – Peer effects on the number of safe choices - heterogeneous effects between pre-existing and new peers - Nonlinear least squares estimation

	Hom. effects (a)	Het. effects (b)
peer effect -	0.783*	
no ambiguity θ_{na}	(0.459)	
peer effect -	0.627***	
ambiguity θ_a	(0.184)	
peer effect from pre-existing peers -		-0.109
no ambiguity θ^p_{na}		(0.317)
peer effect from new peers -		2.047*
no ambiguity θ^n_{na}		(1.118)
peer effect from pre-existing peers -		0.905*
ambiguity θ^p_a		(0.497)
peer effect from new peers -		0.575***
ambiguity θ^n_a		(0.200)
second exp. effect α_2	1.122*	1.565***
	(0.594)	(0.509)
1st exp payoff effect δ	-0.200***	-0.206***
(in thousands of UGX)	(0.069)	(0.064)
p-value H_0 :$\theta^p_{na} = \theta^n_{na}$		0.08
p-value H_0 :$\theta^p_a = \theta^n_a$		0.54
Number of observations	258	258
Ambiguity fixed effects α^g_2	Yes	Yes

*** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$

TABLE 1.12 – Peer effects on the number of safe choices (estimated on subsamples) - Nonlinear least squares estimation

	(a)	(b)	(c)	(d)
peer effect -	0.639	0.776	0.120	2.725*
no ambiguity θ_{na}	(1.082)	(0.577)	(0.208)	(1.532)
peer effect -	0.552*	0.489***	0.608*	0.529***
ambiguity θ_a	(0.326)	(0.188)	(0.326)	(0.198)
second exp. effect α_2	1.867	0.318	1.539**	-0.900
	(1.494)	(0.702)	(0.699)	(0.996)
1st exp payoff effect δ	-0.375*	-0.078	-0.268***	-0.154
(in thousands of UGX)	(0.199)	(0.072)	(0.092)	(0.101)
number of observations	70	188	62	136
Ambiguity fixed effects α_2^g	Yes	Yes	Yes	Yes

*** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$

Columns (a) and (b) : estimation only for unsuccessful and successful participants respectively. **Column (c)** : estimation only for participants who stated that discussion with others had changed their mind. **Column (d)** : estimation only for participants who made consistent choices in the first experiment.

TABLE 1.13 – Test of coefficient restrictions - Nonlinear least squares estimation

	Hom. effects (a)	Het. effects (b)
peer effect - no ambiguity θ_{na}^{A}	0.172	
	(0.639)	
peer effect - no ambiguity θ_{na}^{B}	0.823**	
	(0.396)	
peer effect - ambiguity θ_{a}^{A}	0.058	
	(0.230)	
peer effect - ambiguity θ_{a}^{B}	0.793***	
	(0.180)	
peer effect (successful peers) - no ambiguity θ_{na}^{sA}		0.657
		(0.702)
peer effect (successful peers) - no ambiguity θ_{na}^{sB}		1.329**
		(0.528)
peer effect (successful peers) - ambiguity θ_{a}^{sA}		-0.263
		(0.263)
peer effect (successful peers) - ambiguity θ_{a}^{sB}		0.518***
		(0.193)
peer effect (unsuccessful peers) - no ambiguity θ_{na}^{uA}		–1.383*
		(0.775)
peer effect (unsuccessful peers) - no ambiguity θ_{na}^{uB}		–1.546*
		(0.872)
peer effect (unsuccessful peers) - ambiguity θ_{a}^{uA}		0.886*
		(0.491)
peer effect (unsuccessful peers) - ambiguity θ_{a}^{uB}		1.530***
		(0.463)
Tests of coefficient restrictions		
p-value $H_0: \theta_{na}^{A} = \theta_{na}^{B}$	0.307	
p-value $H_0: \theta_{a}^{A} = \theta_{a}^{B}$	0.004	
p-value $H_0: \theta_{na}^{sA} = \theta_{na}^{sB}$		0.217
p-value $H_0: \theta_{a}^{sA} = \theta_{a}^{sB}$		0.003
p-value $H_0: \theta_{na}^{uA} = \theta_{na}^{uB}$		0.828
p-value $H_0: \theta_{a}^{uA} = \theta_{a}^{uB}$		0.051
Number of observations	258	258

*** $p \leq 0.01$; ** $p \leq 0.05$; * $p \leq 0.1$

Note : Both estimations control for second experiment effect (α_2), first experiment payoff and ambiguity fixed effects (α_2^g).

1.B Details about the experiments

Upon arrival to the workshop, participants answered a questionnaire about their socio-demographic characteristics. They were then gathered in a room for the first experiment. An instructor explained the instructions and verified participants' comprehension by asking a series of questions. When he thought everyone understood, he took the box representing the first lottery and put it in front of the group. The box contained black balls (representing a high payoff) and white balls (representing a low payoff). He briefly explained again the composition of the box and asked participants to write down their first investment choice on a decision sheet. The online appendix provides the exact instructions provided to participants, as well as the decision sheet on which they had to write their choices. The boxes in these decision sheets indicate the exact proportion of each ball and their associated payoffs. When participants were done writing their choice, the instructor took the box representing the second lottery and briefly explained the composition of the box, before participants recorded their second choice of lottery. Then the instructor went on with the third lottery and onward. All choices were made individually and in silence. Once everyone had finished recording their choices, one of the nine lotteries was randomly chosen by drawing from a bag of balls numbered from 1 through 9. Then, a single ball was randomly drawn from the selected lottery and participants were payed according to the choice recorded on their decision sheet.

Approximately 50% of participants were then randomly chosen to participate in a second experiment. Selected participants were randomly divided into two groups, with each group participating in an experiment with a different level of ambiguity (including none, low, medium and high). Only two ambiguity treatments were conducted at each workshop. Table 1.14 shows the number of participants assigned to each ambiguity level at each workshop. Note that there are more participants assigned to the *low* and *medium* levels. This comes from a confusion that arose in the organization of one of the workshops. Specifically, the participants of the "Kampala 2" workshop should have been assigned with *none* and *high* levels of ambiguity, but were mistakenly assigned with *low* and *medium* instead. This, however, does not invalidate our results, as we control for these differences in ambiguity levels in our estimations.

Participants assigned to *none* participated in the same experiment as the first experiment. Those assigned to treatments with ambiguity were presented a box that contained, in addition to white and black balls, balls that were wrapped in opaque bags, so that their color was unknown. The decisions sheets for the low, medium and high ambiguity treatments, as well as the exact instructions that were read and provided in written form to participants, are presented in figures the online appendix.

TABLE 1.14 – Assignment of participants to the second experiment

District		1st exp. only	Ambiguity level in second experiment None	Low	Medium	High	Total
Kampala 1	Obs.	53	0	0	18	19	90
	%	59%	0%	0%	20%	21%	100%
Kampala 2	Obs.	44	0	18	15	0	77
	%	57%	0%	23%	19%	0%	100%
Wakiso	Obs.	46	0	24	21	0	91
	%	51%	0%	26%	23%	0%	100%
M'bale	Obs.	50	24	0	0	26	100
	%	50%	24%	0%	0%	26%	100%
Gulu	Obs.	50	26	27	0	0	103
	%	49%	25%	26%	0%	0%	100%
M'barara	Obs.	39	0	16	24	0	79
	%	49%	0%	20%	30%	0%	100%
Total	Obs.	282	50	85	78	45	540
	%	52%	9%	16%	14%	8%	100%

Chapitre 2

Who Benefits from Tax-Preferred Savings Accounts ?[1]

2.1 Résumé

De nombreux pays utilisent des comptes d'épargne à avantages fiscaux pour inciter les individus à épargner en vue de leur retraite. Les deux principales formes de comptes d'épargne à avantages fiscaux, les TEE et les EET, imposent l'épargne aux années de cotisation et de retrait, respectivement. Ainsi, les rendements relatifs des deux véhicules d'épargne dépendent des taux d'imposition marginaux effectifs au cours de ces deux années, lesquels dépendent à leur tour de la dynamique des revenus. Cet article estime un modèle de dynamique des revenus sur une base de données administrative longitudinale canadienne contenant des millions d'individus, permettant une hétérogénéité substantielle dans l'évolution du revenu entre les groupes de revenu. Le modèle est ensuite utilisé, avec un calculateur de d'impôt et de crédits, pour prédire comment les rendements des EET et des TEE varient entre ces groupes. Les résultats suggèrent que les rendements des TEE sont généralement plus élevés, en particulier pour les groupes à faible revenu. La comparaison des choix d'épargne optimaux prédits par le modèle avec les choix d'épargne observés dans les données suggère que les EET sont sur-choisis, en particulier dans la province de Québec. Ces résultats ont d'importantes implications pour les politiques de « nudging » qui sont actuellement mises en œuvre au Québec, obligeant les employeurs à inscrire automatiquement leurs employés à des comptes d'épargne semblables

1. I thank Vincent Boucher, Bernard Fortin, Derek Messacar and Pierre-Carl Michaud for useful comments. I am grateful to the Industrial Alliance Research Chair on the Economics of Demographic Change for financial support. Part of the analysis presented in this paper was conducted at the Quebec Interuniversity Centre for Social Statistics which is part of the Canadian Research Data Centre Network (CRDCN). The services and activities provided by the QICSS are made possible by the financial or in-kind support of the Social Sciences and Humanities Research Council (SSHRC), the Canadian Institutes of Health Research (CIHR), the Canada Foundation for Innovation (CFI), Statistics Canada, the Fonds de recherche du Québec - Société et culture (FRQSC), the Fonds de recherche du Québec - Santé (FRQS) and the Quebec universities. The views expressed in this paper are those of the author, and not necessarily those of the CRDCN or its partners.

aux EET. Ceux-ci pourraient produire des rendements très faibles pour les personnes à faible revenu, qui sont connues pour être les plus sensibles au « nudging ».

2.2 Abstract

Many countries use tax-preferred saving accounts to incentivize individuals to save for retirement. The two main forms of tax-preferred saving accounts – TEE and EET – tax savings at the contribution and withdrawal years respectively. Thus the relative returns of the two saving vehicles depend on the effective marginal tax rates in these two years, which in turn depend on earning dynamics. This paper estimates a model of earning dynamics on a Canadian longitudinal administrative database containing millions of individuals, allowing for substantial heterogeneity in the evolution of income across income groups. The model is then used, together with a tax and credit calculator, to predict how the returns of EET and TEE vary across these groups. The results suggest that TEE accounts yield in general higher returns, especially for low-income groups. Comparing optimal saving choices predicted by the model with observed saving choices in the data suggests that EET are over-chosen, especially in the province of Quebec. These results have important implications for "nudging" policies that are currently being implemented in Quebec, forcing employers to automatically enrol their employees in savings accounts similar to EET. These could yield very low returns for low-income individuals, which are known to be the most sensitive to nudging.

2.3 Introduction

Many countries use tax-preferred saving accounts to incentivize individuals to save for retirement. These accounts yield returns that depend on effective marginal tax rates (EMTRs) at the contribution and withdrawal years and may therefore benefit disproportionately to individuals with specific career paths. Given the heterogeneity of career paths, combined with the complexity of fiscal systems, it is far from obvious who benefits the most from these plans. The question of whether different types of tax-preferred saving accounts are equally well suited for low-income and high-income individuals, those with children, or those who live alone, has important policy implications for governments using or considering using these instruments to promote saving. Another important question is whether individuals are effectively able to choose the best available tax-preferred saving account for their savings. Again, the answer may vary substantially across income groups or individual characteristics if financial literacy varies across these. Answering these questions could lead to significant policy implications regarding the type of tax-preferred saving account governments should encourage, for example by "nudging" individuals in saving in a specific type of account. Still, the academic literature currently offers little insights to guide these policies.

This paper aims at filling these gaps. I estimate a model of income dynamics on a rich Cana-

dian longitudinal administrative database that comprises millions of individuals, allowing for substantial heterogeneity in income paths across income groups and other variables. I calculate the relative returns of the two main types of tax-preferred savings accounts based on predicted earnings dynamics, retirement incomes and implied EMTRs, shedding light on optimal choices of savings accounts given predicted income paths. I then explore whether individuals effectively tend to choose the "optimal" saving account – as predicted by the model.

I focus on EET and TEE savings vehicles – the two main types of tax-preferred savings accounts. EET and TEE differ mainly by the moment the savings is taxed : EET taxes it at the withdrawal year and TEE taxes it at the contribution year. Therefore comparing the returns from the two mainly involves comparing EMTRs in these two periods (this is explained in more details in the next section). EET is the most widely used plan in most OECD countries (OECD, 2015). In Canada, as in most OECD countries, employers' private pension plans are treated as EET. For individuals wishing to save by themselves, both savings vehicles are available through tax-preferred savings accounts : EET is available through the Registered Retirement Savings Plan (RRSP) and TEE through Tax-free Savings Accounts (TFSA). The possibility of choosing between RRSP and TFSA makes Canada a natural choice to study EET and TEE. Furthermore, the richness of the Longitudinal Administrative Database (LAD), which is discussed in Section 2.5, allows to study how predicted returns and observed choices differ across a large number of groups specific to income levels, province, gender or family status. Despite this, Canada has been the subject of very few studies on tax-preferred savings accounts. [2]

The predictions of the model of income dynamics, combined with a calculator of income taxes and credits that allows to calculate EMTRs, suggests that TEE types of savings vehicles tend in general to yield higher returns in the Canadian population. This finding applies to almost every subpopulation considered, but the predicted benefit of choosing a TEE over an EET is even stronger for low-income groups. The finding that EET policies seem particularly ill-suited for lower income groups is of particular importance for governments developing policies aiming at increasing savings through these incentives. For example, the government of the province of Quebec has recently implemented a policy aimed at increasing individuals' savings through these vehicles. Between 2016 and 2017, employers in Quebec with a least ten employees were progressively required to automatically enrol all employees to contribute a fraction of their wages in a Voluntary Retirement Savings Plan (VRSP), if they were not already using a comparable employer pension plan. This policy could in theory not affect savings choices if individuals ignore choices made for them by their employer and reallocate their savings

2. There are nevertheless some notable studies on Canadian tax-preferred savings accounts. Milligan (2002)'s findings suggest that EET contributions in Canada are sensitive to EMTRs, and that individuals' contributions are in part motivated by tax smoothing considerations. Milligan (2003) studies how contribution limits to tax-preferred accounts affect contribution levels and provides evidence that they affect even individuals no reaching the limit.

themselves. There is however substantial evidence that a large proportion of individuals is sensitive to nudging in their savings decisions (see Beshears et al. (2009)). Chetty and Friedman (2014) show that automatic enrolment in employers pension plans affect savings rates for most individuals in Denmark, especially among those who are the least financially sophisticated and the least prepared for retirement. Messacar (2017) arrives at similar findings in Canada, noting that the propensity of one's savings rate to be affected by nudging is inversely related to her education level. This paper's findings on the non-suitability of EET accounts for low income individuals is clearly complementary to these previous findings. If nudging policies are aimed at improving retirement prospects for individuals who are the least prepared or able to make sound decisions, these policies should not only aim at influencing these individuals' savings rate, but also at favouring good savings choices given these individuals' situation. The results of this paper suggest that making VRSP accounts of the TEE type instead of EET would favour to a greater extent individuals targeted by these policies. While the literature provides little evidence that just changing the type of account in which individuals save, and not the amount saved, would change individuals' savings rates (e.g. see Beshears et al. (2017) who use administrative data from employers introducing TEE on top of EET and find no effect on savings rate), low-income individuals' preparation for retirement could still be improved by better returns from contributing to savings account more suited to their situation.

Comparing the optimal choices predicted by the model to choices made in the LAD reveals several interesting findings. First, even though differences in income dynamics and tax codes across provinces do not result in significant differences in predicted optimal choices, there is a large difference in choices between Quebec and the other provinces considered. TEE in Quebec are chosen only around 30% of the time, but are predicted to be the best choice 70% of the times. In Ontario and British Columbia, TEE are also predicted to be the best choice around 70% of the time and are favoured over EET in more than 50% of the cases. Also, low-income individuals do seem to take their situation into account and favour TEE more often than higher-income individuals.

The next section discusses the link between EMTRs and optimal savings choices. Section 2.5 discusses the data and Section 2.6 presents the income dynamics model. Section 2.7 then uses this model to simulate income paths and calculate the implied EMTRs. Section 2.8 compares optimal contribution choices, as predicted by the model, with observed choices in the data and Section 2.10 concludes.

2.4 Effective marginal tax rates and returns from tax-preferred savings accounts

The two main types of tax-preferred savings accounts are often labelled taxable-exempt-exempt (TEE) and exempt-exempt-taxable (EET), where the three letters of the acronyms,

from left to right, represent three chronological periods : (1) the contribution period, when money is invested in a savings account, (2) the accumulation period, when savings accumulate interests and (3) the withdrawal period, when savings are withdrawn from the account. Thus, with TEE accounts, savings are taxed the year the money is invested, whereas with EET accounts it is taxed the year it is withdrawn. It is well known that the relative returns between the two depend on how the EMTRs differ in the contributory year and in the withdrawal year. Assume a two-period model where τ_0 and τ_1 are the EMTRs at the contribution and withdrawal year respectively. Under a TEE regime, the amount withdrawn in period 1 when giving up one dollar of after-tax income in period 0 is simply $R^{TEE} = (1+r)$, where r is the interest rate. Under an EET regime, one must invest $1/(1-\tau_0)$ to give up one dollar of after-tax income in period 0, so the after-tax amount withdrawn in period 1 is $R^{ETT} = \frac{(1-\tau_1)}{(1-\tau_0)}(1+r)$. Thus, to compare the return of EET relative to TEE, one must simply compare the EMTR that is avoided by contributing to a EET in period 0 with the EMTR that must be paid in period 1 on the withdrawal : R^{TEE} is smaller, equal or larger than R^{EET} if τ_1 is respectively smaller, equal or larger than τ_0.

In practice, the relevant EMTRs must be calculated using the complex fiscal rules that apply to EET contributions and withdrawals. I define the EMTR of the contribution period as follows :

$$EMTR_{contrib}(earn, \mathbf{x}) = 1 + \frac{\Delta dispinc}{\Delta contrib} \mid earn, \mathbf{x}, \tag{2.1}$$

where $earn$ is earnings, $\Delta contrib$ is an increase in EET contributions (I use 100$ in the graphs below), $\Delta dispinc$ is the variation in disposable income (i.e. income after taxes, credits, transfers and contributions) that results from the increase in contribution, and $\mathbf{x}$ is the vector of all characteristics that are taken into account in the calculation of taxes and transfers. Note that $\Delta dispinc$ can potentially vary between $-\Delta contrib$ and zero. An EMTR of zero would mean that investing 100$ in an EET reduces disposable income by 100$, so the individual does not avoid any tax or transfer clawback by contributing. An EMTR of one would mean that disposable income is not reduced by the contribution, implying that the individual avoids a 100% tax or clawback rate on the amount she invests.

I define the EMTR affecting withdrawals as follows :

$$EMTR_{withdraw}(retinc, \mathbf{x}) = 1 - \frac{\Delta dispinc}{\Delta retinc} \mid retinc, \mathbf{x}, \tag{2.2}$$

where $retinc$ is private pension incomes (EET withdrawals in Canada are treated as pension income in tax returns), $\Delta retinc$ is an increase in $retinc$ (I use 100$ in the graphs below), and $\Delta dispinc$ is the resulting increase in disposable income.

Figure 2.1 presents the EMTRs for single individuals and couples without children, in the fiscal year 2015, for Canada's three most populous provinces (Ontario, Quebec and British Columbia).[3] The solid lines depict EMTRs on contributions to EET before age 65. For both

3. I calculate the EMTRs using *SimTax*, a Canadian calculator of taxes and transfers developed by myself and other members of the *Industrial Alliance Research Chair on the Economics of Demographic Change*.

single individuals and couples, and for all provinces, these EMTRs are very low for low-come individuals and tend to increase with income. [4] The dashed lines show the EMTRs on pension withdrawal incomes for individuals over 65 years old. These EMTRs are the highest for low-income individuals – generally higher than EMTRs for contributions. This is due to high clawback rates of public pension schemes targeted at low-income seniors. [5] These high EMTRs on withdrawals can tend to make TEE savings vehicles more interesting for individuals expecting low incomes when withdrawing after age 65.

The next sections of the paper use the Longitudinal Administrative Database to estimate a model of income dynamics. This model is then used to simulate heterogeneous income paths. The EMTRs on contributions and withdrawal can then be computed, shedding light on optimal contribution choices.

2.5 Data

I use data from the Longitudinal Administrative Database (LAD), a Canadian administrative longitudinal database developed by Statistics Canada using T1 family files. The first available year is 1982. A random sample of 20% of Canadian tax filers was selected in 1982. Selected households are followed each year until individuals die or emigrate from Canada, and additional households are added each year to reach 20% of tax filers in all years. The last available year is the 2013 data, so I observe individuals for at most 31 years. Variables include most lines appearing in the Canadian income tax return, so the LAD is very rich in terms of income sources. It is however less rich in terms of other socio-demographic variables. It includes gender, age, province of residence, marital status and the age of all children. The large number of observations (20% of Canadian tax filers represents more than five millions of observations per year) makes the LAD a natural choice to study income dynamics while allowing for substantial heterogeneity in income processes. This heterogeneity will allow us to investigate whether EET and TEE are more or less suited to low-income or high-income individuals, men or women, or whether their returns vary by family status.

Importantly, the LAD also contains information on contributions to RRSP (EET tax-preferred savings account) and TFSA (TEE type). [6] Figure 2.2 presents the proportion of individuals contributing to RRSP and TFSA savings accounts at 30 y/o by year and earnings quintile. The propensity to contribute to both RRSP and TFSA increase with earnings quintile. For the lowest quintile, the proportion of individuals contributing to a TFSA account is higher than it

4. The low EMTRs on contributions for low-incomes contrast with usual high EMTRs on earnings for low incomes. This difference is mainly explained by social assistance : since one cannot claim additional social assistance by contributing to a EET (social assistance calculations ignore contributions to tax-preferred saving accounts), the high EMTRs on earnings caused by social assistance does not affect EMTRs on contributions.

5. Canadians of at least 65 years old are eligible to the Guaranteed Income Supplement (GIS), which is clawed back with income at a rate of 50%, or even 75% in some income ranges.

6. Statistics Canada matched the information from TFSA contribution to the LAD even though TSFA are not recorded in the Canadian tax returns.

is for a RRSP account, while the opposite is true for the highest quintiles. Also, since TFSAs are only available since 2009, the sharp increase in the proportion of individuals contributing to it probably results in part from a period of transition for which this saving vehicle is less known.

I use the LAD to estimate (1) a model of earnings dynamics, (2) a model of private retirement income and (3) to analyze choices between TEE and EET. The data treatment used for each of these analyses is discussed separately in the subsequent sections.

2.6 Modelling income dynamics

This section presents the models used to estimate parameters that will allow, in the next section, to simulate income path. I estimate a model of earnings dynamics, as well as a model of retirement incomes. The following subsections present each of these.

2.6.1 The earnings dynamics model

I estimate a model largely inspired from the earnings dynamics structure used in Gourinchas and Parker (2002). Let $y_{i,t}$ be real earnings of individual i at year t. I assume the following :

$$log(y_{i,t}) = f(age_{i,t}) + \alpha_t + p_{i,t} + \mu_{i,t}, \tag{2.3}$$

$$p_{i,t} = \rho p_{i,t-1} + \epsilon_{i,t}, \tag{2.4}$$

where $\mu_{i,t}$ is i's transitory shock at year t, $\epsilon_{i,t}$ is his permanent shock and ρ is the persistence of the permanent shocks. The model includes year-specific fixed effects α_t, and a parametric function of age $f(age_{i,t})$. I let $f(age_{i,t})$ be the third degree polynomial function $f(age_{i,t}) = \beta_1 age_{i,t} + \beta_2 age_{i,t}^2 + \beta_3 age_{i,t}^3$, where $age_{i,t}$ is i's age minus 30.

With panel data, this model can be estimated in two parts. First, assuming $E\left[\eta_{i,t}|age_{i,t}\right] = 0$, where $\eta_{i,t} \equiv p_{i,t} + \mu_{i,t}$. I estimate the age trends parameters β_1, β_2 and β_3 using a within individual regression with years fixed effects. Second, I estimate the variance of the permanent income shocks (σ_ϵ^2) and of transitory income shocks (σ_μ^2) by a minimum distance estimation. Minimum distance estimation consists of comparing the covariance matrix of residual earnings with the theoretical covariance matrix. Let $\boldsymbol{\theta}$ be the vector of parameters to be estimated $[\rho, \sigma_\epsilon, \sigma_\mu]'$ and $\boldsymbol{\theta}_0$ be the vector of these parameters' real values. Let also $\hat{\boldsymbol{\Omega}}(\boldsymbol{\theta}_0)$ be the observed covariance matrix from the data, and $\Sigma(\boldsymbol{\theta})$ the theoretical covariance matrix implied by the above structure, with the off-diagonal and on diagonal elements respectively given by :

$$E\left[\eta_{i,t}\eta_{i,s}\right] = \rho^{|t-s|}\frac{\sigma_\epsilon^2}{1-\rho^2}, \tag{2.5}$$

$$V\left[\eta_{i,t}\right] = \frac{\sigma_\epsilon^2}{1-\rho^2} + \sigma_\mu^2. \tag{2.6}$$

The (equally weighted) minimum distance estimator is then given by :

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \, vech\left(\Sigma(\boldsymbol{\theta}) - \hat{\boldsymbol{\Omega}}(\boldsymbol{\theta}_0)\right)' vech\left(\Sigma(\boldsymbol{\theta}) - \hat{\boldsymbol{\Omega}}(\boldsymbol{\theta}_0)\right). \tag{2.7}$$

and the standard errors of the estimates are estimated using :[7]

$$V[\hat{\boldsymbol{\theta}}] = \left(\mathbf{G}'\mathbf{G}\right)^{-1}\left(\mathbf{G}'V[vech(\hat{\boldsymbol{\Omega}}(\boldsymbol{\theta}_0))]\mathbf{G}\right)\left(\mathbf{G}'\mathbf{G}\right)^{-1} \tag{2.8}$$

where $G \equiv \partial\left(vech\left[\Sigma(\boldsymbol{\theta})\right]\right)/\partial\boldsymbol{\theta}'|_{\boldsymbol{\theta}_0}$ and $V[vech(\hat{\boldsymbol{\Omega}}(\boldsymbol{\theta}_0))]$ is estimated by bootstraps.

I estimate the above two-part model using the LAD on earnings, converted in real 2010 dollars using Statistics Canada's Consumer Price Index, of individuals born between 1953 and 1958, since these individuals are observed at least from age 30 to age 55. I only use observation from ages 30 to 55 so that the age range is identical across birth cohorts. Importantly, I allow for substantial heterogeneity in earnings dynamics by estimating the model separately for each combination of gender, province, marital status at age 30 and earnings quintile at age 30. Since I use the log of earnings as the dependent variable, only observations with strictly positive earnings are used. Therefore, it is worth keeping in mind that the results presented latter in the paper apply to individuals with uninterrupted careers. I consider the three most populous Canadian provinces (Quebec, Ontario and British Columbia), and the four following marital status :

1. Single individual without children,
2. Couple without children,
3. Single individual with one or more children,
4. Couple with one or more children.

The "single individual with one or more children" category is only considered for women, because of the lower number of observations for men in this category.

The estimates of the parameters of the first part of the model – the age trend parameters – are presented in Table 2.1. Because of the large number of parameters, this section rather summarizes the predicted age trends resulting from these estimates. Note from the appendix that most of the estimates underlying the predicted trends are estimated very precisely because of the large number of observations, with the majority of the estimates having a *p-value* of less than 0.001.

Figure 2.3 shows the predicted earnings for each combination of gender, group of provinces at age 30, marital status at age 30 and earnings quintile at age 30. Note that the slopes of the predicted trends correspond increases of earnings in percentages, since the dependent variable is the log of earnings. Predicted log earnings tend to increase with age, but tends to increase at a slower rate as age increases. Furthermore, the increase tends to be more important for the

7. See Section 6.7 of Cameron and Trivedi (2005) for the proof.

lowest quintile, so a 30-years-old low-income individual should in general expect his earnings to increase significantly throughout his career.

The estimates of the parameters of the error component model are presented in Table 2.2. The most noticeable tendency is that the persistence of persistence shocks tends to decrease with income quintile at 30 y/o. The two other parameters – the variance of persistent and transitory shock – do not display any obvious trend, but differ significantly across groups.

2.6.2 The retirement income model

This section describes how private retirement incomes are predicted using the LAD database. Using a two-part estimation, I predict private retirement taxable incomes from employers' pension plans and or from other private sources, excluding income from RRSPs (individual EET accounts). [8] I use individuals observed at least from ages 45 to 70. This allows to observe both earnings in the end of career and retirement incomes, and thus to predict the latter as a function of the former. Although the analysis models retirement incomes at 70 years old only, all results are robust to using 75 years old instead ; these results are available upon request. The first part of the model estimates the probability that a 70-years-old individual receives any private retirement income using the following model :

$$b_{1i}^{*} = \gamma_0 + \gamma_1 log(\bar{y}_i) + \gamma_2 log(cqpp_i) + \gamma_3 couple_i + \upsilon_i \tag{2.9}$$

where b_{1i}^{*} is a latent variable, $\bar{y}_i$ is the average annual earnings the individual received from 45 to 55 years old, $cqpp_i$ is her Canadian or Quebec Pension Plan income and $couple_i$ is a dummy that equals one if the individual is in a couple and zero otherwise. The observable outcome is b_{1i}, which equals one if $b_{1i}^{*} \geq 0$ and zero otherwise. This model is estimated separately for each combination of provinces and gender. I assume the error term υ_i follows a logistic distribution and use a logit estimation to estimate the model.

In the second step, I estimate the amount of private retirement income – conditioning on receiving an amount strictly greater than zero – using a OLS estimation of the following model :

$$b_{2i} = \delta_0 + \delta_1 log(\bar{y}_i) + \delta_2 log(cqpp_i) + \delta_3 couple_i + \psi_i \tag{2.10}$$

where b_{2i} is the log of private retirement pension income. The model is again estimated separately for each combination of provinces and gender.

Table 2.3 presents the estimates of the models. Earnings are positively related to private pension incomes in both the first and the second step and is the most important predictor for men. For men, a one percent increase in average earnings from 45 to 55 years old is associated

8. More precisely, the variable I define as private retirement income corresponds to the line 115 of the Canadian federal tax return.

with a little more (for Quebec) or a little less (for Ontario and British Columbia) than a one percent increase in private retirement income. For women, income from Canadian or Quebec Pension Plan (CPP or QPP) is a stronger predictor of private retirement income than earnings history.

2.7 Who can potentially benefit from EET and TEE ?

In this Section, I first use the income models from the previous section to simulate heterogeneous private income paths. I then use a calculator of taxes and other government transfers to calculate incomes from public sources and EMTRs on EET contributions and withdrawals. I finally use these EMTRs to investigate whether EET or TEE is the optimal choice given the simulated income path.

I run 10000 simulations. Each one goes as follows :

1. I assume that persistent shocks at age 29 are zero.
2. I generate log earnings from 30 to 55 years old using the coefficients and variances estimated from equations (2.3) and (2.4) – for each combination of gender, province, marital status at age 30 and earnings quintile at age 30, and using the year fixed effect from the last available year (2013). I convert these earnings *ex post* in 2015 constant dollars using Statistics Canada CPI.
3. For each simulation and each group, I calculate CPP or QPP benefits according to earnings history as follows :

$$b_{CPP} = \frac{1}{26} \sum_{t=30}^{55} 0.25 \min(MPE, y_{i,t}), \tag{2.11}$$

   where MPE is the *Maximum annual Pensionable Earnings* used in 2015 (53600$). This formula corresponds to simplified rules from 2015 and assumes the parameters from these rules remain fixed in real terms. [9] It also assumes the individual chooses to start receiving CPP/QPP at the normal age (65 years old).
4. For each simulation, I then use the estimated parameters from equations (2.9) and (2.10) to predict private pension income at 70 years old as a function of the earnings history and CPP/QPP that are generated in the previous steps.
5. Using SimTax – a Canadian calculator of taxes and transfers (see Marchand et al. (2015)), I compute, for each simulation, the EMTRs on a 1000$ contribution to an EET account at each age between 30 and 55. I also compute the EMTR on a 1000$ EET withdrawals at 70 years old. [10]

9. Using the complete set of rules would not be possible, because they depend on earnings history since 18 years old.

10. I use 1000$ contributions and withdrawals instead of 1 or 100$ in order to illustrate more realistic contribution and withdrawal behaviours.

Figure 2.4 shows the average difference in EMTRs on EET withdrawals and on EET contributions across simulations – in percentage points– for each combination of age province, gender, family status at 30 years old and earnings quintile at 30 years old. A positive value means that the EMTR on EET withdrawals at 70 years old tends to be higher than the EMTR on EET contribution at the current age, and thus that TEE should tend to be favoured. For women, EMTRs tend to be most of time higher at 70 years old in all provinces, except for those with children in the highest earnings groups. The picture is slightly more complicated for men. In all provinces, men who are single and without children at 30 years old should most of the time favour TEE. For men in a couple without children, this is only true for the lowest quintile groups. Finally, for men in a couple with children at 30 years old, EET are predicted to be optimal most of the time.

To sum up these findings, Tables 2.4 and 2.5 present the proportions of simulations for which TEE should be favoured over EET as predicted by the model. Table 2.4 shows that the dominance of TEE over EET is varies little across provinces. It does vary, however, across earnings quintiles : TEE is favoured in 73% to 79% of simulations for the lowest quintile, whereas it is only favoured in 55% to 59% of simulations for the highest quintile. Table 2.5 shows how these proportions vary across family status at 30 years old. The negative relation with earnings quintile still arises conditionally on family status. Also, single individuals, with or without children, should tend to favour TEE more often than individuals in couples according to the model. Overall, these results suggest that TEE should in general be favoured, especially for the lowest income groups.

2.8 Are predicted optimal choices in line with observed choices ?

This section compares the predicted optimal choices from last section with observed choices in the LAD. Recall that, in Canada, individuals wishing to invest themselves for retirement in a tax-preferred savings account may contribute to a RRSP (of the EET type) or to a TFSA (of the TEE type), both of which are observable in the LAD. As shown in Figure 2.2 TFSAs were only introduced in 2009 in Canada, and the proportion of individuals contributing to them has kept increasing since then. I therefore only use data from 2013, the last available year in the LAD, in the analysis. Since the analysis focusses on optimal contribution choice conditional on contributing, I exclude individuals who did not contribute to a TFSA or to a RRSP in the year. Furthermore, for simplicity, I focus on individuals who either invested in a RRSP or in a TFSA – and not in both.

Figure 2.5 presents the proportion of simulations where TEE is favoured, as well as the proportion of individuals choosing TEE in the LAD. Figure 2.5a first decomposes these proportions by province. The results are worrying for Quebec : although Quebec's earnings dynamics and

tax code creates no obvious disadvantage of choosing a TEE, it is chosen only around 30% of the times in this province, and around 50 and 55% of the times in British Columbia and Ontario. Figure 2.5b suggests that predicted optimal choices are more in line with observed choices for men than they are for women. It is however important to keep in mind that the income dynamics model might perform worst in predicting income trends that are still relevant in 2013, since the data used to estimate the model go back to 1983, and women's careers have evolved substantially since then. Figure 2.5c suggests that, while single individuals may benefit more often from TEE, they do not seem to choose this account significantly more often. Finally, Figure 2.5d suggests a positive finding for low-income individuals : the lowest earnings quintile chooses TEE almost 80% of the times, a proportion very much in line with that predicted by the model.

2.9 How would risk aversion change the picture ?

The predictions of optimal choices in Section 2.7 implicitly assume risk neutrality. However, at least two sources of uncertainty may affect optimal choices between EET and TEE. First, while the present tax code is known, future tax code may change for policy reasons. That is to say that the tax rate from TEE is given while the one from EET is uncertain. Assume an individual sacrifices one dollar of disposable income in period 0 to save it for period 1. Assume also for now, in order to isolate the effect of this source of uncertainty, that future income is given. Recall from Section 2.4 that the amount withdrawn in period 1 if investing in a TEE is $R_1^{TEE} = (1 + r)$, where r is the interest rate. If investing in an EET, this amount is $\tilde{R}_1^{EET} = \frac{1-\tilde{\tau}_1}{1-\tau_0}(1 + r)$, where $\tilde{\tau}_1 = \tau_1 + \epsilon$. Assume $E(\epsilon) = 0$, so that individuals have no information suggesting that future tax rates should systematically increase or decrease. Then, uncertainty on future tax rate only adds noise to $\tilde{R}_1^{EET}$, diminishing the desirable of EET for any risk averse individual (Rothschild and Stiglitz, 1970). If TEE was already the optimal choice without uncertainty (i.e. if $\tau_1 > \tau_0$), then adding this uncertainty makes TEE an even better choice if the individual is risk averse. If EET was the optimal choice without uncertainty, then TEE could become the optimal choice depending on the individual's risk premium. Therefore, uncertainty on future tax code would favour TEE even more than the results in Section 2.7.

A second source of uncertainty that can affect the return of EET relative to TEE is risk on future income. Assume that the future tax code is known (and thus ignore risks on the tax code discussed in the previous paragraph). Let the total tax amount that the individual will have to pay in period 1 be $T(\tilde{y}_1)$, where $\tilde{y}_1$ is the uncertain future before-tax income. An individual sacrificing one dollar of disposable income in period 0 to invest in a TEE or in an

EET will respectively have the following after-tax income in period 1 :

$$\tilde{c}_1^{TEE} = \tilde{y}_1 - T(\tilde{y}_1) + 1 + r, \tag{2.12}$$

$$\tilde{c}_1^{EET} = \tilde{y}_1 - T(\tilde{y}_1) + \frac{1 - \tau_1(\tilde{y}_1)}{1 - \tau_0(y_0)}(1 + r). \tag{2.13}$$

Note that the marginal tax rate $\tau_1(\tilde{y}_1)$ is the derivative of $T(\tilde{y}_1)$ with respect to $\tilde{y}_1$. After having invested one dollar of after-tax income in TEE and EET, respectively, a one dollar shock in future before-tax income will create the following variations in future disposable income :

$$\frac{\partial c_1^{TEE}}{\partial \tilde{y}_1} = 1 - \tau_1(\tilde{y}_1), \tag{2.14}$$

$$\frac{\partial c_1^{EET}}{\partial \tilde{y}_1} = 1 - \tau_1(\tilde{y}_1) - \tau_1'(\tilde{y}_1)\frac{(1 + r)}{1 - \tau_0(y_0)}. \tag{2.15}$$

Noting that $\frac{(1+r)}{1-\tau_0(y_0)} > 0$, it follows that shocks on future before-tax income y_1 will be attenuated by a EET, relative to a TEE, if $\tau_1' > 0$ and accentuated if $\tau_1' < 0$. The intuition behind this result is straightforward : progressive taxation (i.e. $\tau_1' > 0$) leads to less variable after-tax income, whereas the opposite is true for regressive taxation (i.e. $\tau_1' < 0$). Therefore, for risk-averse individuals, uncertainty on income increases the desirability of EET compared to TEE with progressive taxation and decreases it with regressive taxation.

In practice, as shown in Figure 2.4, EMTRs are neither clearly increasing nor clearly decreasing with income, so the effect of income uncertainty on the relative desirability of EET and TEE for risk-averse individuals is ambiguous. However EMTRs after 65-years-old do show an important decrease for lower income group that results from the high clawback rate of government transfers for low-income seniors. Overall, the progressivity of the Canadian EMTRs seems unlikely to be pronounced enough to invalidate the results from Section 2.7 that favour TEE, especially for the low-income individuals, for which TEE may be even more desirable.

2.10 Discussion and policy implications

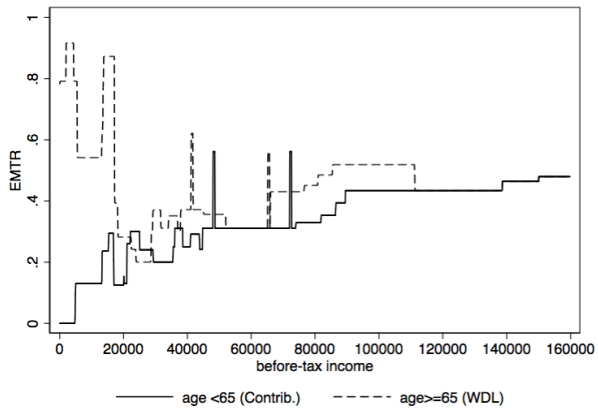
This paper suggests that, given income dynamics across income groups and the Canadian tax code, TEE savings vehicles tend to yield higher returns than EET, especially for the lowest income groups. This is in large part due to the very high EMTRs resulting from the clawbacks of social transfers. While the main analysis considered EMTRs implicitly assumed risk neutrality, it is likely that risk aversion favours TEE accounts even more. First, uncertainty on future tax code adds noise to the future return of EET accounts. Second, uncertainty on income may only favour EET if the progressivity of taxation is significant. For low-income individuals the decreasing EMTRs resulting from the clawbacks of social transfers are therefore likely to favour TEE even more, as before-tax income shocks are accentuated by EET accounts

under regressive taxation. Observed choices in the LAD suggests that low-income individuals do take their situation into account and favour TEE more often than higher-income individuals.

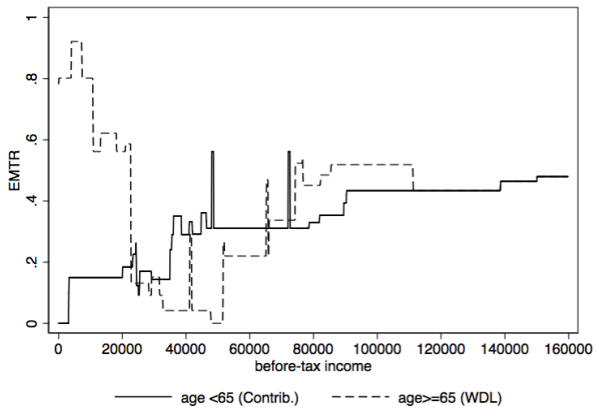
Another finding is that TEE is much less favoured in Quebec than in Ontario or British Columbia, a result than is not explained by differences in taxation or in income dynamics across provinces. These findings are important considering that the government of Quebec is currently implementing policies aimed at nudging more individuals to save in EET accounts. Employers not currently offering an equivalent pension plan are now required to automatically enrol their employees in VRSPs, which are of the EET type. It is likely that these policies will lead low-income individuals to make saving choices that are less suited to their situation. Since richer and more educated individuals are probably more able to ignore default choices made by their employer and make saving choices according to their own situation (see Chetty and Friedman (2014) and Messacar (2017)), it would seem natural that nudging policies be more oriented toward individuals with lower incomes. Thus, making VRSPs of the TEE type instead of EET would be a policy worth considering. Future research should explore in greater length potential financial literacy problems in Quebec and their implications for nudging policies.

FIGURE 2.1 – Effective Marginal Tax Rates (EMTR) for contributions before age 65 and pension withdrawals after age 65 ; fiscal year 2015 ; no child ; graphs for couples assume one individual has all before-tax income

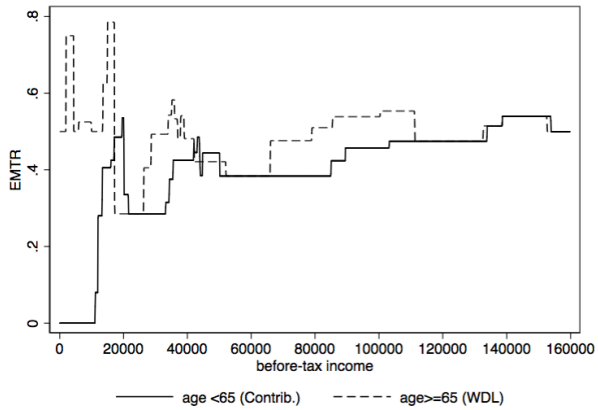
(a) Ontario - Single individuals



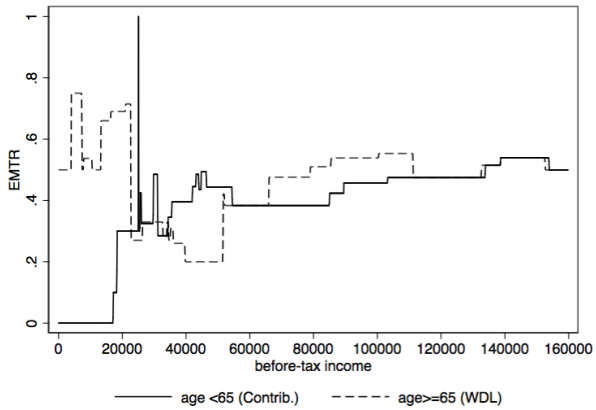
(b) Ontario - Couples



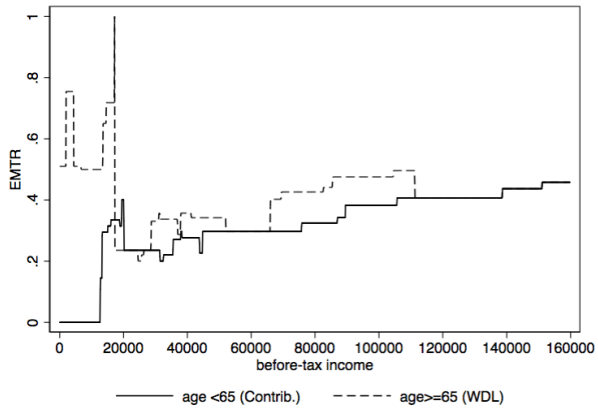
(c) Quebec - Single individuals



(d) Quebec - Couples



(e) British Columbia - Single individuals



(f) British Columbia - Couples

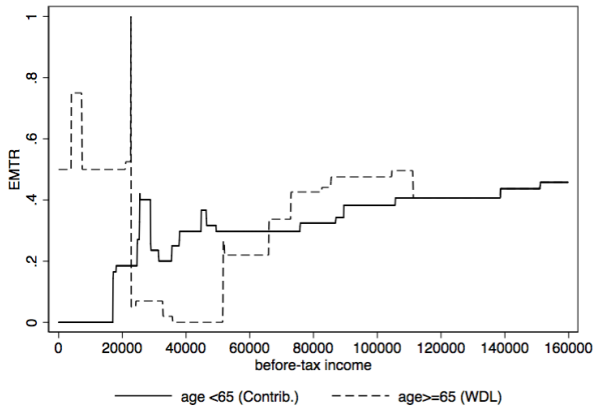


TABLE 2.1 – Estimated coefficients of earnings age trends by family status and earnings quintile at 30 y/o - within individual regressions with year fixed effects (not shown) - p-values in square brackets

Women-Quebec

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1438	0.0529	0.0161	0.0121	0.0119	0.1305	0.0385	0.0025	0.0003	0.0002
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.2651]	[0.8910]	[0.9334]
$(age-30)^2$	-0.0094	-0.0021	-0.0002	0.0011	0.0014	-0.0078	-0.0015	0.0015	0.0023	0.0027
	[0.0000]	[0.0000]	[0.3924]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^3$	0.0002	0.0000	0.0000	-0.0001	-0.0001	0.0002	0.0000	-0.0001	-0.0001	-0.0001
	[0.0000]	[0.0556]	[0.1997]	[0.0000]	[0.0000]	[0.0000]	[0.0184]	[0.0000]	[0.0000]	[0.0000]
constant	9.3530	10.0102	10.2863	10.5619	10.8895	9.1925	9.9281	10.2250	10.5217	10.7926
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	43735	58970	65215	69565	61745	34950	47745	60885	69180	62210
N. indiv.	2005	2540	2620	2735	2510	1625	2080	2510	2760	2565
	Single-with child at 30 y/o					Couple-with child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1357	0.0248	0.0131	0.0071	0.0080	0.1265	0.0196	-0.0078	-0.0162	-0.0178
	[0.0000]	[0.0000]	[0.0005]	[0.0500]	[0.0759]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^2$	-0.0057	0.0020	0.0026	0.0028	0.0045	-0.0042	0.0033	0.0047	0.0058	0.0059
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^3$	0.0001	-0.0001	-0.0001	-0.0001	-0.0002	0.0000	-0.0001	-0.0002	-0.0002	-0.0002
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
constant	8.9728	9.7236	10.1942	10.5152	10.6890	8.8088	9.6755	10.0969	10.3305	10.6444
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	35070	31335	26925	21725	14570	147745	157975	171500	166290	126105
N. indiv.	1705	1365	1120	855	585	6840	6690	6965	6555	4960

Men-Quebec

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1237	0.0398	0.0142	0.0160	-0.0027	0.1583	0.0683	0.0312	0.0227	0.0152
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.2640]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^2$	-0.0069	-0.0016	0.0002	0.0016	0.0027	-0.0105	-0.0036	-0.0006	0.0010	0.0022
	[0.0000]	[0.0000]	[0.3526]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0005]	[0.0000]	[0.0000]
$(age-30)^3$	0.0001	0.0000	-0.0000	-0.0001	-0.0001	0.0002	0.0000	-0.0000	-0.0001	-0.0001
	[0.0000]	[0.2379]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0003]	[0.0000]	[0.0000]
constant	9.5249	10.2970	10.5760	10.8340	11.0059	9.9067	10.4479	10.6562	10.8362	11.1303
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	124675	109275	97960	71175	46380	67835	75755	86110	73950	54530
N. indiv.	5705	4685	4005	2900	1950	2845	2990	3315	2905	2215
	Couple-with child at 30 y/o									
	q1	q2	q3	q4	q5					
$age-30$	0.1388	0.0565	0.0274	0.0228	0.0127					
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]					
$(age-30)^2$	-0.0088	-0.0029	-0.0006	0.0007	0.0027					
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]					
$(age-30)^3$	0.0002	0.0000	-0.0000	-0.0001	-0.0001					
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]					
constant	9.7250	10.3712	10.6248	10.8633	11.0916					
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]					
N. obs.	122320	171895	215615	206895	174670					
N. indiv.	4920	6545	7970	7795	6730					

Table 2.1 (continued) - Estimated coefficients of earnings age trends by family status and earnings quintile at 30 y/o - within individual regressions with year fixed effects (not shown) - p-values in square brackets

Women-Ontario

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1386	0.0540	0.0228	0.0188	0.0098	0.1373	0.0504	0.0031	0.0008	-0.0020
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.1473]	[0.6703]	[0.2284]
$(age-30)^2$	-0.0097	-0.0039	-0.0011	-0.0004	0.0013	-0.0107	-0.0032	0.0000	0.0006	0.0015
	[0.0000]	[0.0000]	[0.0000]	[0.1074]	[0.0000]	[0.0000]	[0.0000]	[0.9684]	[0.0092]	[0.0000]
$(age-30)^3$	0.0002	0.0001	0.0000	0.0000	-0.0001	0.0003	0.0001	0.0000	0.0000	-0.0001
	[0.0000]	[0.0000]	[0.0816]	[0.0008]	[0.0000]	[0.0000]	[0.0000]	[0.3377]	[0.0001]	[0.0000]
constant	9.4181	10.0548	10.3526	10.6421	10.8510	9.4635	10.0445	10.3006	10.5689	10.8568
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	39210	62975	74045	84270	93940	34365	61175	80595	103725	124035
N. indiv.	2195	3160	3390	3710	4280	1890	2995	3640	4495	5495

	Single-with child at 30 y/o					Couple-with child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1474	0.0361	0.0100	-0.0039	-0.0054	0.1318	0.0279	-0.0139	-0.0323	-0.0333
	[0.0000]	[0.0000]	[0.0029]	[0.3002]	[0.2624]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^2$	-0.0066	0.0004	0.0022	0.0031	0.0035	-0.0049	0.0024	0.0051	0.0067	0.0068
	[0.0000]	[0.2713]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^3$	0.0001	-0.0001	-0.0001	-0.0001	-0.0002	0.0001	-0.0001	-0.0002	-0.0002	-0.0002
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
constant	9.0230	9.8024	10.1694	10.3718	10.7707	8.8455	9.7238	10.0594	10.2354	10.5411
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	43715	36000	31570	26205	17335	195315	202830	234905	235245	230180
N. indiv.	2620	1845	1475	1165	795	9780	9180	10175	10000	9670

Men-Ontario

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1240	0.0461	0.0209	0.0163	0.0032	0.1766	0.0786	0.0410	0.0297	0.0133
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0590]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^2$	-0.0080	-0.0022	-0.0006	0.0010	0.0016	-0.0115	-0.0044	-0.0014	-0.0004	0.0015
	[0.0000]	[0.0000]	[0.0008]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0069]	[0.0000]
$(age-30)^3$	0.0002	0.0000	-0.0000	-0.0001	-0.0001	0.0002	0.0001	0.0000	-0.0000	-0.0001
	[0.0000]	[0.0000]	[0.0946]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.7618]	[0.0000]	[0.0000]
constant	9.6998	10.3793	10.6569	10.8628	11.0362	10.0309	10.5822	10.7309	10.9433	11.0884
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	111900	112320	109665	103695	89370	72660	87285	104640	119325	114510
N. indiv.	6490	5695	5135	4780	4140	3570	3945	4450	5060	5010

	Couple-with child at 30 y/o				
	q1	q2	q3	q4	q5
$age-30$	0.1495	0.0653	0.0391	0.0294	0.0128
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^2$	-0.0103	-0.0039	-0.0012	-0.0001	0.0018
	[0.0000]	[0.0000]	[0.0000]	[0.3039]	[0.0000]
$(age-30)^3$	0.0002	0.0001	-0.0000	-0.0000	-0.0001
	[0.0000]	[0.0000]	[0.0001]	[0.0000]	[0.0000]
constant	9.8846	10.4694	10.7368	10.9294	11.1028
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	102255	159135	229695	305435	326810
N. indiv.	5070	6930	9400	12480	13640

Table 2.1 (continued) - Estimated coefficients of earnings age trends by family status and earnings quintile at 30 y/o - within individual regressions with year fixed effects (not shown) - p-values in square brackets

Women-British-Colombia

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1212	0.0601	0.0181	0.0042	-0.0121	0.1185	0.0428	-0.0035	-0.0143	-0.0298
	[0.0000]	[0.0000]	[0.0000]	[0.2053]	[0.0001]	[0.0000]	[0.0000]	[0.4024]	[0.0000]	[0.0000]
$(age-30)^2$	-0.0084	-0.0045	-0.0010	0.0012	0.0023	-0.0080	-0.0017	0.0008	0.0016	0.0035
	[0.0000]	[0.0000]	[0.0349]	[0.0038]	[0.0000]	[0.0000]	[0.0036]	[0.0930]	[0.0001]	[0.0000]
$(age-30)^3$	0.0002	0.0001	0.0000	-0.0001	-0.0001	0.0002	0.0000	-0.0000	-0.0000	-0.0001
	[0.0000]	[0.0000]	[0.7039]	[0.0000]	[0.0000]	[0.0000]	[0.1177]	[0.4158]	[0.0001]	[0.0000]
constant	9.2553	10.0556	10.4042	10.5601	10.7418	9.3190	9.9085	10.2606	10.5242	10.6776
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	15455	21695	22110	30170	38210	13020	20875	23240	31285	44715
N. indiv.	935	1255	1205	1490	1880	805	1150	1205	1540	2185
	Single-with child at 30 y/o					Couple-with child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1381	0.0310	0.0096	-0.0104	-0.0110	0.1345	0.0222	-0.0146	-0.0490	-0.0673
	[0.0000]	[0.0000]	[0.1697]	[0.1603]	[0.1730]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^2$	-0.0073	-0.0007	0.0029	0.0039	0.0038	-0.0056	0.0039	0.0062	0.0089	0.0097
	[0.0000]	[0.3174]	[0.0004]	[0.0000]	[0.0001]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
$(age-30)^3$	0.0001	-0.0000	-0.0001	-0.0001	-0.0002	0.0001	-0.0002	-0.0002	-0.0003	-0.0003
	[0.0000]	[0.8401]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
constant	8.9394	9.8772	10.2215	10.3016	10.7052	8.9174	9.6508	10.0082	10.1446	10.3265
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	16510	12995	8745	6915	6135	66160	57435	51210	56165	69660
N. indiv.	1080	755	480	365	305	3685	3025	2585	2710	3245

Men-British Colombia

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$age-30$	0.1207	0.0484	0.0158	0.0111	-0.0100	0.1604	0.0779	0.0341	0.0205	-0.0035
	[0.0000]	[0.0000]	[0.0000]	[0.0002]	[0.0001]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.1448]
$(age-30)^2$	-0.0068	-0.0021	0.0004	0.0018	0.0021	-0.0092	-0.0045	-0.0018	-0.0000	0.0024
	[0.0000]	[0.0000]	[0.2625]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.9705]	[0.0000]
$(age-30)^3$	0.0001	0.0000	-0.0000	-0.0001	-0.0001	0.0002	0.0001	0.0000	-0.0000	-0.0001
	[0.0000]	[0.1272]	[0.0003]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0069]	[0.0020]	[0.0000]
constant	9.6479	10.3565	10.6518	10.8806	11.0340	9.8387	10.5723	10.7762	10.9338	11.0590
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]
N. obs.	40860	38460	32480	34615	42340	25490	27350	26930	32580	41995
N. indiv.	2650	2265	1830	1805	2120	1430	1430	1330	1575	2005
	Couple-with child at 30 y/o									
	q1	q2	q3	q4	q5					
$age-30$	0.1507	0.0677	0.0315	0.0203	0.0016					
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.2390]					
$(age-30)^2$	-0.0093	-0.0039	-0.0013	-0.0004	0.0014					
	[0.0000]	[0.0000]	[0.0000]	[0.0874]	[0.0000]					
$(age-30)^3$	0.0002	0.0001	0.0000	-0.0000	-0.0001					
	[0.0000]	[0.0000]	[0.6069]	[0.0001]	[0.0000]					
constant	9.8880	10.5110	10.7324	10.9515	11.1354					
	[0.0000]	[0.0000]	[0.0000]	[0.0000]	[0.0000]					
N. obs.	35935	46855	52610	69935	116405					
N. indiv.	1960	2350	2575	3345	5425					

TABLE 2.2 – Estimated persistence (ρ) variance of persistent shocks (σ^2_ϵ) and of transitory shocks (σ^2_μ) by family status and earnings quintile at 30 y/o of residuals from within individual regressions with year fixed effects - p-values in square brackets

Women-Quebec										
	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.9586	0.9486	0.9381	0.9087	0.9115	0.9723	0.9538	0.9397	0.8885	0.8988
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.0338	0.0347	0.0326	0.0430	0.0481	0.0218	0.0320	0.0317	0.0512	0.0558
	[0.022]	[0.030]	[0.002]	[0.329]	[0.405]	[0.370]	[0.139]	[0.006]	[0.609]	[0.307]
$\hat{\sigma}^2_\mu$	0.1630	0.1282	0.1097	0.1647	0.2072	0.1922	0.1008	0.1361	0.1392	0.2051
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
N. obs.	43735	58970	65215	69565	61745	34950	47745	60885	69180	62210
N. indiv.	2005	2540	2620	2735	2510	1625	2080	2510	2760	2565
	Single-with child at 30 y/o					Couple-with child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.9460	0.9443	0.8945	0.9305	0.8762	0.9485	0.9412	0.9330	0.9270	0.9176
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.0439	0.0385	0.0422	0.0213	0.0487	0.0437	0.0379	0.0397	0.0261	0.0443
	[0.411]	[0.397]	[0.582]	[0.766]	[0.838]	[0.004]	[0.008]	[0.016]	[0.066]	[0.168]
$\hat{\sigma}^2_\mu$	0.1760	0.1567	0.1581	0.1454	0.1994	0.2106	0.1647	0.1546	0.1899	0.2097
	[0.000]	[0.000]	[0.000]	[0.001]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
Men-Quebec										
	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.955	0.950	0.932	0.930	0.921	0.953	0.951	0.950	0.910	0.917
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.036	0.030	0.032	0.036	0.042	0.032	0.028	0.023	0.051	0.055
	[0.007]	[0.052]	[0.007]	[0.381]	[0.682]	[0.062]	[0.020]	[0.062]	[0.633]	[0.389]
$\hat{\sigma}^2_\mu$	0.195	0.129	0.145	0.173	0.206	0.196	0.117	0.113	0.134	0.233
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
N. obs.	124675	109275	97960	71175	46380	67835	75755	86110	73950	54530
N. indiv.	5705	4685	4005	2900	1950	2845	2990	3315	2905	2215
	Couple-with child at 30 y/o									
	q1	q2	q3	q4	q5					
$\hat{\rho}$	0.958	0.951	0.941	0.923	0.926					
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]					
$\hat{\sigma}^2_\epsilon$	0.025	0.027	0.023	0.041	0.041					
	[0.003]	[0.000]	[0.000]	[0.122]	[0.000]					
$\hat{\sigma}^2_\mu$	0.165	0.125	0.097	0.125	0.201					
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]					
N. obs.	122320	171895	215615	206895	174670					
N. indiv.	4920	6545	7970	7795	6730					

Table 2.2 (continued) - Estimated persistence (ρ) variance of persistent shocks (σ^2_ϵ) and of transitory shocks (σ^2_μ) by family status and earnings quintile at 30 y/o of residuals from within individual regressions with year fixed effects - p-values in square brackets

Women-Ontario										
	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.977	0.959	0.948	0.901	0.918	0.953	0.963	0.935	0.911	0.938
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.022	0.035	0.030	0.061	0.064	0.043	0.039	0.041	0.061	0.040
	[0.655]	[0.401]	[0.150]	[0.234]	[0.112]	[0.244]	[0.093]	[0.034]	[0.303]	[0.230]
$\hat{\sigma}^2_\mu$	0.165	0.135	0.098	0.096	0.177	0.119	0.116	0.084	0.088	0.127
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.002]	[0.006]	[0.000]
N. obs.	39210	62975	74045	84270	93940	34365	61175	80595	103725	124035
N. indiv.	2195	3160	3390	3710	4280	1890	2995	3640	4495	5495
	Single-with child at 30 y/o					Couple-with child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.920	0.917	0.921	0.872	0.841	0.937	0.937	0.928	0.927	0.920
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.060	0.041	0.035	0.071	0.099	0.048	0.047	0.044	0.043	0.044
	[0.501]	[0.846]	[0.696]	[0.829]	[0.672]	[0.003]	[0.141]	[0.022]	[0.000]	[0.002]
$\hat{\sigma}^2_\mu$	0.279	0.215	0.129	0.130	0.058	0.205	0.132	0.107	0.121	0.155
	[0.000]	[0.000]	[0.006]	[0.175]	[0.871]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
N. obs.	43715	36000	31570	26205	17335	195315	202830	234905	235245	230180
N. indiv.	2620	1845	1475	1165	795	9780	9180	10175	10000	9670
Men-Ontario										
	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.946	0.958	0.942	0.884	0.921	0.959	0.959	0.939	0.941	0.922
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.045	0.034	0.028	0.068	0.055	0.043	0.034	0.039	0.033	0.057
	[0.010]	[0.005]	[0.095]	[0.043]	[0.217]	[0.094]	[0.063]	[0.331]	[0.010]	[0.195]
$\hat{\sigma}^2_\mu$	0.179	0.124	0.103	0.123	0.129	0.151	0.093	0.092	0.100	0.175
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.010]	[0.000]	[0.000]	[0.000]
N. obs.	111900	112320	109665	103695	89370	72660	87285	104640	119325	114510
N. indiv.	6490	5695	5135	4780	4140	3570	3945	4450	5060	5010
	Couple-with child at 30 y/o									
	q1	q2	q3	q4	q5					
$\hat{\rho}$	0.950	0.940	0.929	0.911	0.908					
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]					
$\hat{\sigma}^2_\epsilon$	0.038	0.039	0.036	0.053	0.062					
	[0.001]	[0.055]	[0.000]	[0.000]	[0.000]					
$\hat{\sigma}^2_\mu$	0.196	0.137	0.123	0.109	0.164					
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]					
N. obs.	102255	159135	229695	305435	326810					
N. indiv.	5070	6930	9400	12480	13640					

Table 2.2 (continued) - Estimated persistence (ρ) variance of persistent shocks (σ^2_ϵ) and of transitory shocks (σ^2_μ) by family status and earnings quintile at 30 y/o of residuals from within individual regressions with year fixed effects - p-values in square brackets

Women-British-Columbia

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.923	0.967	0.935	0.933	0.888	0.959	0.952	0.892	0.905	0.932
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.060	0.020	0.037	0.029	0.081	0.043	0.044	0.054	0.087	0.038
	[0.827]	[0.846]	[0.781]	[0.222]	[0.441]	[0.869]	[0.554]	[0.621]	[0.567]	[0.208]
$\hat{\sigma}^2_\mu$	0.061	0.107	0.132	0.153	0.115	0.176	0.178	0.109	0.099	0.131
	[0.872]	[0.086]	[0.005]	[0.000]	[0.015]	[0.000]	[0.000]	[0.064]	[0.298]	[0.012]
N. obs.	15455	21695	22110	30170	38210	13020	20875	23240	31285	44715
N. indiv.	935	1255	1205	1490	1880	805	1150	1205	1540	2185
	Single-with child at 30 y/o					Couple-with child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.813	0.984	0.930	0.915	0.923	0.931	0.926	0.932	0.902	0.945
	[0.000]	[0.064]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.164	0.003	0.057	0.029	0.019	0.056	0.052	0.051	0.065	0.030
	[0.801]	[0.998]	[0.763]	[0.905]	[0.775]	[0.288]	[0.446]	[0.357]	[0.268]	[0.136]
$\hat{\sigma}^2_\mu$	0.000	0.134	0.105	0.154	0.148	0.213	0.139	0.141	0.112	0.132
	[1.000]	[0.378]	[0.813]	[0.042]	[0.466]	[0.000]	[0.000]	[0.000]	[0.003]	[0.000]
N. obs.	16510	12995	8745	6915	6135	66160	57435	51210	56165	69660
N. indiv.	1080	755	480	365	305	3685	3025	2585	2710	3245

Men-British Colombia

	Single-no child at 30 y/o					Couple-no child at 30 y/o				
	q1	q2	q3	q4	q5	q1	q2	q3	q4	q5
$\hat{\rho}$	0.961	0.921	0.926	0.934	0.918	0.969	0.939	0.939	0.934	0.918
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]
$\hat{\sigma}^2_\epsilon$	0.034	0.036	0.034	0.043	0.039	0.020	0.042	0.020	0.029	0.042
	[0.525]	[0.777]	[0.772]	[0.406]	[0.142]	[0.734]	[0.671]	[0.677]	[0.651]	[0.702]
$\hat{\sigma}^2_\mu$	0.174	0.185	0.098	0.112	0.168	0.189	0.117	0.079	0.118	0.120
	[0.000]	[0.000]	[0.115]	[0.003]	[0.000]	[0.002]	[0.000]	[0.029]	[0.042]	[0.005]
N. obs.	40860	38460	32480	34615	42340	25490	27350	26930	32580	41995
N. indiv.	2650	2265	1830	1805	2120	1430	1430	1330	1575	2005
	Couple-with child at 30 y/o									
	q1	q2	q3	q4	q5					
$\hat{\rho}$	0.947	0.959	0.929	0.919	0.922					
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]					
$\hat{\sigma}^2_\epsilon$	0.035	0.022	0.037	0.033	0.031					
	[0.593]	[0.166]	[0.549]	[0.517]	[0.023]					
$\hat{\sigma}^2_\mu$	0.199	0.146	0.112	0.112	0.128					
	[0.000]	[0.000]	[0.000]	[0.000]	[0.000]					
N. obs.	35935	46855	52610	69935	116405					
N. indiv.	1960	2350	2575	3345	5425					

TABLE 2.3 – Private retirement income models - Standard errors in parentheses

	Women					
	First step : Logit estimation			**Second step : OLS estimation**		
	Quebec	Ontario	BC	Quebec	Ontario	BC
$log(avg.earnings)$	0.467	0.389	0.332	0.388	0.355	0.307
	(0.015)	(0.013)	(0.022)	(0.010)	(0.008)	(0.013)
$log(CQPP)$	1.100	1.030	1.086	0.738	0.544	0.579
	(0.031)	(0.025)	(0.042)	(0.021)	(0.017)	(0.027)
couple	-0.116	-0.096	0.030	-0.172	-0.278	-0.269
	(0.031)	(0.025)	(0.042)	(0.019)	(0.015)	(0.024)
constant	-13.565	-12.239	-12.232	-1.404	0.739	0.868
	(0.243)	(0.198)	(0.335)	(0.163)	(0.132)	(0.213)
std. dev. of residuals				1.070	1.079	1.017
N. obs.	25245	40125	13860	14455	25180	8695
	Men					
	First step : Logit estimation			**Second step : OLS estimation**		
	Quebec	Ontario	BC	Quebec	Ontario	BC
$log(avg.earnings)$	1.096	0.977	0.895	1.072	0.959	0.860
	(0.022)	(0.018)	(0.030)	(0.012)	(0.009)	(0.015)
$log(CQPP)$	-0.012	0.255	0.289	-0.060	-0.277	-0.232
	(0.040)	(0.036)	(0.061)	(0.025)	(0.020)	(0.034)
couple	0.048	0.189	0.240	-0.007	-0.073	-0.011
	(0.033)	(0.031)	(0.048)	(0.020)	(0.016)	(0.026)
constant	-10.764	-11.985	-11.525	-1.829	1.605	2.143
	(0.313)	(0.304)	(0.490)	(0.224)	(0.196)	(0.306)
std. dev. of residuals				1.066	0.973	0.956
N. obs.	29820	42410	15260	20240	30905	10850

TABLE 2.4 – Proportion of simulations favouring TEE over EET by earnings quintile at age 30 and province

Earnings quintile	**Quebec**	**Ontario**	**BC**
1	0.73	0.79	0.77
2	0.73	0.75	0.71
3	0.65	0.70	0.66
4	0.60	0.61	0.59
5	0.59	0.55	0.57

TABLE 2.5 – Proportion of simulations favouring TEE over EET by earnings quintile at age 30 and family status at age 30

Earnings quintile	**Single-no child**	**Couple-no child**	**Single with children**	**Couple with children**
1	0.88	0.74	0.92	0.59
2	0.88	0.67	0.81	0.59
3	0.86	0.67	0.52	0.56
4	0.81	0.60	0.47	0.47
5	0.77	0.50	0.53	0.44

FIGURE 2.2 – Proportion of individuals contributing to RRSP and TFSA savings accounts at 30 y/o by year and earnings quintile

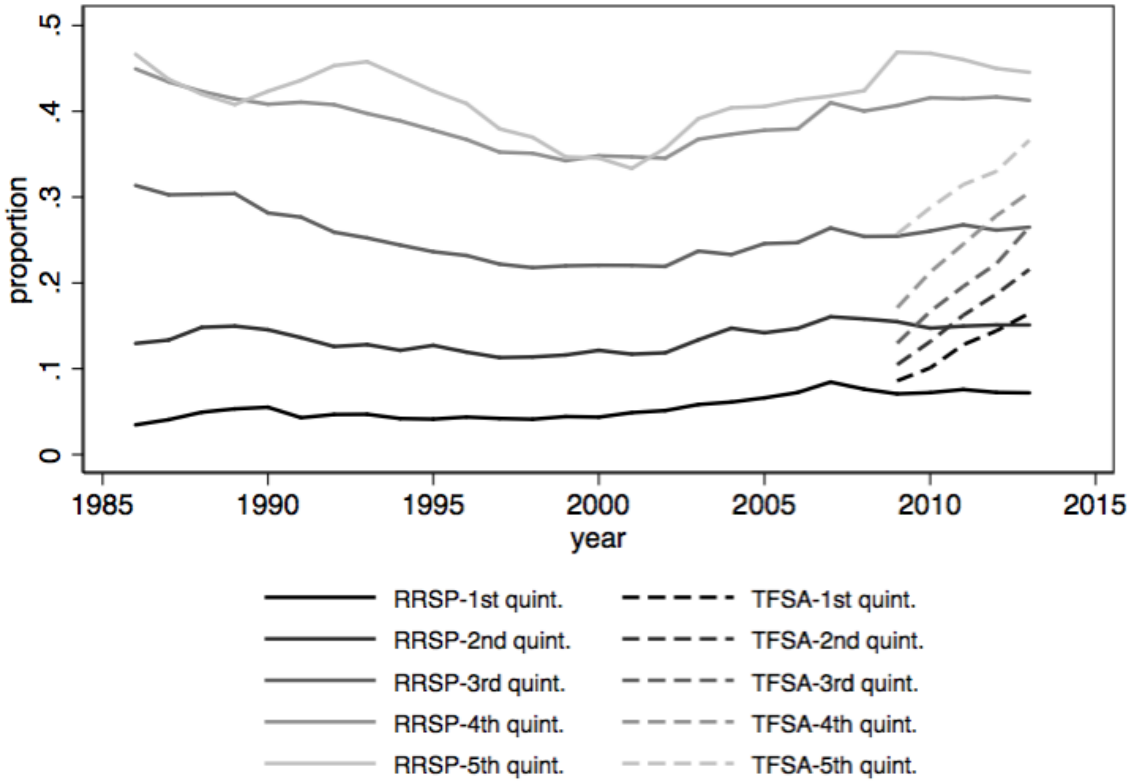
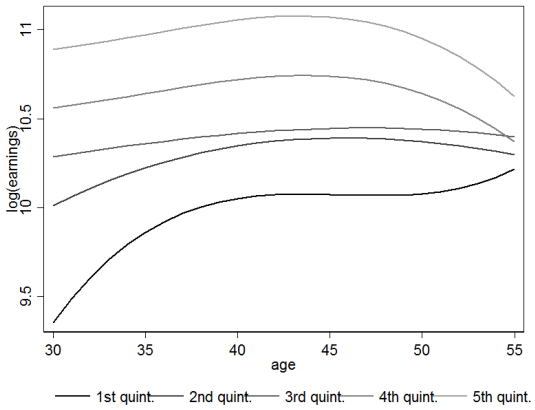
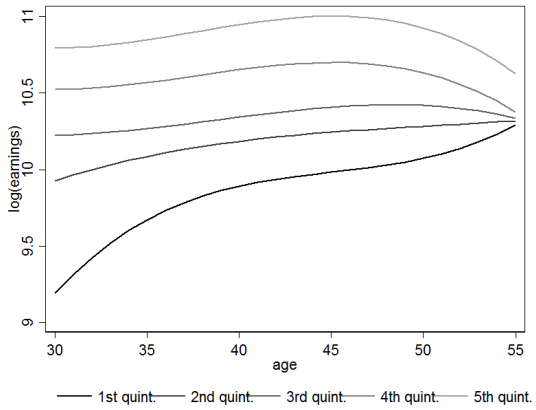


FIGURE 2.3 – Predicted earnings ($ 2010) by province group, family status at 30 y/o and earnings quintile at 30 y/o

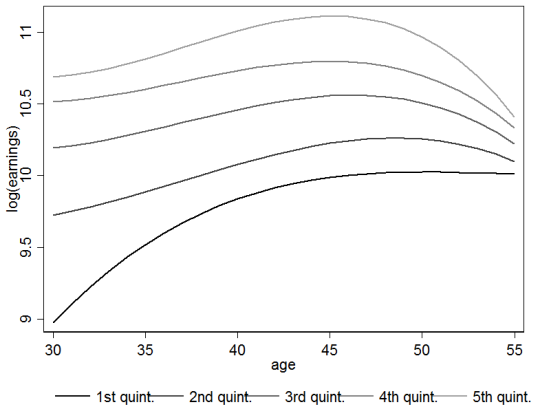
(a) Women - Quebec - single- no child



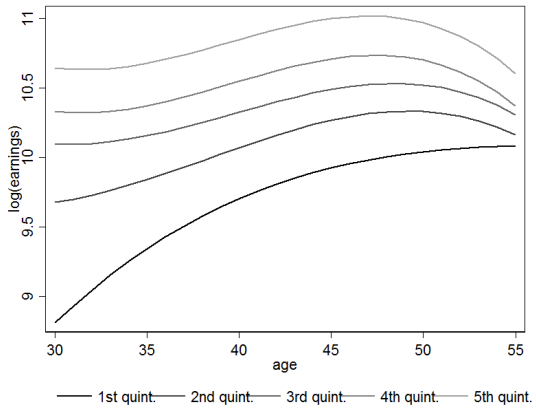
(b) Women - Quebec - couple- no child



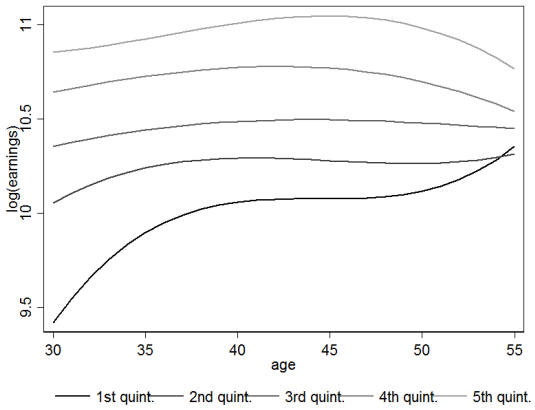
(c) Women - Quebec - single- with child



(d) Women - Quebec - couple- with child



(e) Women - Ontario - single- no child



(f) Women - Ontario - couple- no child

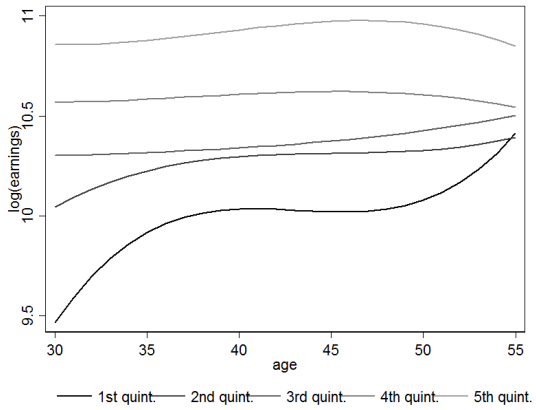
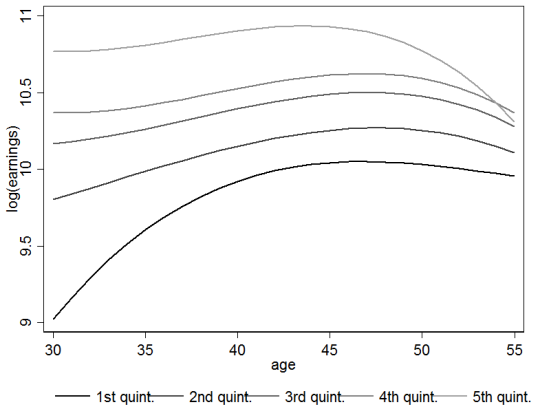
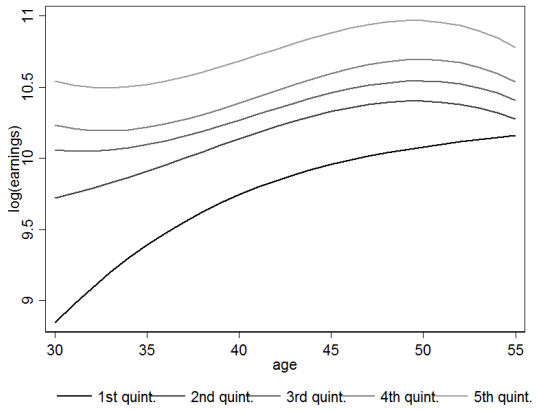


Figure 2.3 (continued) – Predicted earnings ($ 2010) by province group, family status at 30 y/o and earnings quintile at 30 y/o

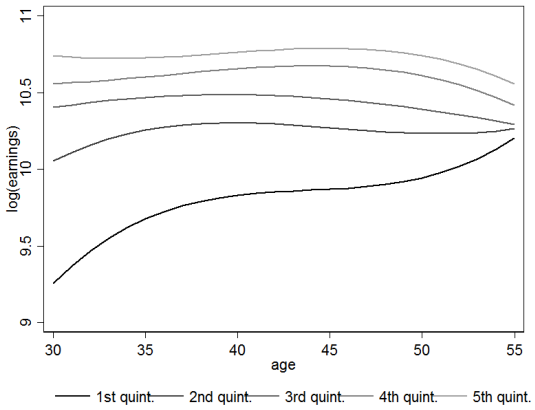
(g) Women - Ontario - single- with child



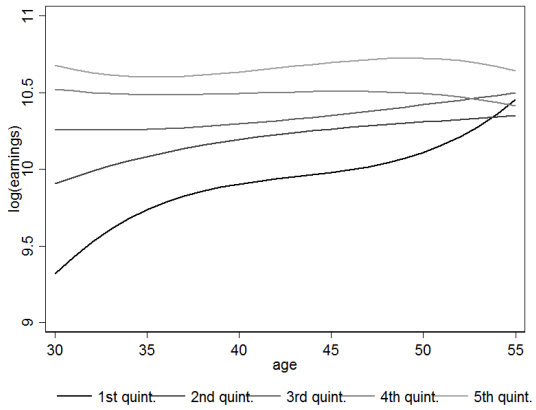
(h) Women - Ontario - couple- with child



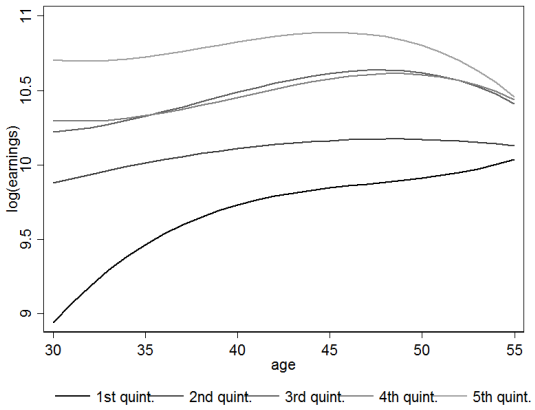
(i) Women - British Columbia - single- no child



(j) Women - British Columbia - couple- no child



(k) Women - British Columbia - single- with child



(l) Women - British Columbia - couple- with child

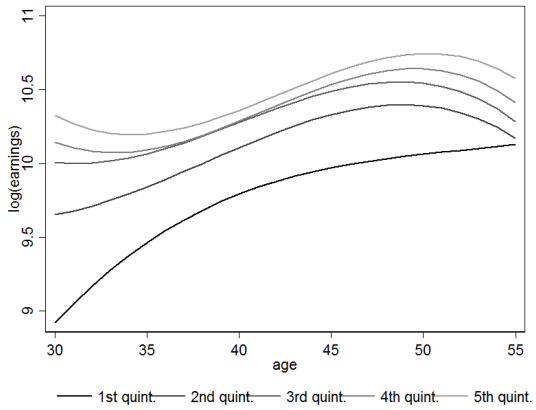
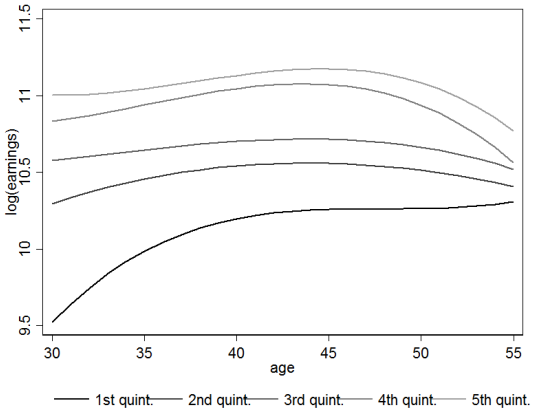
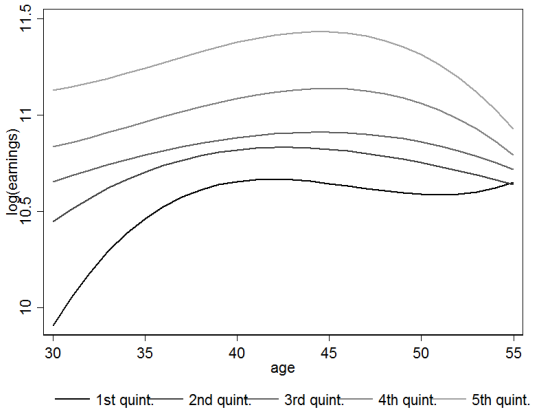


Figure 2.3 (continued) – Predicted earnings ($ 2010) by province group, family status at 30 y/o and earnings quintile at 30 y/o

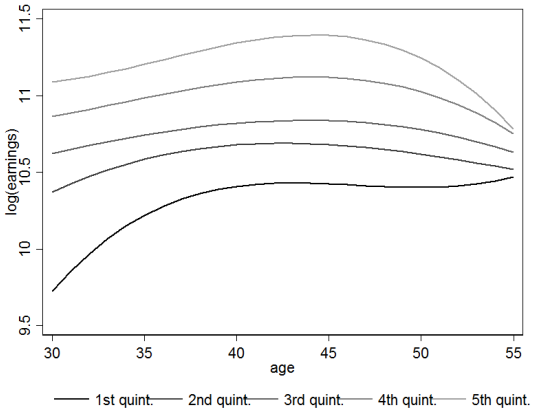
(m) Men - Quebec - single- no child



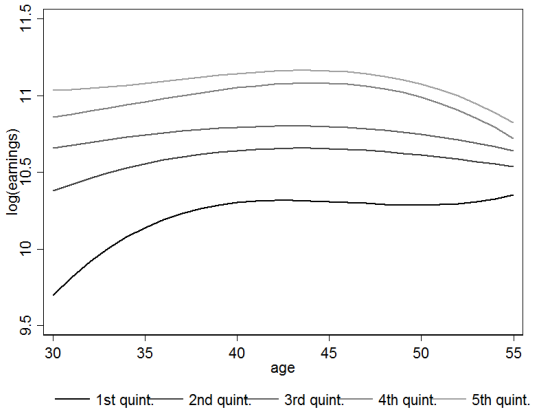
(n) Men - Quebec - couple- no child



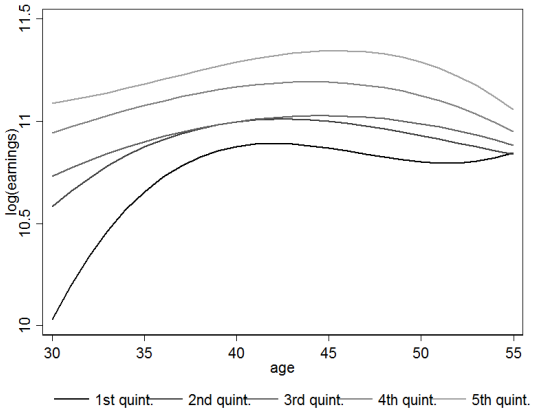
(o) Men - Quebec - couple- with child



(p) Men - Ontario - single- no child



(q) Men - Ontario - couple- no child



(r) Men - Ontario - couple- with child

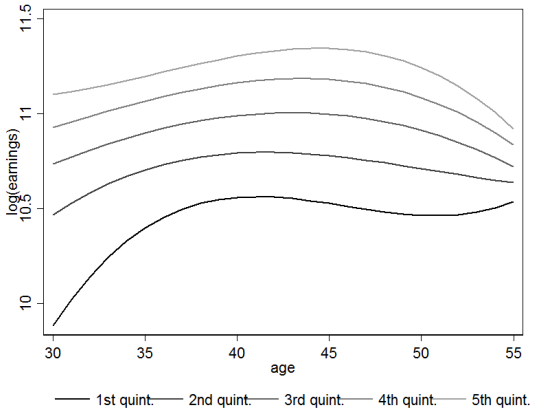
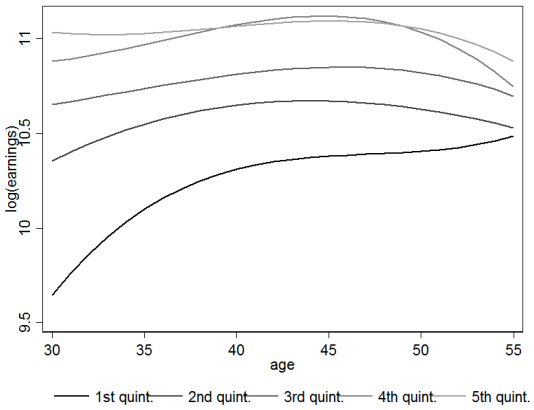
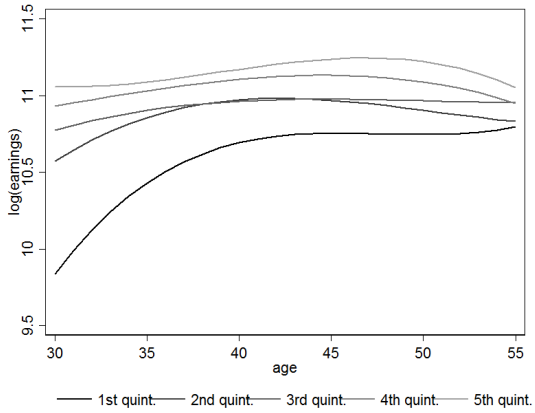


Figure 2.3 (continued) – Predicted earnings ($ 2010) by province group, family status at 30 y/o and earnings quintile at 30 y/o

(s) Men - British Columbia - single- no child



(t) Men - British Columbia - couple- no child



(u) Men - British Columbia - couple- with child

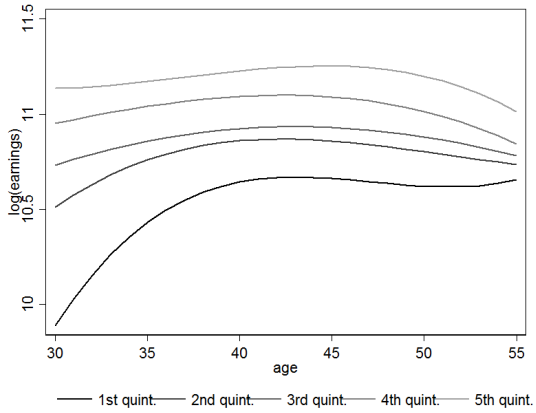
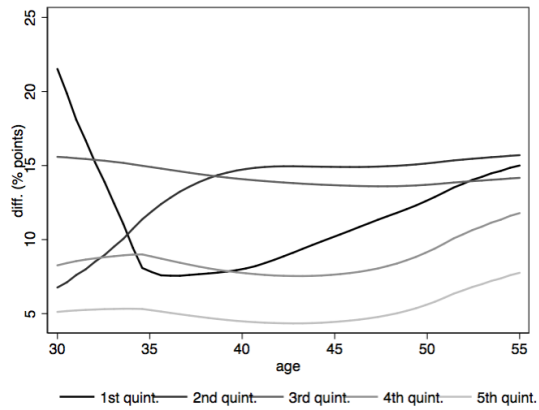
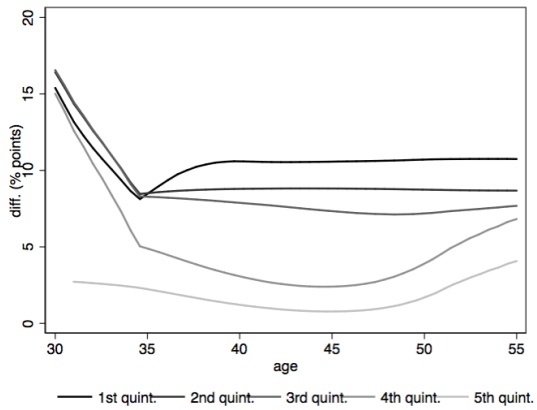


FIGURE 2.4 – Predicted difference between EMTRs on EET withdrawals and on EET contributions (in % points), by province, gender, family status at 30 y/o and earnings quintile at 30 y/o

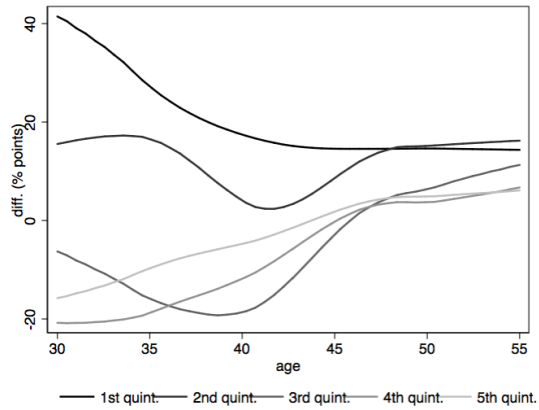
(a) Women - Quebec - single- no child



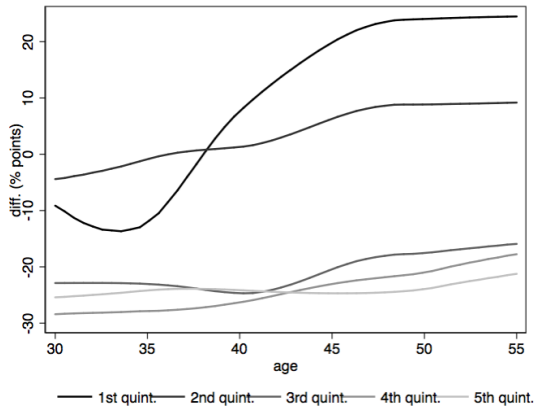
(b) Women - Quebec - couple- no child



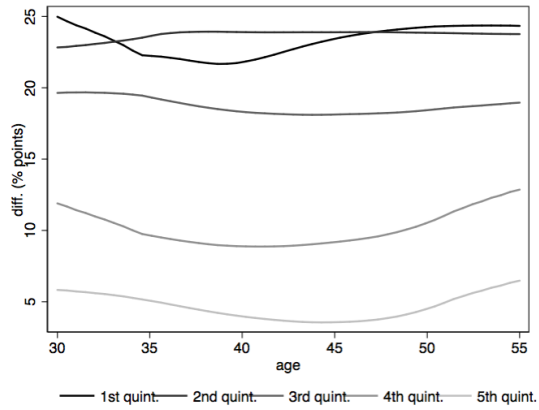
(c) Women - Quebec - single- with child



(d) Women - Quebec - couple- with child



(e) Women - Ontario - single- no child



(f) Women - Ontario - couple- no child

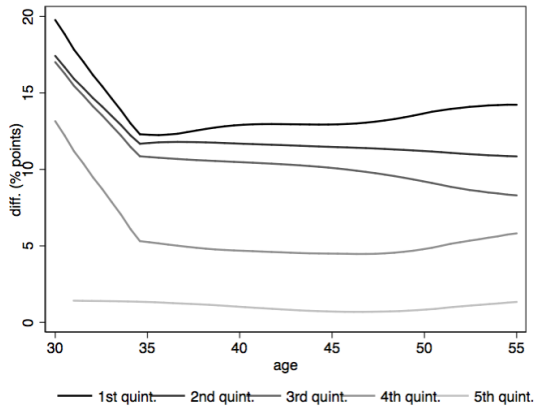
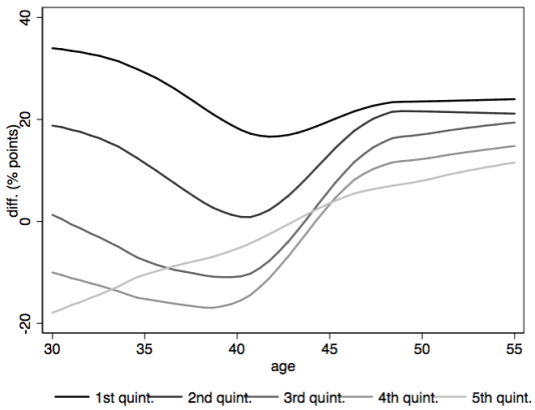
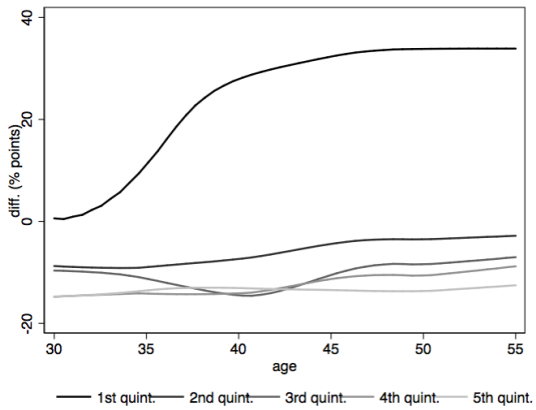


Figure 2.4 (continued) – Predicted difference between EMTRs on EET withdrawals and on EET contributions (in % points), by province, gender, family status at 30 y/o and earnings quintile at 30 y/o

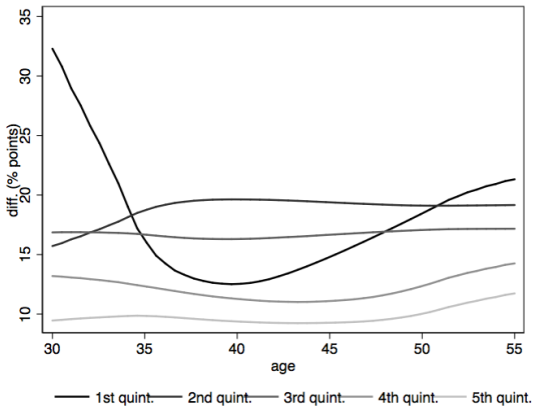
(g) Women - Ontario - single- with child



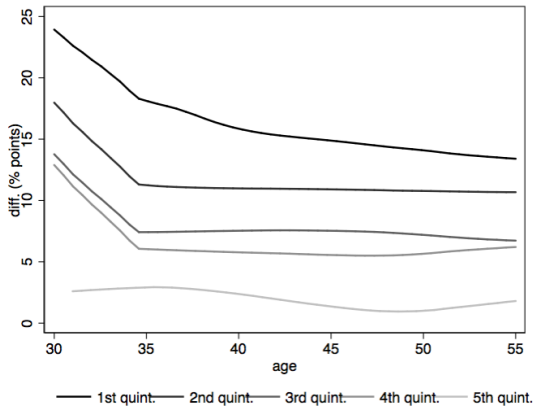
(h) Women - Ontario - couple- with child



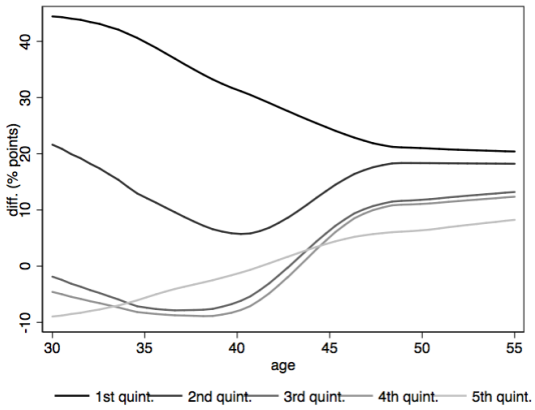
(i) Women - British Columbia - single- no child



(j) Women - British Columbia - couple- no child



(k) Women - British Columbia - single- with child



(l) Women - British Columbia - couple- with child

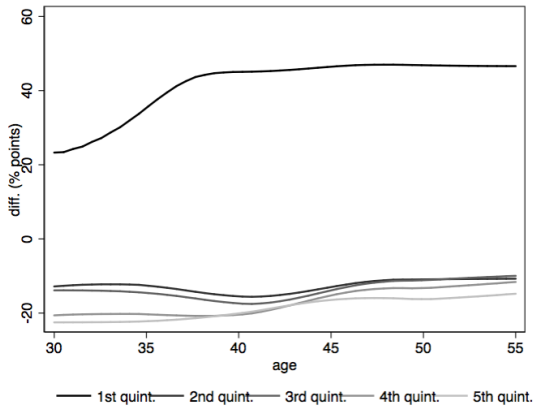
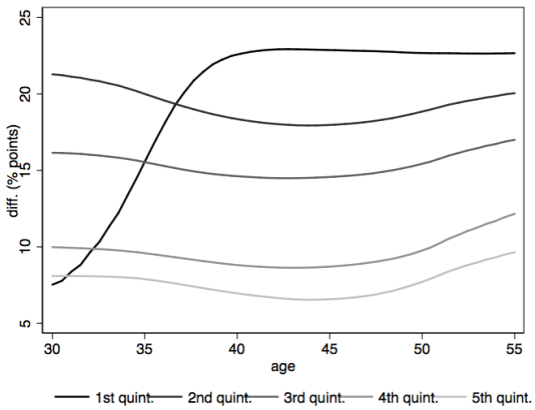
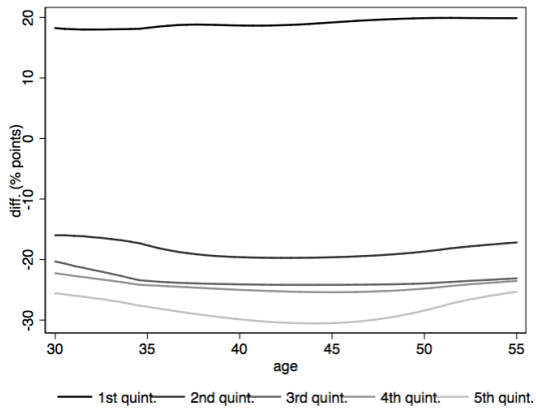


Figure 2.4 (continued) – Predicted difference between EMTRs on EET withdrawals and on EET contributions (in % points), by province, gender, family status at 30 y/o and earnings quintile at 30 y/o

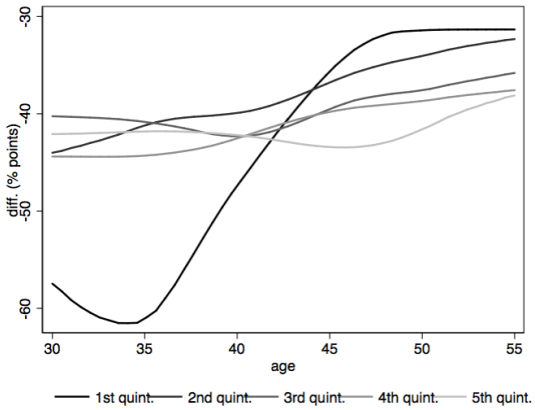
(m) Men - Quebec - single- no child



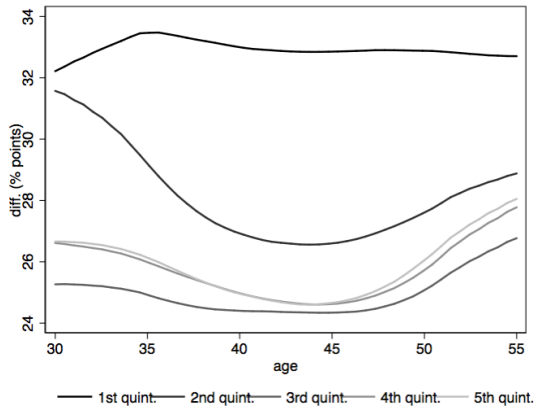
(n) Men - Quebec - couple- no child



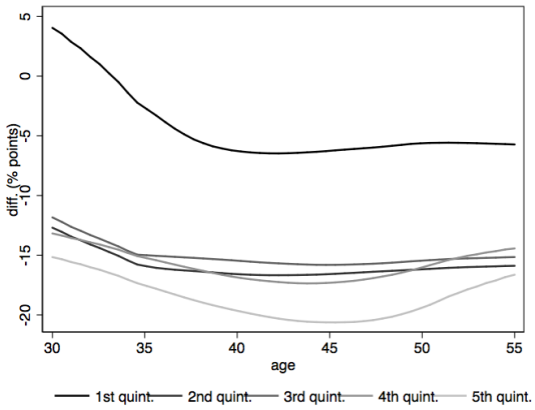
(o) Men - Quebec - couple- with child



(p) Men - Ontario - single- no child



(q) Men - Ontario - couple- no child



(r) Men - Ontario - couple- with child

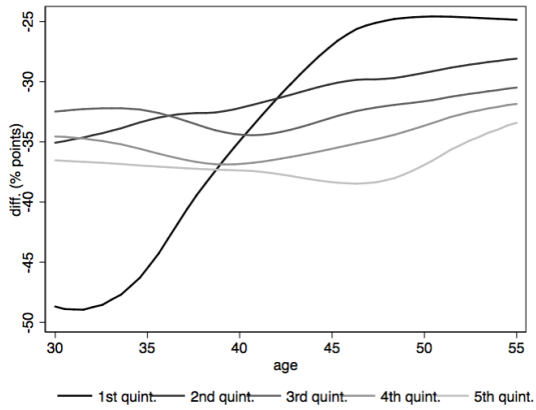
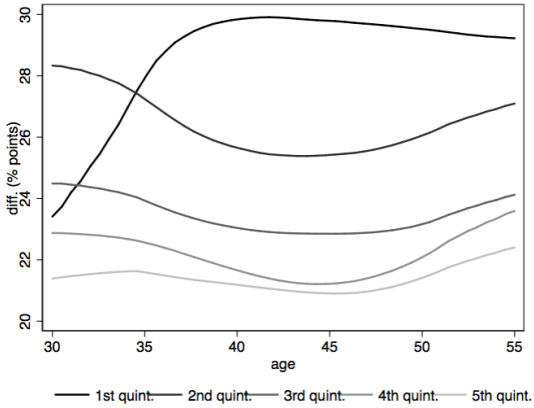
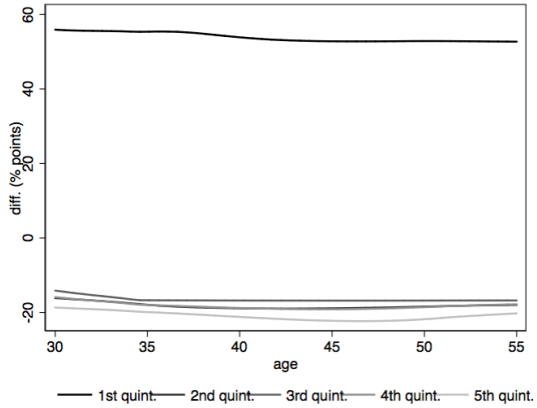


Figure 2.4 (continued) – Predicted difference between EMTRs on EET withdrawals and on EET contributions (in % points), by province, gender, family status at 30 y/o and earnings quintile at 30 y/o

(s) Men - British Columbia - single- no child



(t) Men - British Columbia - couple- no child



(u) Men - British Columbia - couple- with child

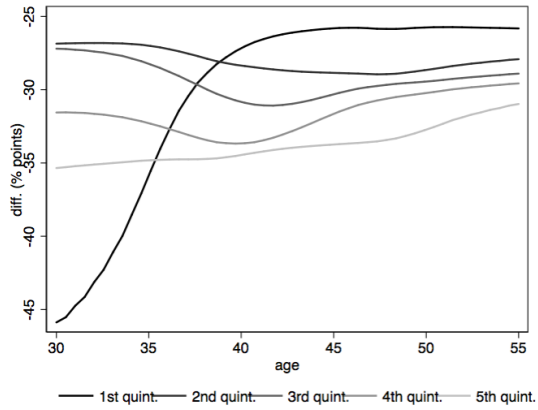
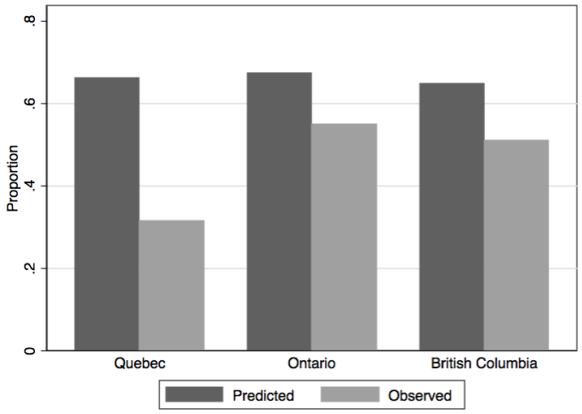
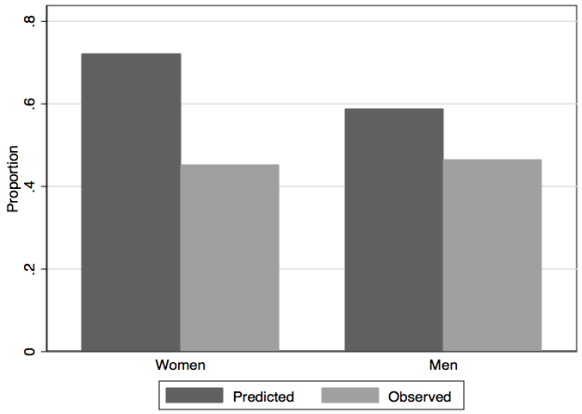


FIGURE 2.5 – Proportion of simulations for which TEE is predicted to be the optimal choice versus proportion of observations choosing TEE in the LAD

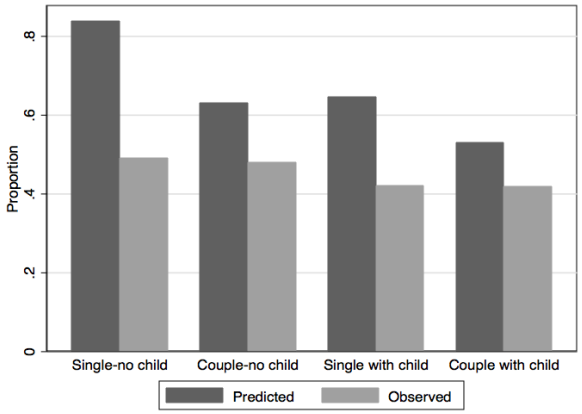
(a) Proportions by province



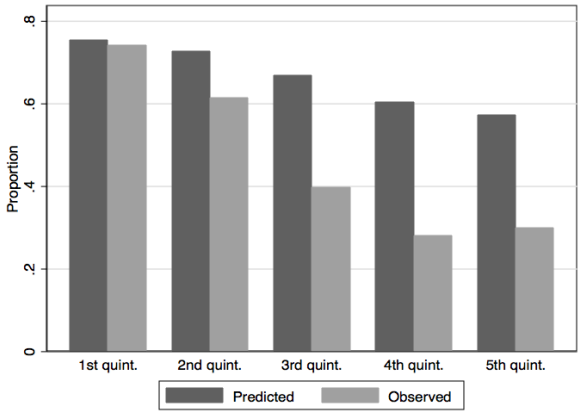
(b) Proportions by gender



(c) Proportions by family status



(d) Proportions by earnings quintile



2.11 Bibliography for Chapter 2

Beshears, J., J. J. Choi, D. Laibson, and B. C. Madrian (2009). The importance of default options for retirement saving outcomes : Evidence from the united states. In *Social security policy in a changing environment*, pp. 167–195. University of Chicago Press.

Beshears, J., J. J. Choi, D. Laibson, and B. C. Madrian (2017). Does front-loading taxation increase savings ? evidence from roth 401 (k) introductions. *Journal of public economics 151*, 84–95.

Cameron, A. C. and P. K. Trivedi (2005). *Microeconometrics : methods and applications*. Cambridge University Press.

Chetty, R. and J. N. Friedman (2014). Active vs. passive decisions and crowd-out in retirement savings accounts : Evidence from denmark. *Quarterly Journal of Economics 129*(3), 1141–1219.

Gourinchas, P.-O. and J. A. Parker (2002). Consumption over the life cycle. *Econometrica 70*(1), 47–89.

Marchand, S., L. Bissonnette, A. Blancquaert, and D. Jean-Yves (2015). Simtax documentation. Technical document, Industrial Alliance Research Chair on the Economics of Demographic Change.

Messacar, D. (2017). Crowd-out, education, and employer contributions to workplace pensions : Evidence from canadian tax records. *Review of Economics and Statistics* (0).

Milligan, K. (2002). Tax–preferred savings accounts and marginal tax rates : evidence on rrsp participation. *Canadian Journal of Economics/Revue canadienne d'économique 35*(3), 436–456.

Milligan, K. (2003). How do contribution limits affect contributions to tax-preferred savings accounts ? *Journal of Public Economics 87*(2), 253–281.

OECD (2015). Stocktaking of the tax treatment of funded private pension plans in oecd and eu countries.

Rothschild, M. and J. E. Stiglitz (1970). Increasing risk : I. a definition. *Journal of Economic theory 2*(3), 225–243.

Chapitre 3

Regression discontinuity designs with rounding errors and mismeasured treatment[1]

3.1 Résumé

Les erreurs d'arrondissement dans les régressions par discontinuité rendent souvent la variable de traitement inobservable pour certaines observations autour du seuil. Alors que les chercheurs rejettent généralement ces observations, je montre qu'elles contiennent des informations importantes, car la distribution de la variable dépendante se divise en deux en fonction de l'effet du traitement. L'intégration de cette information dans des critères standards de sélection de modèle aide à choisir la meilleure spécification du modèle et à éviter les biais de spécification.

3.2 Abstract

Rounding errors in the running variable of regression discontinuity designs often make the treatment variable unobservable for some observations around the threshold. While researchers usually discard these observations, I show that they contain valuable information because the outcome's distribution splits in two as a function of the treatment effect. Integrating this information in standard data driven criteria helps in choosing the best model specification and avoid specification biases.

1. I thank Luc Bissonnette, Charles Bellemare, Vincent Boucher, David Card, Bernard Fortin, Guy Lacroix and Thomas Lemieux for useful comments. I also thank the Fonds de recherche du Québec - Société et culture and the Social Sciences and Humanities Research Council for my scholarships.

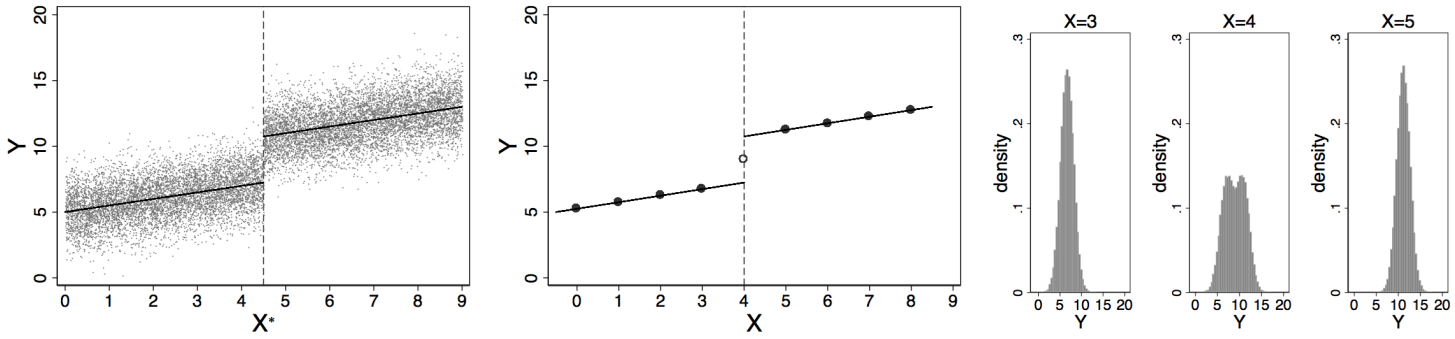
3.3 Introduction

In regression discontinuity designs (RDD), the probability of being treated changes exogeneously when a running variable crosses some threshold, so comparing the predicted outcome just below and above this latter allows to estimate the treatment's causal effect (Thistlethwaite and Campbell, 1960). In practice, the running variable is often rounded, and the treatment variable is unknown for observations closest to the threshold. A common example occurs when individuals are treated depending on their precise birth date, but observed age is rounded (e.g. Leuven and Oosterbeek (2004) and Dong (2015)). While researchers usually discard observations for which the treatment variable is unknown, this paper argues that they contain valuable information. Since these observations comprise both treated and untreated individuals, the distribution of the outcome splits as a function of the treatment effect. Beside providing additional graphical evidence of this treatment effect, these observations can be used for estimation under standard distributional assumptions.

As noted by Lee and Card (2008), when the running variable is discrete (or rounded), it is necessary to specify a parametric relationship between the running variable and the outcome. Since assuming a wrong specification biases the estimate of the treatment effect, one could mistake a sharper slope around the threshold for a discontinuity. I show that the observations with mismeasured treatment are useful to distinguish between the two, because a discontinuity splits the outcome's distribution, while a change in slope simply spreads it. I provide Monte Carlo evidence to the effect that integrating the information from these observations in standard data-driven model selection criteria improves performance and helps avoid specification bias.

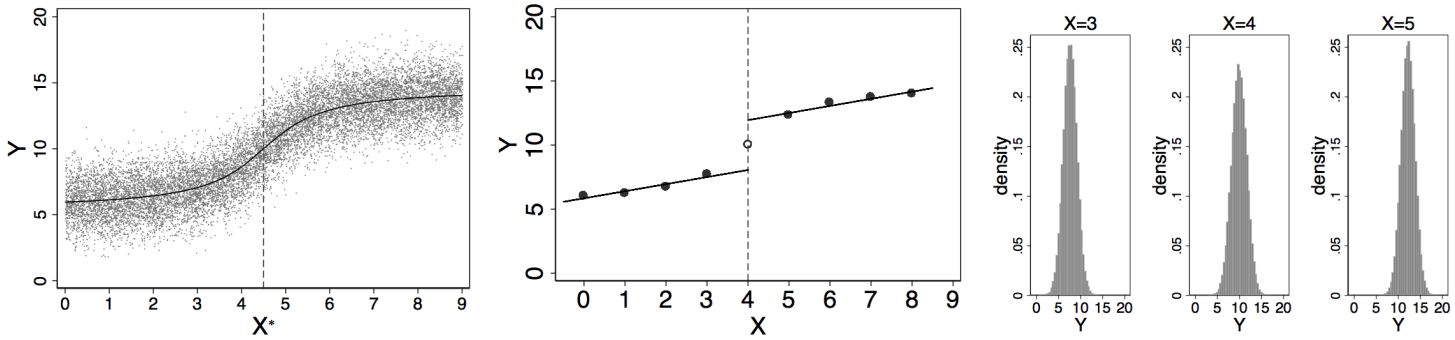
This paper contributes to the literature on measurement errors in RDD (see Hullegie and Klein (2010), Davezies and Le Barbanchon (2017) and Pei and Shen (2017) who study continuous measurement error problems) and, more precisely, to the literature on rounding error in RDD. While Lee and Card (2008) and Dong (2015), respectively, address inference issues and biases caused by rounding errors, they both assume that the threshold is an integer, so the researcher always knows whether the running variable is above or below the threshold. This paper is, to my knowledge, the first to address the importance of observations with mismeasured treatment caused by rounding errors.

The next section provides the intuition through graphical analyses, Section 3.5 suggests an estimation method, Section 3.6 highlights the method's performance through Monte Carlo simulations and Section 3.8 concludes.



(a) Discontinuity in $E(Y|X^*)$ (b) Discontinuity in $\hat{Y}|X$ (c) Distributions of $Y|X$

FIGURE 3.1 – Discontinuity in expected outcome



(a) Nonlinearity in $E(Y|X^*)$ (b) Nonlinearity mistaken for discontinuity (c) Distributions of $Y|X$

FIGURE 3.2 – Nonlinearity in expected outcome

3.4 Graphical analysis

Assume a sharp RDD : an individual receives a treatment $T(x^*)$ if x^* is greater than some threshold c. We are interested in the effect of $T(x^*)$ on a continuous outcome variable y. Figure 1(a) depicts a potential relationship. The solid line is $E(y|x^*)$ and the dashed line is c which is set to 4.5. The discontinuity in $E(y|x^*)$ at c results from the treatment effect. Now assume we observe x instead, which equals x^* rounded down to the nearest integer. A common practice is to present the average of y for each value of x and to show the discontinuity in the conditional predicted outcome at the rounded down threshold as shown in Figure 1(b). The observations at $x = 4$ cannot be used to depict the discontinuity because we do not know whether x^* lies before or after the threshold. However Figure 1(c) reveals an irregularity in the distribution of y when x equals 4 because these observations comprise both treated and untreated individuals, so two distributions overlap each other. This irregularity arises from the effect of the treatment and thus provides additional graphical evidence of its effect on y.

Assume now that the relationship between y and x^* is nonlinear as depicted in Figure 2(a).

The treatment effect is zero so there is no discontinuity. However, specifying the wrong model could lead one to mistake the sharper slope near the threshold for a discontinuity. The difficulty of choosing the right specification is further increased with rounding errors, especially when discarding observations near the threshold. Observing x instead of x^* could lead one to wrongly assume linearity, as shown in Figure 2(b). Notice how this figure resembles Figure 1(b), despite the difference between the true data generating processes. There is a difference, however, in the way the nonlinearity affects the distribution of y at $x = 4$, as depicted in Figure 2(c). The nonlinearity spreads the distribution but does not split it like the discontinuity does. Thus, data at $x = 4$ contain useful information to choose a polynomial specification that avoids biases in the estimate of the treatment effect.

3.5 Estimation method

Assume the following :

$$y_i = h(x_i^*) + \delta T(x_i^*) + \epsilon_i, \tag{3.1}$$

where $h(x_i^*)$ is a continuous function, δ is the treatment effect and ϵ_i is an error term. Assume a sharp RDD (i.e. $T(x_i^*) = 1$ if $x_i^* \geq c$ and 0 otherwise), where we only observe y_i and x_i, which equals x_i^* rounded down to the nearest integer. The threshold c is not an integer, so $T(x_i^*)$ is unobserved around c. I assume the following :

Assumption 1 *$f_{x_i^*}(x_i^*|x_i)$ is uniform on the interval $[x_i, x_i + 1)$.*

Assumption 2 *$\epsilon_i \sim N(0, \sigma^2)$.*

Both assumptions are easily adaptable. An alternative to Assumption 1 could be to use an empirical distribution observed from an external database (e.g. an observed distribution of birth dates). Assumption 2 can be tested using the observed distributions of y_i at values of x_i that are not at the threshold. An alternative distribution that would better fit the data could be used instead. Under Assumption 2, individual i's contribution to the likelihood is :

$$L_i(\boldsymbol{\beta}, \delta|y_i, x_i^*) = f(y_i|x_i^*) = \frac{1}{\sigma}\phi\left(\frac{y_i - h(x_i^*) - \delta T(x_i^*)}{\sigma}\right), \tag{3.2}$$

where ϕ is the density of the standard normal. Because we observe x_i instead of x_i^*, I use individual i's expected contribution to the likelihood which, from Assumption 1, is :

$$EL_i(\boldsymbol{\beta}, \delta|y_i, x_i) = \begin{cases} \frac{1}{\sigma}\int_{x_i}^{x_i+1} \phi\left(\frac{y_i - h(u)}{\sigma}\right) du & \text{if } x_i < \underline{c}, \\ \frac{1}{\sigma}\int_{x_i}^{c} \phi\left(\frac{y_i - h(u)}{\sigma}\right) du + \frac{1}{\sigma}\int_{c}^{x_i+1} \phi\left(\frac{y_i - h(u) - \delta}{\sigma}\right) du & \text{if } x_i = \underline{c}, \\ \frac{1}{\sigma}\int_{x_i}^{x_i+1} \phi\left(\frac{y_i - h(u) - \delta}{\sigma}\right) du & \text{if } x_i > \underline{c}, \end{cases} \tag{3.3}$$

where $\underline{c}$ equals c rounded down to the nearest integer. The integrals are algebraically solvable if $h''(x_i^*) = 0$. For instance, assuming [2]

$$h(x_i^*) = \beta_0 + \beta_1 x_i^* + \beta_1^a (x_i^* - c) T(x_i^*), \tag{3.4}$$

we have :

$$EL_i(\boldsymbol{\beta}, \delta | y_i, x_i) = \begin{cases} \frac{-1}{\beta_1}\left[\Phi\left(\frac{y_i - h(x_i+1)}{\sigma}\right) - \Phi\left(\frac{y_i - h(x_i)}{\sigma}\right)\right] & \text{if } x_i < \underline{c}, \\ \frac{-1}{\beta_1}\left[\Phi\left(\frac{y_i - h(c)}{\sigma}\right) - \Phi\left(\frac{y_i - h(x_i)}{\sigma}\right)\right] - & \\ \quad \frac{1}{(\beta_1 + \beta_1^a)}\left[\Phi\left(\frac{y_i - h(x_i+1) - \delta}{\sigma}\right) - \Phi\left(\frac{y_i - h(c) - \delta}{\sigma}\right)\right] & \text{if } x_i = \underline{c}, \\ \frac{-1}{(\beta_1 + \beta_1^a)}\left[\Phi\left(\frac{y_i - h(x_i+1) - \delta}{\sigma}\right) - \Phi\left(\frac{y_i - h(x_i) - \delta}{\sigma}\right)\right] & \text{if } x_i > \underline{c}. \end{cases} \tag{3.5}$$

If $h''(x_i^*) \neq 0$, numerical approximations can be used instead. It is then straightforward to maximize $\sum_i^N log\left(EL_i(\boldsymbol{\beta}, \delta | y_i, x_i)\right)$.

3.6 Monte Carlo simulations

I generate x_i^* from a uniform distribution on the range $[-5, 6]$, its rounded down value x_i, and a threshold of 0.5. As is standard practice, I normalize the threshold to zero, and x_i^* and x_i to $\tilde{x}_i^* = x_i^* - 0.5$ and $\tilde{x}_i = x_i - 0.5$ respectively. I generate y_i according to :

$$y_i = 3T(\tilde{x}_i^*) + h(\tilde{x}_i^*) + \epsilon_i \tag{3.6}$$

$$h(\tilde{x}_i^*) = 1.5\tilde{x}_i^* + 0.15\tilde{x}_i^{*2} + \left[0.5\tilde{x}_i^* - 0.1\tilde{x}_i^{*2}\right] T(\tilde{x}_i^*) \tag{3.7}$$

where $T(\tilde{x}_i^*)$ equals one if $\tilde{x}_i^* \geq 0$, the treatment effect equals three, and $\epsilon_i \sim N(0, 3)$. Note that, with rounding, $T(\tilde{x}_i^*)$ is unobserved at $\tilde{x}_i = -0.5$. The sample size is set to 1000. I seek to discriminate between the two following specifications :

$$h(\tilde{x}_i^*) = h_1(\tilde{x}_i^*) \equiv \beta_0 + \beta_1 \tilde{x}_i^* + \beta_1^a \tilde{x}_i^* T(\tilde{x}_i^*), \tag{3.8}$$

$$h(\tilde{x}_i^*) = h_2(\tilde{x}_i^*) \equiv \beta_0 + \beta_1 \tilde{x}_i^* + \beta_2 \tilde{x}_i^2 + T(\tilde{x}_i^*)\left(\beta_1^a \tilde{x}_i^* + \beta_2^a \tilde{x}_i^{*2}\right). \tag{3.9}$$

Naturally, assuming $h(\tilde{x}_i^*) = h_1(\tilde{x}_i^*)$ will yield biased estimates because of the specification error. I estimate the model using each specification for three estimation methods. Importantly, all of them consist of maximum likelihood and assume that ϵ_i follows a normal distribution, even though this assumption is not needed in conventional methods. Therefore, differences in results across estimation methods will not result from this assumption.

2. It is widely acknowledged that RDD estimators should allow the slope to adjust before and after the threshold (e.g. see Lee and Lemieux (2010)).

a) Full information estimation – There is no rounding problem ; the objective function is :

$$Q_{full} = \sum_{i=1}^{1000} ln \left[\frac{1}{\sigma} \phi \left(\frac{y_i - h(\tilde{x}_i^*) - \delta T(\tilde{x}_i^*)}{\sigma} \right) \right]. \tag{3.10}$$

b) Rounding error - conventional estimation – We observe $\tilde{x}_i$ rather than $\tilde{x}_i^*$. The 1000 observations are all generated at values of $\tilde{x}_i$ other than -0.5, so the performance of this estimation relative to the others does not result from a smaller number of observations because of discarding. The objective function is :

$$Q_{err} = \sum_{i=1}^{1000} ln \left[\frac{1}{\sigma} \phi \left(\frac{y_i - h(\tilde{x}_i) - \delta T(\tilde{x}_i)}{\sigma} \right) \right]. \tag{3.11}$$

Dong (2015) showed that the estimate of the treatment effect based on this approach is biased even when guessing the right specification if the slope or higher derivatives change at the threshold. I therefore use her correction to correct the estimate of each simulation. [3]

c) Rounding error - proposed estimation – We observe $\tilde{x}_i$ instead of $\tilde{x}_i^*$ and the method from Section 3.5 is used. For the first order polynomial specification, I maximize $\sum_{i=1}^{1000} ln\left(EL_i(\boldsymbol{\beta}, \delta | y_i, \tilde{x}_i)\right)$ in equation (3.5). For the second order polynomial specification, because the integrals of equation (3.14) have no algebraic solution, I rather maximize $\sum_{i=1}^{1000} ln\left(EL_i(\boldsymbol{\beta}, \delta | y_i, \tilde{x}_i)\right)$ in equation (3.14), approximating the integrals numerically. [4]

Table 3.1 presents the average estimates across 1000 simulations for these three estimations in Columns (a), (b) and (c), respectively. As expected, the first order polynomial model leads to biased estimates of the treatment effect for all estimations, while the second order polynomial specification leads to essentially unbiased estimates. Columns (b) and (c) show that rounding errors result in a loss of efficiency, but that the proposed estimation method attenuates this loss compared to the conventional estimation. The proposed method also significantly improves the capacity of both the Akaike information criterion (AIC) and the Likelihood-ratio (LR) test to favour the right model, relative to the conventional estimation.

3.7 Extension to RDD with binary outcome

If the outcome of interest is binary rather than continuous, one may not observe a split or a spread in the outcome's distribution, so the method proposed in this paper will provide no

3. For the first order specification, the formula to obtain the unbiased estimate is $\hat{\delta} = \hat{\delta}_u - \frac{\hat{\beta_1^a}}{2}$, where $\hat{\delta}_u$ is the uncorrected estimate obtained from maximizing equation (3.11). For the second order specification, it is $\hat{\delta} = \hat{\delta}_u - \frac{\hat{\beta_1^a}}{2} + \frac{\hat{\beta_2^a}}{6}$.

4. I use the composite Simpson's rule where each range starting from $\tilde{x}_i$ and ending at $\tilde{x}_i + 1$ is split in 100 subintervals.

TABLE 3.1 – Monte Carlo simulations : average estimate of the treatment effect ($\delta = 3$)

	Full information	**Rounding errors**	
		Conventional	Proposed
	(a)	(b)	(c)
1^{st} order polynomial	3.51	3.69	3.59
	(0.39)	(0.46)	(0.41)
2^{nd} order polynomial	3.01	3.00	3.01
	(0.59)	(0.95)	(0.72)
AIC succes rate*	0.81	0.56	0.72
LR test 10% reject rate**	0.77	0.50	0.66
N. obs.	1000	1000	1000

Standard deviation of estimated coefficients in parentheses

1000 replications

$*$ Prop. of replications where AIC favours the 2^{nd} order polynomial

$**$ Prop. of replications where LR test rejects the 1^{st} order polynomial with a 10% confidence level

additional information to help in distinguishing between a true discontinuity and a change in slope around the threshold. Therefore, the only benefit the approach may yield is an increase in the precision of the estimate resulting from the observations that are not discarded. However, even though the benefit of the approach will be less important with a binary outcome, it is important to note that the cost of the approach – in terms of additional assumptions that need to be made – will also be less important. Indeed, in binary outcome models, researchers usually already make distributional assumptions on the error term (e.g. a normality assumption for probit estimations). Thus, if one is confident that the running variable is drawn from a uniform distribution (or any other distribution), there is no reason to discard observations around the threshold. The remainder of this subsection provides an estimation method for RDDs with binary outcome variables and the Monte Carlo simulations that confirm the intuition above.

Assume the following :

$$y_i^* = h(x_i^*) + \delta T(x_i^*) + \epsilon_i, \tag{3.12}$$

where y_i^* is a latent variable and all other variables and parameters are as defined in Section 3.5. The observable outcome variable of interest is y_i which equals one if $y_i^* \geq 0$ and zero otherwise. I again let x_i be equal to x_i^* rounded down to the nearest integer. I also again make assumptions 1 ($f_{x_i^*}(x_i^*|x_i)$ is uniform on the interval $[x_i, x_i + 1)$) and 2 ($\epsilon_i \sim N(0, \sigma^2)$),

Without rounding error, under assumption 2 (normality of ϵ_i), and normalizing σ^2 to one, the contribution of individual i to the likelihood simply corresponds to the likelihood of a probit

model :

$$L_i(\boldsymbol{\beta}, \delta | y_i, x_i^*) = \Phi\Big((2y_i - 1)\Big[h(x_i^*) + \delta T(x_i^*)\Big]\Big). \tag{3.13}$$

With rounding errors, under assumption 1, we can use the following expected likelihood :

$$EL_i(\boldsymbol{\beta}, \delta | y_i, x_i) = \begin{cases} \int_{x_i}^{x_i+1} \Phi\Big((2y_i - 1)h(x_i^*)\Big)du & \text{if } x_i < \underline{c}, \\ \int_{x_i}^{c} \Phi\Big((2y_i - 1)h(u)\Big)du + & \\ \int_{c}^{x_i+1} \Phi\Big((2y_i - 1)\Big[h(u) + \delta\Big]\Big)du & \text{if } x_i = \underline{c}, \\ \int_{x_i}^{x_i+1} \Phi\Big((2y_i - 1)\Big[h(u) + \delta\Big]\Big)du & \text{if } x_i > \underline{c}, \end{cases} \tag{3.14}$$

where the integrals may be approximated numerically. I conduct Monte Carlo simulations using the following data generation process :

$$y_i^* = T(\tilde{x}_i^*) + \frac{1}{3}h(\tilde{x}_i^*) + \epsilon_i \tag{3.15}$$

$$h(\tilde{x}_i^*) = -2 + 1.5\tilde{x}_i^* + 0.15\tilde{x}_i^{*2} + \left[0.5\tilde{x}_i^* - 0.1\tilde{x}_i^{*2}\right] T(\tilde{x}_i^*), \tag{3.16}$$

where y_i^* is a latent variable and ϵ_i follows a normal distribution with a variance of one. Note that the parameters are different than those of Section 3.6, because they are chosen to yield a significant proportion of values of y_i^* both above and below zero for the whole domain of x_i^*. I generate the observable outcome y_i, which equals one if $y_i^* \geq 0$ and zero otherwise. I estimate the model using the three estimation methods and the two polynomial specifications for $h(x_i^*)$ described in Section 3.6.

Table 3.2 presents the results of the simulations. Comparing columns (b) and (c) reveals no benefit of the approach when the dependent variable is binary. It is important to note that the proposed approach would in practice yield the additional benefit of an increased number of observations. This is not shown in the results below : all estimations are made using the same number of observations (because observations from column (b) are all generated elsewhere than at the threshold and are thus not discarded). Since the proposed method would in practice increase the number of observations used for estimation, the approach could be used to increase precision of the estimate, and this benefit would be more important the more observations there is at the threshold.

3.8 Discussion and potential extensions

The estimation method suggested above yields more precise estimates of the treatment effect than conventional methods and helps chose the right model specification to avoid biases. It

TABLE 3.2 – Monte Carlo simulations : average estimate of the treatment effect ($\delta = 1$) with binary outcome

	Full information	**Rounding errors**	
		Conventional	Proposed
	(a)	(b)	(c)
1^{st} order polynomial	1.23	1.36	1.32
	(0.08)	(0.10)	(0.10)
2^{nd} order polynomial	1.00	1.00	0.99
	(0.12)	(0.20)	(0.21)
AIC succes rate*	0.84	0.62	0.63
LR test 10% reject rate**	0.80	0.56	0.56
N. obs.	5000	5000	5000

Standard deviation of estimated coefficients in parentheses

1000 replications

$*$ Prop. of replications where AIC favours the 2^{nd} order polynomial

$**$ Prop. of replications where LR test rejects the 1^{st} order polynomial with a 10% confidence level

should contribute to expand the RDD methodology to applications for which the rounding errors are currently deemed too important to provide convincing quasi-experimental designs.

The proposed methodology is built on parametric assumptions, which contrasts with non-parametric methods often favoured in the RDD literature (see Hahn et al. (2001)). However, many points are important to keep in mind. First, as noted by Lee and Card (2008), with rounding errors in the running variable of RDD, one has no choice but to assume a parametric relationship between the running variable and the outcome. Secondly, the distributional assumption made on the true value of the running variable is easily adaptable and will often be easy to justify. For example, in a RDD where the true running variable is the date of birth, but where the data only provides the year of birth, one could probably assume that births are uniformly distributed across all dates. What is more, this assumption could be tested on an external database that would provide the information on the distribution of birth dates. If this assumption does not seem appropriated, the researcher can easily use the empirical distribution of birth dates instead, an approach also used by Davezies and Le Barbanchon (2017) for continuous measurement errors. Lastly, the distributional assumption made on the outcome can be tested at values of the running variable elsewhere than the threshold, and an alternative distribution that fits the data better can easily be used. With binary outcome, distributional assumptions on the latent variable cannot be tested, but such assumptions are in any case already imposed in probit or logit estimations.

The proposed estimation method also assumed the homoscedasticity of the error term for simplicity. With continuous outcome, it would be possible to relax this assumption if assuming that the variance of the error term is a more general– and continuous– function of the running variable. The function would need to be continuous to rule out that the split in the distribution of the outcome at the threshold simply results from a local jump in the variance of the error term at this threshold. Allowing for heteroskedasticity would however require to estimate the additional parameters from the parametric relationship.

The method proposed in this paper will probably be more appealing to researchers when the number of observations for each rounded value of the running variable is very large. The larger it is, the higher is the number of observations discarded by conventional methods. It should prove useful given the increasing use of administrative data, which sometimes comprise millions of observations per year (e.g. the LAD database used in the previous chapter of this thesis). It is likely that discarding millions of observations will result in a significant loss of relevant information on the outcome around the threshold.

Finally, it is important to keep in mind that the analysis assumes a sharp RDD. Therefore, a natural extension would be to extend the method to the more general fuzzy RDD. In a fuzzy RDD, the treatment variable is not determined by whether or not the running variable crosses the threshold, but crossing the threshold causes an exogenous change in the probability of being treated (the sharp RDD is the special case in which this probability changes from zero to one or from one to zero depending on whether or not the running variable crosses the threshold). Therefore, in a fuzzy RDD, the dummy variable indicating whether the running variable in greater than the threshold is used as an instrumental variable for the treatment, rather than as the treatment itself. In such a design, one would therefore observe the treatment variable at the rounded value of the threshold, but not the exact value of its instrument. This would therefore result in a different problem that is left for future research.

3.9 Bibliography for Chapter 3

Davezies, L. and T. Le Barbanchon (2017). Regression discontinuity design with continuous measurement error in the running variable. *Journal of Econometrics 200*, 260–281.

Dong, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics 30*(3), 422–446.

Hahn, J., P. Todd, and W. Van der Klaauw (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica 69*(1), 201–209.

Hullegie, P. and T. J. Klein (2010). The effect of private health insurance on medical care utilization and self-assessed health in germany. *Health economics 19*(9), 1048–1062.

Lee, D. S. and D. Card (2008). Regression discontinuity inference with specification error. *Journal of Econometrics 142*(2), 655–674.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*, 281–355.

Leuven, E. and H. Oosterbeek (2004). Evaluating the effect of tax deductions on training. *Journal of Labor Economics 22*(2), 461–488.

Pei, Z. and Y. Shen (2017). The devil is in the tails : Regression discontinuity design with measurement error in the assignment variable. In M. D. Cattaneo and J. C. Escanciano (Eds.), *Regression Discontinuity Designs (Advances in Econometrics, Volume 38)*, Chapter 12, pp. 455–502. Emerald Publishing Limited.

Thistlethwaite, D. L. and D. T. Campbell (1960). Regression-discontinuity analysis : An alternative to the ex post facto experiment. *Journal of Educational Psychology 51*(6), 309.

Conclusion

In this thesis, I have used and developed microeconometrics methods and applied them to innovative and large databases. Applying these methods to risk-taking and savings decisions resulted in many interesting findings. By combining information on the formation of a network of entrepreneurs with observations from a lab-in-the-field, I estimated social conformity effects while controlling for homophily. I found that entrepreneurs tend to conform with their peers' choices, which suggests that social interactions play a role in shaping risk-taking behaviours. I also found that individuals tend to develop relationships with others based on some characteristics linked to cognitive ability that are not easily observable. These findings open exciting paths for future research. For example, the social networking activity that was organized within which we conducted our experiments could be conducted again with a focus on discussing important risk-taking decisions, such as insurance choices and loans possibilities. If social conformity can also affect these decisions, it may push behaviours toward peers' average behaviour, reducing excessive risk-taking and increasing risk tolerance for excessively risk averse individuals, possibly improving these entrepreneurs' outcomes.

Also, by estimating an econometric model of income dynamics on very rich administrative data, I provided new insights on the suitability of the two main types tax-preferred savings accounts. My results suggest that TEE savings vehicles tend to yield higher returns than EET in Canada, especially for the lowest income groups. Considering this, my other finding that TEE is much less favoured in Quebec than in other provinces should be taken seriously. This difference does not seem to arise from differences in income dynamics or tax codes. Future research should explore whether this stems from lower financial literacy or other unobserved factors specific to Quebec.

Finally, the method proposed in the third chapter for regression discontinuity designs with rounding errors in the running variable should also stimulate new research. Administrative database, such as the LAD used in the second chapter, are increasingly available and often comprise millions of observations per year. Regression discontinuity estimations focus on estimating a local treatment effect of a reform y comparing observations for which the running variable is just above the threshold to those for which it is just below. But since conventional methods often discard all observations closest to this threshold, they probably discard signi-

ficant information very much relevant to measuring this local treatment effect. Generalizing the proposed method to fuzzy regression discontinuity designs would widen potential applications vastly. For example, educational reforms often lead to fuzzy designs, and the students are affected or not by the reform depending on their exact birth date. Large administrative databases with data on income, and in which only birth year is observed, could then provide a compelling setting for taking advantage of the proposed estimation method.