Sommaire

Introduction générale **1**

1 Analyses du comportement de corruption **15**
1.1 Jeu de corruption à trois joueurs 54
1.2 Protocole de l'expérience 59
1.3 Déterminants du comportement de corruption 66
1.4 Conclusion 77
Annexes 81

2 Coût et qualité de l'offre de soins, quelle(s) rémunération(s) ? **89**
2.1 Institutions : la contrainte budgétaire des médecins du Québec 97
2.2 Analyse théorique du passage à la Rémunération Mixte 119
2.3 Modèle économétrique 138
2.4 Résultats : les vertus de la flexibilité 148
2.5 Conclusion 167
Annexes 171

3 Demande de travail au noir en environnement concurrentiel **195**
3.1 Demande de travail au noir et concurrence à la Bertrand 203
3.2 Dénonciation : l'équilibre de silence collusif 213
3.3 Présentation des marchés expérimentaux 225
3.4 Malédiction de Bertrand et évasion collusive : résultats empiriques . . 232
3.5 Conclusion 256
Annexes 259

Conclusion générale **280**

Bibliographie **281**

Table des matières **317**

Introduction Générale

> *Although economists may speak of 'the agency problem,' agency is in fact a solution, a neat kind of social plumbing. The problem is the ancient and ineluctable one of how to attain and maintain control in order to carry out definite, yet varying purposes.*
>
> H.C. White (1985, p. 188)

Dans le numéro spécial de l'*European Journal of the History of Economic Thought* destiné à tirer le bilan du dernier siècle, Kenneth Arrow – prix Nobel d'économie 1972 – comptait l'économie de l'information pour l'un des cinq progrès les plus importants de la discipline (Arrow, 2001a). Cette appellation, très large, recouvre moins un champ à part entière que la reconnaissance de l'importance que revêt la distribution de l'information entre les agents économiques (*[. . .] new point is that information is a variable*»Arrow, 2001a p.300). Sous l'impulsion, entres autres, de ses propres travaux (Arrow, 1963) et grâce aux outils d'analyse issus des avancées de la théorie des jeux,[1] la théorie de l'agence a fourni un cadre unifié permettant de formaliser les problèmes économiques que posent ces asymétries d'information. Au regard des objectifs que le recul a permis de lui assigner, et qui servent d'exergue à ce travail, cette littérature doit en particulier son succès à la compréhension des conditions sous lesquelles les incitations parviennent à

[1] Selten (1965 ; 1975) et Harsanyi (1967a ; 1967b ; 1968) notamment. Myerson (2004) propose une discussion de l'apport de la théorie des jeux aux développements de l'économie de l'information et singulièrement de la théorie de l'agence.

réconcilier les intérêts divergents du principal et de l'agent.[2] Ces conditions dépendent de façon cruciale du contexte institutionnel dans lequel s'inscrivent les transactions (Smith, 1962). Offrant une description plus fine et techniquement rigoureuse de cet environnement, ces travaux fournissent alors des outils puissants d'aide à la décision comme en témoigne, par exemple, la synthèse récente proposée par Laffont (2000).

De fait, le souci de doter les travaux scientifiques d'une représentation enrichie de l'environnement économique constitue le point de départ de cette littérature. En suivant Salanié (1997, p.2), on peut en effet considérer l'abandon du cadre simplifié offert par la théorie de l'équilibre général comme l'acte fondateur de la théorie de l'agence :

> *« The theory of contracts originates in the failures of general equilibrium theory [. . .]. The idea is to turn away temporarily from general equilibrium models, whose description of the economy is consistent but not realistic enough, and to focus on necessarily partial models that take into account the full complexity of strategic interactions between privately informed agents in well-defined institutional. It was hoped then that lessons drawn from these studies could later be integrated inside a better theory of general equilibrium settings.»*

Désagréger les transactions économiques accroît cependant considérablement la complexité des problèmes étudiés. Ce mouvement a donc été rendu possible par un certain nombre d'hypothèses restreignant le champ de l'analyse et, parfois, sa pertinence empirique. Le pouvoir de décision quant aux modalités du contrat, en particulier, est en général supposé détenu intégralement par le principal (Mirrlees (1975) ou Grossman & Hart (1983), par exemple). Dans la conférence donnée à l'occasion de la remise de son prix Nobel (1996), Mirrlees (1997, p.132) soulignait le rôle central de cette hypothèse,

[2]Le champ d'application de cette littérature est trop large pour faire l'objet ici d'une recension exhaustive. On pourra consulter, par exemple, Macho-Stadler & Pérez-Castrillo (2001) pour une introduction aux problèmes traités et Bolton & Dewatripont (2005) pour une présentation détaillée des développements les plus récents. Laffont & Martimort (2002) proposent une synthèse alliant un traitement techniquement poussé à de nombreux exemples d'application.

comme la fécondité des perspectives ouvertes par sa remise en cause :

> *« the account given here has been one-sided, for the principal always set the terms of the contract, and the agent took the action. It is true that many economic relationships are one-sided [. . .]. Many others are not, and involve cooperative arrangements or bargain between people in similar situations. It is not so much the asymmetry of information that is special about principal–agent relationships, but the asymmetries of responsibilities, with the principal moving first, the agent following. We have made some progress [. . . and] now we can better appreciate that anonymous market relationships are only a part of economic reality : perhaps not even the largest part. Most economic problems and possibilities involve instead relationships between and among agents, whether taxes, contracts and bargain, fights and thefts, learning and search. It is a world still only imperfectly explored.»*

Conformément à cette recommandation, plusieurs travaux récents ont exploré le champ de recherche qui s'ouvre en l'absence de cette hypothèse. Un premier développement naturel consiste à étudier la robustesse des propriétés du contrat à des hypothèses alternatives quant au pouvoir de négociation des parties.[3] Au delà même de la distribution du pouvoir entre le principal et l'agent, pourtant, plusieurs hypothèses importantes du modèle peuvent être considérées comme une fondation indirecte du contrôle du principal sur les modalités du contrat.

Développements récents : les limites du pouvoir du principal

Comme le souligne Sappington (1991), la théorie des incitations s'est en effet développée en imposant, d'une part, une hypothèse forte de rationalité du comportement de l'agent et en considérant, d'autre part, des contrats parfaitement isolés et indépendants

[3]Les travaux consacrés explicitement au rôle joué par la distribution du pouvoir de négociation entre les joueurs ont permis d'en montrer l'importance dans la résolution des problèmes de sélection adverse (Quiggin & Chambers, 2003) comme d'aléa moral (Pitchford, 1998 ; Mookherjee & Ray , 2002 ; Demougin & Helm, 2005).

des autres. Etudier le rôle de ces hypothèses revient à mettre en question le pouvoir du principal sur les modalités et le fonctionnement du contrat.

La première hypothèse impose que l'ensemble des déterminants du comportement se trouvent inclus dans les modalités du contrat. Ce cadre d'analyse élimine par conséquent les déterminants informels du comportement, qui entrent en interaction avec les instruments de la politique d'incitation. Par définition, ils échappent au contrôle du principal et en limitent donc le pouvoir d'influence. En ce sens, les travaux destinés à enrichir la description du comportement de l'agent apportent un éclairage complémentaire à celui de la théorie des contrats incomplets, qui met en lumière l'impossibilité d'inclure dans le contrat la liste exhaustive des contingences réalisables (Tirole, 1999).

Un certain nombre de travaux empiriques ont effet remis en cause la capacité du contrat conçu par le principal à résoudre les problèmes posés par les situations d'aléa moral (Keser & Willinger, 2000) comme de sélection adverse (Cabrales & Charness, 2003). L'accent est mis, en particulier, sur les effets possiblement contre-productifs des instruments contractuels traditionnellement utilisés que sont les incitations (Frey & Oberholzer-Gee, 1997 ; Gneezy & Rustichini, 2000 ; Fehr & Gächter, 2002) et le contrôle (Frey, 1993 ; Dickinson & Villeval, 2004 ; Falk & Kosfeld, 2004). Suite à cette impulsion initiale, qui se rattache assez largement aux travaux d'analyse économique du comportement (*Behavioral Economics*, voir par exemple Mullainathan & Thaler, 2001) un vaste courant de travaux récents s'est alors efforcé d'élargir le champ de l'analyse aux déterminants du comportement qui échappent au contrat. Sont alors mis en évidence, par exemple, l'influence de l'identité (Akerlof & Kranton, 2005) et des normes éthiques de l'agent (Stevens & Thevaranjan, 2005) sur les propriétés du contrat ; ou encore la mesure dans laquelle la motivation intrinsèque peut s'opposer à l'efficacité des incitations extrinsèques (Kreps, 1997), notamment si elle se trouve découragée par l'information que révèlent les motivations extrinsèques (Benabou & Tirole (2003) ou par l'inadéquation des missions confiées à l'agent (Besley & Ghatak, 2005).

Une seconde limite importante provient de l'existence simultanée de contrats différents, pouvant engager l'agent ou le principal envers d'autres agents économiques, indépendants du contrat qui les lie. Cette intervention d'agents externes à la relation principal–agent se traduit alors par une interdépendance entre contrats concurrents – ou par la dépendance d'un même contrat sur la volonté de plusieurs parties – dont le principal n'a pas toujours la maîtrise. Intégrer ces interdépendances apparaît pourtant comme un prolongement naturel de l'analyse, permettant de franchir autant de pas supplémentaires en direction d'une intégration des résultats de la théorie de l'agence dans la perspective plus large du fonctionnement général d'une économie («*The incentive litterature [. . .] focuses on isolated, independent agency relationships, which precludes a complete understanding of complex organizations like firms and government.*»Sappington, 1991 p.63).

Les travaux théoriques consacrés à cette question s'intéressent aux propriétés de contrats engageant plus de deux joueurs. Il s'agit alors de décliner les résultats de la théorie de l'agence aux situations de contrats multipartites, pour une configuration déterminée des relations qu'entretiennent les joueurs et de leur position relative. Les modèles de collusion s'intéressent ainsi à l'efficacité des incitations dans les relations de nature hiérarchique, où les parties au contrat jouent simultanément le rôle de principal et d'agent (voir les contributions théoriques de Tirole (1986 ; 1988 ; 1992) et Barankay, Bandiera & Rasul (2005) pour une analyse empirique). Les travaux consacrés aux problèmes d'agence commune s'intéressent quant à eux à l'existence de plusieurs principaux se disputant les services d'un même agent (Bernheim & Whinston (1986) et Kirchsteiger & Prat, 1999).[4] C'est au contraire l'existence de plusieurs agents, dont l'action est coordonnée par un même principal, que l'analyse du travail en équipe permet de prendre en compte (Holmstrom (1982) et Meidinger, Rullière & Villeval, 2003). Bien qu'ils soient originellement consacrés à la répartition de l'activité de l'agent entre plusieurs tâches, les modèles multitâches (Holmstrom & Milgrom, 1991), enfin, permettent de prendre en compte la relation de l'agent avec plusieurs utilisateurs destinataires de ses décisions

[4]Ces travaux feront l'objet d'une présentation plus approfondie dans le Chapitre 1.

(Crifo & Rullière (2004) et Fehr & Schmidt (2004) pour une analyse empirique des développements originels de Holmstrom & Milgrom).

S'ils constituent un élargissement du champ de l'analyse aux contrats conclus entre plus de deux joueurs, ces travaux ne permettent pas de prendre en compte explicitement les perturbations engendrées, sur le contrat passé entre un agent et un principal, par d'autres individus affectés par ce contrat. Traiter cette question permettrait pourtant d'enrichir la description du contexte institutionnel dans lequel s'inscrivent les transactions, en intégrant l'influence des agents économiques périphériques à la relation contractuelle. Cet aspect a par ailleurs fait l'objet de vives critiques, qui ouvrent autant de questions de recherche importantes :[5]

> *« Conflicts between the interests of the agents and those of the principal are the least of the agent's problems. The real problem is that the agent is most likely serving many masters, many of them with conflicting interests. Even if the agent is able to silence his or her own interests, there is the matter of how to maneuver through the tangled loyalties he or she owes to many different principals and how to negotiate through their competing interests and sometimes irreconcilable differences. [...] Only the rare agent has the luxury of aligning her interests with a single principal. Conflict of interest is hardly about shirking or opportunism with guile ; it is about wrenching choices among the legitimate interests of multiple principals by agents who cannot extricate themselves from acting for so many.*
>
> Shapiro (2005, p.278)

S'inscrivant dans cet effort destiné à élargir le champ d'analyse couvert par la théorie de l'agence, les essais présentés ici se proposent d'évaluer l'influence de l'intervention d'une tierce partie sur les propriétés et l'efficacité du contrat qui lie l'agent et le principal. Cette thématique très générale recouvre une déclinaison infinie de réalités, qui se distinguent tant par la nature du contrat – et des agents économiques qu'il engage

[5] Issues, en l'occurence, de la sociologie.

– que par la position occupée par le troisième joueur.[6] La question sera donc traitée à travers trois applications types, que sont le comportement de corruption, les choix de pratique des médecins spécialistes et la demande de travail au noir.

Méthodologie

Comme le souligne Jensen (1983), les développements originels de la théorie de l'agence ont d'abord donné naissance à deux branches se distinguant principalement par l'importance accordée à la confrontation aux faits. Tandis qu'une première branche, appelée *théorie principal–agent*, se caractérisait selon l'auteur par un niveau de formalisme élevé mais négligeant la pertinence empirique des résultats ; la seconde, *théorie positive de l'agence*, mettait ainsi un outillage mathématiquement plus simple au service de la compréhension de problèmes pratiques et de la recherche de solutions réalisables.[7]

Bien que des différences subsistent, un double mouvement a par la suite permis de réconcilier ces approches. En réaction aux critiques de plus en plus nombreuses quant à l'absence de réfutation empirique (Butz (1999), par exemple), un important effort a d'abord été conduit pour tester les prédictions issues de la théorie principal–agent.[8] Ces résultats ont également été intégrés de plus en plus fréquemment dans des analyses tournées vers la résolution de problèmes pratiques. Un exemple frappant en est,

[6] Il convient de préciser que ce travail de thèse n'a en aucun cas prétention à formuler une théorie générale de la relation d'agence en présence d'un troisième joueur, mais, plus modestement, d'élargir la palette des contextes institutionnels ayant reçu l'éclairage de l'analyse.

[7] « *The principal–agent litterature is generally mathematical and non–empirically oriented, while the positive agency litterature is generally non–mathematical and empirically oriented* »(Jensen, 1983 p.334).

[8] Outre les travaux cités plus haut, majoritairement expérimentaux, cet effort inclus également des investigations à partir de données réelles portant tant sur le rôle des incitations (Prendergast, 1996) que sur l'analyse de l'arbitrage entre risque et incitation. Concernant ce dernier aspect, le travail de référence est celui de Chiappori & Salanié (2000). Prendergast (2000a ; 2002a ; 2002b) contient un certain nombre de propositions destinées à réconcilier la théorie avec des résultats empiriques mitigés, dont Chiappori & Salanié (2003) et Salanié (2003) proposent un survol.

notamment, la formation d'un champ de recherche consacré à l'Economie du Personnel (*Personnel Economics*, Lazear, 1995 ; 1999), défini comme l'application de l'analyse microéconomique aux problèmes inhérents à la gestion des ressources humaines. Pour reprendre les termes de Lazear (2000b p.611), ce champ a pu émerger «*as a result of some breakthroughs, particularly those dealing with agency theory and contract theory*»et consacre donc la réconciliation des deux approches, en mettant la rigueur de l'analyse de la théorie principal–agent au service du souci empirique porté originellement par la théorie positive de l'agence.

La démarche adoptée ici poursuit cette tendance, et les travaux présentés dans cette thèse proposent **une évaluation de la pertinence empirique des résultats de la théorie de l'agence quant aux propriétés des incitations lorsque le contrat subit l'influence d'un tiers**.

Les travaux d'économie appliquée se sont récemment ouverts à un large spectre de méthodes de recueil des données – incluant des données réelles, des expériences de terrain (*field experiments*), des expériences naturelles (*natural* ou *social experiments*) et des expériences en laboratoire (*laboratory experiments*) – permettant chacune de répondre à des questions spécifiques (Harrison & List, 2005). Compte tenu de la nature des problèmes abordés, deux méthodes différentes de collecte des données seront mobilisées dans ce travail.

D'une part, deux des applications qui retiennent notre attention concernent les motivations sous-jacentes à l'adoption de comportements illégaux (corruption et travail au noir). L'analyse empirique se heurte alors aux difficultés d'observation inhérentes à ce type d'activité (Levitt & Miles, 2005). Elle souffre en particulier du risque de biais de sélection encouru en se limitant aux activités découvertes et rend donc difficile l'observation de l'influence des variations exogènes de l'environnement sur le comportement illégal. Chacun de ces thèmes se prête par conséquent plus facilement à un traitement d'économie expérimentale, dont l'application à l'économie du crime a déja permis d'im-

Tableau 1 – Structure de l'analyse

		Principal	*Agent*	*Tiers*	*Structure d'intérêts*
I	Corruption	Délégant	Délégué	Corrupteur	Divergents
II	Offre de soins	Etat	Médecin	Patient	Convergents mais contradictoires
III	Travail au noir et concurrence	Employeur	Employé	Concurrent	Divergents avec mécanisme de réconciliation

portants progrès (Camerer & Talley, 2005). La troisième application est d'autre part consacrée à la pratique médicale et, plus particulièrement, à la qualité des soins délivrés. Dans ce cadre, nous nous intéressons moins aux réactions d'un agent représentatif aux variations des incitations qu'aux choix d'une catégorie professionnelle clairement identifiée (les médecins) face à un environnement institutionnel donné. Ce troisième travail est donc mené en utilisant les choix observés contenus dans des données réelles et exploitent une expérience naturelle.

Au delà de la méthodologie adoptée, les applications retenues font en outre échos à un certain nombre de débats contemporains de politique économique, et apporteront en particulier des éléments de réponse aux questions suivantes :

- *Quelle est l'influence des salaires sur le comportement de corruption ?*
- *La rémunération des médecins peut-elle permettre d'accroître la qualité des soins tout en contenant les coûts du système de santé ?*
- *La dénonciation par une firme concurrente peut-elle lutter efficacement contre l'usage du travail au noir par les entreprises ?*

La structure d'analyse adoptée pour y répondre est résumée dans le Tableau 1.

Intervention d'une tierce partie : applications

A travers une analyse des déterminants du comportement de corruption, le **Premier Chapitre** s'intéresse au comportement de l'agent lorsqu'une même décision est gouvernée par plusieurs contrats concurrents. Les situations de corruption correspondent en effet au détournement d'un pouvoir discrétionnaire au bénéfice d'un tiers qui en offre rétribution. Au regard du contrat de délégation, qui lie le principal et l'agent, cette tierce partie – appelée corrupteur – constitue donc un troisième joueur dont les intérêts divergent de ceux du principal. Le corrupteur tente alors d'influencer le comportement de l'agent, pour en obtenir la décision attendue, en greffant sur le contrat de délégation un second contrat, appelé pacte de corruption.

Nous proposons dans un premier temps une synthèse de la littérature récente consacrée à cette question, mettant en parallèle les résultats empiriques et théoriques quant aux déterminants du comportement de corruption et les propriétés de chacun de ces deux contrats. Dans un deuxième temps, nous étudions l'influence de leur imbrication sur le comportement adopté par l'agent. Pour ce faire, nous proposons un jeu de corruption à trois joueurs qui exploite la position particulière de l'agent, à l'intersection de deux accords contradictoires entre eux. Plus particulièrement, nous montrons que l'agent fait face à un conflit de réciprocités lorsque le Principal opte pour un salaire d'efficience. Cet *effet de délégation* devrait tendre à diminuer le niveau de corruption, et conduirait donc à surestimer l'incitation à être corrompu dans les jeux à deux joueurs. Nous construisons deux expériences qui se distinguent par l'exogénéité de la relation de délégation.

Les résultats expérimentaux confirment l'influence de l'effet de délégation sur la décision de participer et de se conformer à un contrat de corruption. Ce mécanisme fournit en outre une explication de l'influence du niveau du salaire sur le comportement de corruption. Si la corruption des individus participant à l'expérience s'avère en effet croissante du salaire lorsque les conditions de délégation sont exogènes, l'existence d'un

principal décidant de ce niveau de salaire tend non seulement à rompre cette relation positive, mais à la renverser.

Le comportement de pratique des médecins, auquel est consacré le **Deuxième Chapitre**, fournit l'exemple d'une relation de délégation influencée par une tierce partie dont les intérêts, bien que contradictoires, restent convergents avec ceux du principal. La politique de santé, décidée par les autorités administratives, est en effet destinée à contrôler la relation entre l'offre et la demande de soins. Pour les autorités, la maîtrise des coûts de santé est une préoccupation importante et guide en partie la conception des contrats de rémunération des médecins, en charge de l'offre de soins. Pourtant, dans un système administré, où les frais afférents aux soins de santé sont assez largement pris en charge par la socialisation des risques, les exigences des patients sont principalement orientées vers la qualité des soins de santé. Sont ainsi valorisés le temps consacré à la réalisation des actes médicaux, la consultation de spécialistes du problème soigné, etc. La conception des contrats de rémunération qui contrôlent l'offre de soins doit donc tenir compte de ce double objectif, et promouvoir la santé tout en assurant l'efficacité de l'offre de soins.

Afin d'évaluer la capacité des incitations à répondre à cette double exigence, nous proposons d'abord une analyse théorique du rôle de la rémunération dans l'arbitrage réalisé par les médecins entre quantité et qualité des soins. Cette analyse met en particulier en évidence les ambiguïtés qu'engendre la prise en compte de ces dimensions, et isole des conditions suffisantes sous lesquelles la rémunération peut permettre d'améliorer la qualité. Ces indéterminations sont levées par l'analyse empirique, consistant à estimer un modèle structurel des choix de pratique des médecins en termes de marges extensives (quantité de travail) comme de marge intensive (temps consacré aux actes). Le modèle est estimé grâce à une base de données originale, qui contient des informations à la fois sur l'offre de soins et les revenus de tous les médecins spécialistes exerçant au Québec entre 1996 et 2002. Cette période recouvre l'introduction d'une réforme majeure des modes de rémunération des médecins du Québec. Alors que la rémunération à l'acte est

très largement majoritaire jusqu'en 1999, le gouvernement du Québec offre à cette date la possibilité d'adopter un mode de rémunération mixte (associant une rémunération à l'acte réduite à un *per diem*). Les préférences des médecins spécialistes sont décrites par une fonction d'utilité directe de forme Translog et les comportements de pratique sont modélisés comme des choix discrets.

Les résultats économétriques montrent que le passage à la rémunération mixte accroît le temps consacré à chaque acte et encourage la diversification des activités professionnelles des médecins, mais au prix d'un accroissement important du coût. Les paramètres estimés sont également utilisés pour simuler l'effet de réformes alternatives, telle qu'une rémunération mixte obligatoire. Les résultats mettent en évidence l'importance de la liberté de choix, qui engendre une auto-sélection des médecins fondée sur leurs préférences à l'égard des choix de pratique.

Dans le cadre de la demande de travail au noir, enfin, le **Troisième Chapitre** analyse les propriétés de l'interaction entre des principaux aux intérêts contraires mais disposant d'un mécanisme permettant de réconcilier ces divergences. Lorsqu'il recourt au travail au noir, le principal propose à l'agent un contrat illégal lui assurant les bénéfices de l'évasion fiscale. La demande de travail au noir se distingue pourtant de l'approche classique de l'économie du crime en ce que le bénéfice de l'illégalité dépend du comportement des firmes concurrentes. La dénonciation du travail au noir, fréquemment évoquée pour lutter contre la fraude, apparaît alors comme un instrument permettant aux principaux de mettre en œuvre un accord préservant les bénéfices de l'illégalité.

Nous proposons d'abord une analyse théorique de la demande de travail au noir prenant en compte l'influence de la concurrence. Nous considérons pour ce faire une industrie en situation d'oligopole où les producteurs se font concurrence en prix de façon répétée à horizon infini. Cette première étape de l'analyse conduit à la *malédiction de Bertrand* : lorsque les firmes se livrent une concurrence en prix, l'évasion devient le seul équilibre du marché, alors même que la guerre des prix qui en résulte annule le

bénéfice de l'évasion. Introduisant la possibilité pour les firmes de dénoncer le travail au noir de leur concurrentes, nous montrons ensuite que la dénonciation constitue une menace crédible contre les baisses de prix. Elle peut alors permettre de mettre en œuvre un état d'*évasion collusive*, dans lequel le prix est durablement maintenu à un niveau garantissant des profits positifs. Ces prédictions théoriques sont testées par des traitements expérimentaux faisant varier tant la taille du marché que le coût de la dénonciation. Les comportements observés confirment la pertinence empirique de la malédiction de Bertrand ainsi que les conditions de mise en œuvre de l'évasion collusive.

Ces résultats établissent donc, en particulier, le caractère fortement contre-productif de la dénonciation dans la lutte contre le travail au noir. L'instauration de ce type d'instrument fournirait en réalité aux fraudeurs un instrument puissant d'apaisement de la concurrence, accroissant ainsi la rentabilité de la fraude fiscale. Les clauses de clémence, qui consistent à exonérer des sanctions qu'il encourt un fraudeur dénonçant l'évasion d'un tiers, sont de nature à faciliter l'usage de la dénonciation. Pour cette raison, on ne peut donc attendre de leur mise en pratique qu'une aggravation de l'emprise du travail au noir.

A travers ces applications, il apparaît donc que l'analyse de l'effet des incitations en présence d'un troisième joueur dépend de façon cruciale de la structure d'intérêts qu'entretiennent les joueurs en présence. Le Chapitre 1, consacré aux déterminants du comportement de corruption, étudie d'abord le comportement de l'agent lorsqu'un second contrat, divergent du premier, lui est proposé par un corrupteur. Le Chapitre 2, ensuite, se limite à l'analyse des propriétés d'un contrat unique – régissant la rémunération des médecins – mais subissant l'influence contradictoire de deux principaux. Enfin, à travers le rôle de la dénonciation dans la demande de travail au noir, le Chapitre 3 s'intéresse au contrat choisi par le principal lorsque le bénéfice qui en est tiré dépend du comportement d'autres principaux agissant sur le même marché.

Chapitre 1

Analyses du comportement de corruption

«[. . .] The fundamental puzzle is how to create state and market institutions that are reliable and trustworthy at the same time as interpersonal relations based on mutual trust (or distrust) are kept from undermining these reform efforts. »

S. Rose-Ackerman (2001, p.27)

La délégation constitue un mécanisme puissant de coordination, assurant la fluidité des transactions. Le fonctionnement d'une économie repose en effet sur deux piliers : la régulation par le marché et la constitution d'organisations, dans lesquelles la délégation se substitue aux prix pour assurer la coordination des échanges (*«[. . .] While economists treat the price mechanism as a co-ordinating instrument, they also admit the co-ordinating function of the "entrepreneur" »* R.H. Coase, 1937 p.389). En ce sens, la notion de corruption est intimement liée à l'activité économique. Selon un assez large consensus parmi les économistes, la caractéristique centrale d'une situation de corruption est en effet d'organiser le détournement d'un pouvoir discrétionnaire au bénéfice

d'un tiers, qui en offre rétribution.[1] En ce sens, la corruption constitue un défi supplémentaire lancé à la théorie des incitations. Leur objectif est de réconcilier les intérêts divergents du délégué et du délégant par l'intermédiaire des dispositions établies par le contrat de délégation.[2] Or la corruption greffe un second accord, sur ce premier contrat, dont l'objectif est d'instaurer un motif additionnel de divergence grâce au versement d'un "pot-de-vin" (*bribe*). Cet accord illégal, le "pacte de corruption", créé donc de nouvelles incitations, orientées vers le détournement du pouvoir discrétionnaire.

Le propos de ce chapitre est de montrer que l'analyse micro-économique de la corruption repose sur les propriétés de ces accords et les caractéristiques des joueurs qui y participent. Les situations que recouvre le concept de corruption ont en effet en commun une structure particulière d'interactions. En conséquence, nous retenons une définition "contractuelle" de la corruption, en nous appuyant sur la description proposée par Banfield (1975, p. 587) :

> *« [. . .] An* ***agent*** *serves (or fails to serve) the* ***interest*** *of a* ***principal****. The agent is a person who has accepted an obligation (as in an employment contract) to act on behalf of his principal in some range of matters and, in doing so, to serve the principal's interest as if it were his own. The principal may be a person or an entity such as an organization or public. In acting in behalf of his principal an agent must exercise some* ***discretion****; the wider the range (measured in terms of effects on the principal's interest) among which he may choose, the broader is his discretion. The situation includes* ***third parties*** *(persons or abstract entities)*

[1] L'accent est mis sur cet aspect dès les premiers développements de l'analyse de la corruption (Rose-Ackerman, 1975 ; 1978 ; par exemple). Le champ des relations auxquelles cette propriété s'applique a cependant suscité d'intenses débats. Le principal motif de division tient à la question de savoir si la définition économique de la corruption doit ou non se conformer à la tradition juridique, et être circonscrite aux détournements des seuls pouvoirs publics. Goudie & Stasavage (1998) présentent les différentes définitions retenues dans la littérature. Dans ce travail, la corruption est définie exclusivement par les relations économiques qu'entretiennent les individus, sans distinction sectorielle.

[2] « *Conflicting objectives and decentralized information are the two basic ingredients of incentive theory.* », Laffont & Martimort (2002), p. 2.

> *who stand to gain or lose by the action of the agent. There are* ***rules*** *(both laws and generally accepted standards of right conduct) violation of which entails some probability of a penalty (cost) being imposed upon the violator. A rule may be more or less indefinite (vague, ambiguous or both), and there is more or less uncertainty as to whether it will be enforced. An agent is personally* ***corrupt*** *if he knowingly sacrifices his principal's interest to his own, that is, if he betrays his trust.»*[3]

Les mécanismes à l'œuvre peuvent être illustrés par l'exemple, abondamment utilisé dans la littérature (Manion, 1996 ; 1998 ; Yava, 1998 ; Cadot, 1987), de l'attribution de permis de production par un agent public. Dans ce type de situation, un fonctionnaire est chargé par l'État de choisir les entreprises qui se verront autorisées à entrer dans un secteur d'activité réglementé. L'État souhaite que les firmes qui se voient attribuer un permis respectent les réglementations en vigueur, et laisse le soin à l'agent d'apprécier les aptitudes des firmes candidates. Pour chacune d'entre elles, le profit dépend exclusivement de l'attribution d'un permis : il ne saurait être positif sans que la firme se trouve en position de produire. Le principal (l'Etat) et le corrupteur (toute firme candidate qui ne respecte pas les critères d'attribution des permis) ont donc des intérêts opposés. Si l'agent n'a pas de préférences quant à l'identité de l'entreprise sélectionnée, il peut accepter un pot-de-vin et choisir en échange l'entreprise qui le lui a versé. A l'inverse, le système de délégation choisi par l'Etat peut le porter à une conscience professionnelle suffisamment forte pour refuser toute relation de corruption et choisir les firmes qui méritent un permis. Ces deux situations sont à l'évidence incompatibles et l'agent, s'il a accepté un pot-de-vin, devra trahir la confiance ou du principal ou du corrupteur.

Comprise comme l'imbrication de deux accords aux motifs contradictoires, la corruption tire ses spécificités des incitations divergentes qu'elle instaure. La démarche adoptée ici pour en exposer les conséquences comprendra deux temps. Dans un premier temps, nous essaierons de montrer comment l'analyse des relations bipartites a permis à l'analyse microéconomique de comprendre les conditions d'émergence et de mise en

[3] Les emphases sont de l'auteur.

œuvre de la corruption. Comme le suggère la définition que nous avons adopté, l'existence d'un contrat de délégation est une condition préalable à son émergence. La Section *(i)* est consacrée à décrire les propriétés du contrat de délégation et les déterminants du comportement de corruption qui en résultent. La possibilité que s'instaure une relation de corruption repose sur l'existence d'un corrupteur. La position qu'il occupe vis-à-vis du principal et, en particulier, le conflit d'intérêt qui les oppose permet de raffiner la définition contractuelle de la corruption et de comprendre les motivations du corrupteur (Section *(ii)*). Elles guident son comportement qui, conjointement à celui de l'agent, détermine les propriétés du pacte de corruption et leurs conséquences sur sa mise en œuvre effective (Section *(iii)*). Cette revue de la littérature récente[4] fera apparaître, en particulier, que les deux accords, quoique radicalement différents, peuvent être amenés à mobiliser des mécanismes identiques au service d'objectifs divergents (Section *(iv)*). Dans un second temps, nous proposerons une analyse expérimentale qui exploite cette propriété et met en évidence l'influence de la nature de la relation instaurée entre le principal et l'agent sur les propriétés du pacte de corruption.

(i) Contrat de délégation : la relation principal – agent

L'élément de base pour que s'instaure une relation de corruption est l'existence d'un contrat de délégation, assorti de marges discrétionnaires. Ce contrat est conclu entre un principal (délégant) et un agent (délégué). Nous décrivons dans cette section les déterminants du comportements de corruption qui peuvent être déduits des propriétés de ce contrat de délégation.

[4]L'engouement pour l'analyse de la corruption a justifié récemment la publication d'un nombre important de revues de la littérature . La synthèse proposée ici se démarque des précédentes en se concentrant sur l'ensemble des relations de corruption – plutôt que les seules relations qui impliquent un fonctionnaire (au contraire de Jain, 2000 et Tanzi, 1998) – afin de couvrir les développements les plus récents de l'analyse micro-économique – plutôt que l'influence de la corruption sur l'équilibre macro-économique et le développement (Bardhan, 1997) – ainsi que la volonté de confronter les résultats théoriques aux faits (Aidt, 2003).

Si l'usage du pouvoir discrétionnaire était parfaitement observable, toute tentative de corruption (*i.e.* détournement du pouvoir délégué à des fins contraires aux intérêts du principal) serait immédiatement détectée. Nous supposons donc que le principal et l'agent se trouvent en situation d'aléa moral. Ce type de situation contractuelle a fait l'objet de très nombreuses analyses[5] qui montrent, en particulier, qu'il existe un schéma de rémunérations contingentes capable de résoudre les problèmes liés à l'asymétrie d'information. Si le principal dispose d'une mesure vérifiable de l'usage du pouvoir discrétionnaire, le contrat qui organise sa délégation peut donc contraindre l'agent à servir les intérêts du principal.

Comme le souligne Prendergast (2000b), l'existence de mesures de performance vérifiables est cependant d'autant moins probable que le pouvoir délégué comporte d'importantes marges discrétionnaires.[6] Parallèlement, l'analyse économique de la corruption a, dès ses premiers développements, identifié l'existence de marges discrétionnaires comme l'une des conditions fondamentales permettant l'émergence de la corruption. En conséquence de ces résultats, l'analyse économique de la corruption s'est concentrée sur les relations d'agence dans lesquelles le système de rémunération échoue à réconcilier les intérêts respectifs du principal et de l'agent.

En l'absence d'instruments d'incitation, c'est vers le comportement de l'agent et donc les motivations à adopter un comportement illégal que se tourne l'analyse. En suivant la tradition initiée par Becker (1968), on considère que l'agent adopte un com-

[5] On pourra consulter, par exemple, la synthèse proposée par Laffont & Martimort (2002).

[6] Pour reprendre l'un des exemples considérés par l'auteur, il est ainsi difficile d'identifier la mesure de performance à utiliser pour un fonctionnaire en charge de l'attribution des passeports. Dans le cadre de cette délégation, le principal (l'Etat) souhaite que les passeports soient attribués aux immigrants respectant les critères définis par la loi, et refusés dans le cas contraire. Le pouvoir discrétionnaire de l'agent (le fonctionnaire) consiste alors à évaluer l'adéquation des candidatures à ces critères. Pour encourager l'agent à un choix pertinent au regard de cette mission, faut-il récompenser l'attribution des passeports ou, au contraire, le nombre de refus ? Répondre à la question nécessiterait un classement ordinal des performances – refus, attribution – que l'existence de marges discrétionnaires rend impossible. Tirole (1994) développe et formalise également des arguments qui vont dans ce sens.

portement illégal dès lors que la valeur de l'illégalité domine celle de l'honnêteté. La valeur de l'honnêteté est déterminée non seulement par le salaire reçu du principal, w, mais également par la préférence de l'agent pour l'honnêteté, θ, reflétant l'ensemble des bénéfices non monétaires associés à un comportement légal (bonne conscience, estime de l'entourage, etc. ...).[7] Cette caractéristique est une information privée de l'agent et le principal n'en connaît, en conséquence, que la distribution au sein de la population, de densité $g(\theta)$ et de fonction de répartition $G(\theta)$.

En matière de corruption, le bénéfice de l'illégalité consiste pour l'agent à recevoir un pot-de-vin, noté b. S'il est impuissant à contrôler le comportement de l'agent par des incitations salariales, le principal peut en revanche mettre en œuvre un mécanisme de surveillance, par lequel l'illégalité lui est révélée avec une probabilité p. L'agent subit dans ce cas le coût de la sanction qui est, dans la version la plus simple du modèle, assimilée à un renvoi définitif. Le cas échéant, l'agent perd donc le salaire et obtient son salaire externe, w_0.

Ces hypothèses correspondent au cadre adopté dans le modèle fondateur de Becker & Stigler (1974) et permettent de mettre en évidence les déterminants essentiels de la décision de l'agent. Un agent décide en effet d'être corrompu si la valeur de la corruption excède celle de l'honnêteté, c'est à dire si son type θ est tel que $\theta + w < (1-p)\ (w + b) + p\ w_0$, soit :

$$\theta + p\ (w - w_0) < (1-p)\ b \tag{1.1}$$

Etant donnée la densité des types au sein de la population d'agents, cette condition définit la proportion d'agents corrompus, y, comme une fonction des décisions du principal : $y = P[\theta < (1-p)\ b - p\ (w - w_0)] = G[\theta^*]$, où θ^* désigne le niveau de

[7] Le fait que la corruption soit un acte intrinsèquement immoral peut paraître discutable. Comme le souligne Bardhan (2005, p.2) : *[...] «if you bribe a police officer for not torturing a suspect, that kind of corruption has been justified by some people as not immoral»*. La littérature a cependant très largement conservé l'assimilation de la corruption à une activité moralement condamnable ($\theta > 0$).

préférences pour l'honnêteté à partir duquel les agents renoncent à être corrompus, $\theta^* = (1-p)\ b - p\ (w - w_0)$. Les résultats présentés dans cette section exploitent la relation de délégation entre le principal et l'agent, et considèrent donc un niveau de pot-de-vin exogène[8]. Sous cette hypothèse, la statique comparative de la proportion d'agents corrompus permet de mettre en évidence le rôle des instruments de lutte contre la corruption. La diffusion de la corruption au sein de l'organisation apparaît en effet décroissante de la probabilité de détection comme du salaire relatif : $y = y[\underset{(-)}{p}, \underset{(-)}{(w - w_0)}]$. Ce résultat est à l'origine d'une large littérature, analysant la capacité du contrôle (p) et des incitations ($w - w_0$) à décourager la corruption. Par ailleurs, la condition d'arbitrage (1.1) qui détermine la décision de l'agent est d'autant plus contraignante que la préférence pour l'honnêteté, θ, est forte. Une seconde tradition de recherche s'est donc intéressée aux déterminants du coût moral de la corruption.

a) Détection, le rôle de p

La probabilité de détection influence le comportement de corruption par deux canaux : $\dfrac{\partial y}{\partial p} = -g(\theta^*)\, b - g(\theta^*)\,(w - w_0)$. Le premier est un effet direct, par lequel la détection agit comme un taux d'escompte sur le pot-de-vin perçu par l'agent. Ainsi, lorsque le corrupteur verse un montant monétaire b, le bénéfice espéré qu'en retire l'agent correspond à $(1-p)\ b$ en raison du risque de découverte de la fraude. En conséquence, le bien-être de l'agent est d'autant moins amélioré par le versement d'un niveau de pot-de-vin donné que le risque de détection est élevé. Le coût de la sanction constitue un second effet, indirect, correspondant à la diminution de bien-être subie par l'agent s'il doit payer l'amende. Dans le cas où la sanction consiste en un renvoi définitif, cette sanction s'interprète comme le coût d'opportunité de la corruption, puisque l'agent perd alors l'avantage salarial offert par l'emploi $(w - w_0)$.

Chacun de ces deux effets relie négativement la propension à être corrompu et le

[8] Cette hypothèse est levée dans la Section *(iii)*.

risque de détection. Cette relation causale est confirmée empiriquement par l'analyse expérimentale proposée par Abbink, Irlenbusch & Renner (2002). Entre autres traitements, les auteurs étudient la variation de comportement engendrée par l'introduction du risque de détection. Bien que sa probabilité soit très faible (0.3%), et typiquement sous-estimée par les participants, elle conduit un nombre significatif d'entre eux à renoncer à être corrompus. Une confirmation partielle provient également d'études consacrées à des variables réputées accroitre la transparence des transactions et, par conséquent, faciliter la détection de la corruption. Ainsi, la durée d'exposition à un régime démocratique (Treisman, 2000), le niveau de libéralisme économique (Goel & Nelson, 2005), le degré de liberté de la presse (Ahrend, 2002 ; Brunetti & Weder, 2003) et l'intensité de la concurrence entre les médias (Suphachalasai, 2005) s'avèrent chacun corrélés négativement avec le niveau de corruption.

Surtout, l'efficacité de la probabilité de détection a été confirmée *a contrario* à plusieurs reprises, à travers la corrélation empirique entre l'efficacité du système juridique et le niveau de corruption dans la fonction publique. Quel que soit le niveau de répression souhaité, c'est en effet l'efficacité du pouvoir judiciaire, en charge de sa mise en œuvre, qui détermine la détection effective. Le premier constitue donc une mesure indirecte de la seconde. A cet égard, Levin & Satarov (2000) mettent en évidence l'important niveau de corruption associé au système juridique embryonnaire qui caractérise la transition économique russe. Herzfeld & Weiss (2003) proposent une analyse plus systématique de ce phénomène en combinant les données de différentes enquêtes, issues de 59 pays observés en panel. L'efficacité du système juridique est mesurée en interrogeant les personnes sondées sur la tradition de conformité aux lois ; la corruption par trois indices de perceptions. Les auteurs établissent l'existence d'une forte corrélation entre les deux variables : une réduction exogène de 10% dans l'efficacité de la mise en œuvre des lois augmente de 13% le niveau de corruption. Ils suggèrent en outre que l'efficacité de la détection tient à son effet direct plutôt qu'à l'effet indirect qui transite par le coût de la sanction. Le coût d'opportunité de la corruption, mesuré par le niveau de salaire dans la fonction publique, ne semble en effet avoir qu'un effet très mitigé sur

le niveau de corruption. Les modèles de salaire d'efficience reconnaissent pourtant à ce second effet un rôle central.

b) Salaire d'efficience, le rôle de $w - w_0$

En suivant l'analyse de Becker & Stigler (1974), la saturation de la condition (1.1) définit le salaire relatif suffisant à dissuader la corruption $(w^* - w_0)$, pour un niveau donné, θ_0, de préférence pour l'honnêteté : $w^* - w_0 = \dfrac{(1-p)}{p}\left(b - \dfrac{\theta_0}{1-p}\right)$. Le salaire d'efficience capable d'empêcher la corruption peut donc s'interpréter comme le versement d'une prime, égale à l'espérance de gain associée à la corruption, ou "tentation de la malversation" (*temptation of malfeasance*). Comme le soulignent les auteurs, cet instrument est efficace pour toute probabilité de détection, aussi faible soit-elle.

Tant que la probabilité reste strictement positive, salaire et détection apparaissent de fait comme des substituts dans la lutte contre la corruption, puisque : $\dfrac{\partial w^*}{\partial p} = -\dfrac{1}{p^2} < 0$. Ainsi, une augmentation de la probabilité de détection permet de réduire le niveau de salaire nécessaire à dissuader la corruption, et réciproquement. Sous l'hypothèse d'un continuum de types d'agent, un accroissement de salaire devrait donc, à probabilité donnée, réduire le niveau de corruption dans l'organisation.

Plutôt qu'en termes d'incitations, cette équation d'arbitrage peut également être interprétée en termes de sélection (Besley & McLaren, 1993). Si l'on suppose que seule une proportion γ de la population dont sont issus les agents se comporte selon la règle d'optimisation (1.1) – et donc qu'une proportion $1 - \gamma$ d'entre eux sont, en toutes circonstances, incorruptibles – la probabilité qu'un agent appartenant à l'organisation soit honnête devient : $\delta = (1 - \gamma) + \gamma\ (1 - y)$. Le salaire devient alors un instrument permettant de modifier la composition de l'organisation, en raison de l'hétérogénéité du salaire de réserve. Même si le salaire de réserve est supposé commun aux deux types d'agents, les agents potentiellement corrompus ont en effet un salaire de réserve

implicite inférieur à celui des agents honnêtes : la perspective d'obtenir les gains de corruption porte les agents opportunistes à accepter plus facilement l'emploi. A mesure que le salaire s'accroît – en supposant la probabilité de corruption *ex post*, y, constante – l'organisation attire alors de plus en plus d'agents irrémédiablement honnêtes, et la proportion d'agents corrompus diminue.

Outre cet effet de composition, Haque & Sahay (1996) mettent également en évidence la complémentarité entre la sélection en termes de propension à être corrompu et la capacité de salaires élevés à attirer des employés de meilleure qualité dans l'organisation. La lutte contre la corruption est alors assortie d'une externalité positive, à travers l'accroissement induit des compétences moyennes dans l'organisation. Réciproquement, la dégradation du niveau de compétence engendrée par un abaissement du salaire constitue donc un coût social indirect de la corruption. Cette conclusion doit cependant être nuancée si l'on tient compte de l'allocation des compétences entre les secteurs de l'économie (Acemoglu & Verdier, 1998).

Qu'ils s'interprètent en termes d'incitations ou de sélection, les modèles de salaire d'efficience prédisent donc une corrélation négative entre le niveau de salaire et la diffusion de la corruption au sein de l'organisation. La validation empirique de l'efficacité de cet instrument a, récemment, fait l'objet d'intenses débats (Di Tella & Schargrodsky, 2003a). Comme nous l'avons déja signalé, Herzfeld & Weiss (2003) obtiennent à cet égard des résultats très mitigés, puisque, quoique négatif, l'effet estimé du salaire est non-significatif dans la plupart des spécifications de l'équation estimée. Rauch & Evans (2000) confirment cette conclusion pessimiste, puisque les résultats de l'enquête qu'ils réalisent, auprès d'experts de 39 pays en voie de développement, échouent à trouver une corrélation entre le niveau de salaire et le niveau de corruption. Enfin, c'est également le résultat qu'obtient Treisman (2000) à partir de l'indice de perception fournit par l'organisation *Transparency International*.[9]

[9]Schargrodsky (2003) présente les méthodes de construction de l'indice et leur évolution, Lambsdorff (1999) une synthèse des résultats qui en sont issus. Voir Reinikka & Svensson (2005) pour une

En première analyse, ces résultats semblent infirmer que le coût d'opportunité de la détection puisse être à l'origine d'une relation d'efficience entre le niveau de salaire et l'inclination à être corrompu. En approfondissant l'analyse du salaire d'efficience, un certain nombre d'arguments ont cependant été avancés pour dépasser cet échec relatif de la théorie. Une première explication tient au coût associé au versement de salaires d'efficiences. En raison de la relation de substitution qu'ils entretiennent, nous avons vu que le salaire nécessaire à dissuader la corruption s'accroît à mesure que la probabilité de détection diminue. Lorsque celle-ci devient très faible, le coût salarial de la lutte contre la corruption devient donc prohibitif : $\lim_{p\to 0}(w^*-w_0) = \lim_{p\to 0}\left[\frac{(1-p)}{p}\ \left(b-\frac{\theta_0}{1-p}\right)\right] = +\infty$. Pour le principal, il devient ainsi préférable de renoncer à lutter contre la corruption dès que le surplus retiré d'un comportement honnête, S, est tel que : $S < w_0 + \frac{(1-p)}{p}\ \left(b-\frac{\theta_0}{1-p}\right)$. Selon la terminologie proposée par Besley & McLaren (1993), la meilleure stratégie peut alors consister à choisir des "salaires de capitulation" (*capitulation wages*).

L'étude menée par Di Tella & Schargrodsky (2003b) propose une seconde explication, fondée sur la complémentarité entre incitations et contrôle. Les auteurs exploitent les résultats d'une importante expérience naturelle de lutte contre la corruption dans les hôpitaux de Buenos Aires, au cours de laquelle salaire et probabilité évoluent simultanément. Ce plan de lutte contre la corruption peut être décomposé en trois phases. Avant sa mise en œuvre, en Septembre 1996, les achats de médicaments, qui sont laissés à la discrétion des hôpitaux, font l'objet de très nombreux versements de pot-de-vin de la part des laboratoires pharmaceutiques sous forme d'augmentations artificielles du prix de vente. Ces détournements ne font l'objet d'aucun contrôle. La réforme combine contrôles systématiques et accroissement du salaire des directeurs d'hôpitaux. Jusqu'en Décembre 1997, tous les hôpitaux sont en effet tenus d'informer les autorités municipales du prix payé pour les médicaments. Le seul usage fait des résultats de ces contrôles est un rapport mensuel envoyé aux hôpitaux, contenant les bornes inférieures et supérieures du prix payé pour différents médicaments, choisis pour leur homogénéité. Ce

discussion méthodologique de l'utilisation des données d'enquête dans les travaux empiriques consacrés à la corruption.

critère garantit que les variations de prix ne puissent pas être attribuées à des différences de qualité. La diffusion de l'éventail des prix payés annonce donc aux directeurs d'hôpitaux que les différences de prix peuvent être tenues pour un signal de corruption par les autorités. A partir de Mai 1997, cependant, une importante campagne de presse stigmatise l'absence de sanctions associées à ces contrôles. Au total, les trois phases de la réforme se distinguent donc par le niveau de contrôle perçu par les directeurs d'hôpitaux, puisque celui-ci peut être considéré comme absent pendant la première phase (avant Septembre 1996), parfait au cours de la deuxième et intermédiaire au cours de la troisième (après Mai 1997).

Ces variations de la probabilité de détection influencent considérablement l'effet estimé du salaire. Lorsque seul le niveau du salaire est pris en compte, les auteurs confirment les résultats des études précédentes selon lesquels salaire et niveau de corruption – mesuré ici par le prix payé – n'entretiennent aucune corrélation. Cette observation cache cependant une importante diversité entre les trois phases de la réforme. Si l'effet du salaire est différencié selon les phases, il reste en effet non significatif pendant les deux premières, mais s'avère significativement négatif pendant la troisième. L'effet estimé est important, puisque l'élasticité du niveau de corruption au salaire dépasse 0.2. Ces résultats mettent donc en évidence la complémentarité entre salaire et probabilité de détection, lorsque celle-ci atteint les bornes de son intervalle (Shapiro & Stiglitz, 1984). Le salaire d'efficience n'est en effet efficace que pour des valeurs intermédiaires de la probabilité ($p \in]\,0, 1\,[$) et perd toute capacité d'influence dès lors que la détection est absente ($p = 0$) ou parfaite ($p = 1$).

La reconnaissance de cette complémentarité entre incitation et contrôle confirme les prédictions du modèle de Becker & Stigler[10] en établissant que la relation d'efficience

[10]Formellement, la condition (1.1) devient : $\theta < b$ lorsque $p = 0$, et $\theta + (w - w_0) < 0$ si $p = 1$. Cette dernière condition constitue une contradiction pour les agents qui appartiennent à l'organisation puisque, dans ce cas, la valeur du salaire externe excède la valeur de l'emploi : $w_0 > w + \theta$. Dans un cas comme dans l'autre, le salaire est donc théoriquement non pertinent pour expliquer le comportement de corruption.

entre le salaire et le niveau de corruption est rompue dès lors que le contrôle est absent. De fait, on peut identifier au moins trois raisons à l'origine de l'absence de contrôle. Les travaux de Levin & Satarov (2000) et Herzfeld & Weiss (2003), présentés plus haut, montrent d'abord que ce problème peut résulter des défaillances du système juridique. Une deuxième cause d'échec de la volonté de contrôle provient de la possibilité que les agents chargés du contrôle soient eux-même corrompus. Cette dilution du contrôle en raison de la corruption est plus particulièrement traitée dans la Section *(ii)*. Enfin, il faut ajouter à ces causes d'échec le très grand nombre de relations économiques dans lesquelles la possibilité de la corruption est tout simplement ignorée. A titre d'exemple, les professeurs d'université ne font l'objet d'aucune surveillance quant aux faveurs obtenues des étudiants afin d'améliorer leurs résultats scolaires. Ces manipulations de la délégation de l'éducation aux enseignants se font pourtant au détriment du système éducatif, à travers notamment la dévaluation des diplômes.

L'hypothèse de salaire d'efficience fondée sur le coût d'opportunité de la corruption rencontre donc d'importantes difficultés empiriques, en raison de l'insuffisance du contrôle notamment. Cet échec relatif n'épuise pas cependant l'ensemble des mécanismes qui relient le comportement de corruption au salaire versé. Comme le soulignent Akerlof (1984) et Yellen (1984) plusieurs explications peuvent en réalité justifier qu'une relation d'efficience soit établie grâce au salaire. Son influence sur le coût moral de la corruption en est une deuxième, qui a été largement retenue dans la littérature.[11]

c) Coût moral, le rôle de θ

Suivant la tradition initiée par Akerlof (1982), un certain nombre de travaux mettent l'accent sur les motifs sociologiques à l'origine d'une relation d'efficience entre le salaire

[11] Les auteurs recensent quatre explications. Au coût d'opportunité et aux motifs sociologiques traités ci-dessous, ils ajoutent la sélection adverse (en partie abordée dans cette section) et la rotation du personnel.

et le comportement dans l'emploi, en l'absence même d'un risque de détection. Selon cette approche, la relation de délégation est le lieu d'un échange de dons et contre-dons entre le principal et l'agent. C'est alors par le biais de son coût moral que la corruption et le salaire entretiennent une relation d'efficience. Conformément à la formulation originelle de l'hypothèse de salaire juste–effort (*fair wage–effort hypothesis*), cette tradition de recherche suppose en effet une corrélation positive entre le bénéfice de l'honnêteté et le salaire relatif. Cette hypothèse consiste donc à endogénéiser le coût moral de la corruption comme une fonction croissante de $w - w^*$, notée $\theta\,(w - w^*)$. Le bénéfice de l'honnêteté est alors d'autant plus élevé (faible) que le salaire reçu est supérieur (inférieur) au salaire désiré, w^*. Pour la clarté de la présentation, nous nous concentrons ici sur la version la plus simple du modèle, dans laquelle la satisfaction dans l'emploi dépend directement de l'écart entre salaires désiré et reçu, soit : $\theta\,(w - w^*) = w - w^*$. Sous cette hypothèse, la condition (1.1) décrivant la décision d'être corrompu, devient : $(w - w^*) + p\ (w - w_0) < (1 - p)\ b$.

En l'absence de détection, $(p = 0)$ un agent renonce donc à être honnête tant que : $w - w^* < b$ ou encore $b < w^* - w$. Si le salaire versé excède le salaire désiré $(w^* - w < 0)$, cette condition constitue une contradiction ($b > 0$ par définition) et l'agent n'est jamais corrompu. Dans le cas contraire $(w^* - w > 0)$, l'agent accepte le pot-de-vin tant que celui-ci compense le différentiel entre la rémunération souhaitée et la rémunération effective. Dans cette version du salaire d'efficience, le pot-de-vin est donc utilisé par l'agent comme un complément de rémunération. Quelle que soit la probabilité effective de détection, le principal peut donc éviter que l'agent y recoure en versant un salaire au moins égal au salaire désiré. En deçà de cette limite, à l'inverse, la propension de l'agent à être corrompu est décroissante du salaire pour un niveau de pot-de-vin donné.

Pour des niveaux intermédiaires de la probabilité de détection $(p \in\,]\,0, 1\,[\,)$, les deux versions du salaire d'efficience se renforcent mutuellement : le salaire diminue l'incitation à être corrompu en proportion du coût d'opportunité comme de l'écart avec le salaire désiré. Dans cet intervalle de la politique de détection, elles sont donc équivalentes

du point de vue de l'observation. Ainsi, bien que leur étude conclue à une corrélation négative entre le salaire et le niveau de corruption (échantillon de 31 pays développés ou en voie de développement, observés en panel), Van Rijckeghem & Weder (2001) se trouvent dans l'impossibilité de discriminer entre les deux hypothèses.

Face à ces difficultés d'observation, un certain nombre d'auteurs se sont tournés vers la méthode expérimentale qui permet de tester séparément la pertinence empirique de chacun de ces mécanismes. Pour les distinguer, Schulze & Frank (2003) étudient ainsi l'impact des accroissements de salaire sur les décisions de corruption selon qu'est introduit, ou non, un risque de détection. En son absence, en effet, seule la relation d'efficience fondée sur le coût moral de la corruption – qualifiée de *satisfacing* par les auteurs – subsiste. L'expérience consiste en un problème de prise de décision individuelle et contextualisée sous la forme d'une situation d'appel d'offre. Le corrupteur (firmes candidates) est un automate, aux décisions duquel les participants, jouant le rôle d'agents (fonctionnaires en charge de l'attribution du marché), réagissent. Un salaire fixe et non contingent est versé par l'expérimentateur. Son effet sur les décisions de corruption est observé sous deux traitements. Dans le premier, seule la moitié des participants se voient offrir un salaire non nul et indépendant de leurs décisions. Un risque de détection aléatoire, assortit de l'annulation des gains, est ajouté dans le second. Dans chacun des deux traitements, le comportement de corruption se montre assez peu sensible au niveau du salaire reçu. Il est cependant d'autant plus non-significatif que le risque de détection est absent.

La spécificité du protocole utilisé rend délicate la généralisation de ces résultats. Abbink (2002) en fournit pourtant une confirmation partielle. L'expérience proposée met en présence deux catégories de joueurs : les agents, possiblement corrompus, et des salariés affectés à une tâche indépendante de la leur. Le salaire désiré des agents corruptibles (w^*) est modélisé comme une fonction du salaire offert aux employés des autres secteurs de l'économie (salaire externe), représentés par la seconde catégorie. L'hypothèse que la satisfaction est à l'origine d'une relation d'efficience est alors testée

en observant l'effet sur le comportement de corruption des variations du salaire externe. Le niveau de corruption comme le montant du pot-de-vin reçu s'avèrent indépendants du niveau de ce salaire relatif. Au total, ces travaux expérimentaux corroborent donc les conclusions de Di Tella & Schargrodsky (2003b), présentées plus haut, à l'encontre de l'hypothèse que le salaire puisse influencer le comportement de corruption en l'absence de détection.

L'ensemble des travaux présentés jusqu'à présent mettent l'accent sur les mécanismes par lesquels les incitations offertes influencent le comportement de corruption. Si ces résultats permettent de comprendre pourquoi un même individu peut, ou non, décider de se livrer à la corruption, elles laissent en revanche sans réponse la question de savoir pourquoi deux individus mis face aux mêmes incitations peuvent adopter des comportements différents. Pour y répondre, plusieurs travaux ont donc levé l'hypothèse d'homogénéité, imposée implicitement jusqu'ici, pour mettre en évidence les caractéristiques individuelles – et collectives – qui influencent la prédisposition à participer à la corruption.

Une première source importante d'hétérogénéité est la différence hommes/femmes, déjà documentée dans le cas d'autres comportements illégaux[12], en termes d'attitude face au crime. Swamy, Knack, Lee & *al.* (2001) établissent en effet que les femmes tendent non seulement à être moins impliquées dans des relations de corruption, mais également qu'elles manifestent une tolérance moindre à son égard. Ces résultats s'avèrent remarquablement robustes aux changements de spécification comme à l'addition de variables explicatives indépendantes du sexe. Ils sont d'ailleurs confirmés par divers travaux empiriques consacrés à l'attitude des femmes vis-à-vis de la corruption (Frank & Schulze, 2000 ; Rigolini, Gatti & Paternostro, 2003). Dollar, Fisman & Gatti (2001) obtiennent en outre une corrélation significative entre la représentation des femmes dans les instances politiques et le niveau de corruption national. S'il est difficile d'isoler les

[12]Voir, par exemple, Mocan & Rees (2005), Gottfredson & Hirshi (1990) et Kalb & Williams (2003).

causes d'une différence aussi marquée,[13] un certain nombre de décisions politiques en ont déjà pris acte, confiant préférentiellement à des femmes certaines missions susceptibles de donner lieu à une relation de corruption. Ainsi, Swamy & *al.* (2001) signalent, par exemple, que les autorités de Mexico comme de Lima (Pérou) ont retiré la délivrance des contraventions aux policiers pour les confier à des équipes formées uniquement de femmes. Le succès de cette entreprise n'a pas, à notre connaissance, fait l'objet d'une évaluation. La nécessité en est d'autant plus forte que des travaux récents tendent à attribuer l'effet du sexe à un problème de sélection, plutôt qu'à une causalité intrinsèque. Les résultats d'estimation obtenus par Sung (2003) confirment en effet que la corrélation observée entre la représentation féminine et le niveau de corruption transite par la propension de pays plus riches et plus développés à être simultanément moins corrompu et plus enclins à promouvoir des femmes.

Le second groupe de caractéristiques ayant retenu l'attention de la littérature concerne la culture et ses déterminants. Hauk & Saez-Marti (2002) proposent ainsi un modèle de transmission des valeurs entre générations, dans lequel la culture est partiellement endogène en raison des efforts d'éducation entrepris. Les auteurs montrent que le niveau d'éthique – déterminé par la culture et donc, indirectement, par l'éducation – qui règne dans la population influence considérablement le niveau de corruption de long terme. Les efforts d'éducation peuvent également altérer durablement la stabilité de niveaux élevés de corruption. Ces résultats tendent donc à militer en faveur de larges campagnes d'information auprès de la population. Les résultats expérimentaux obtenus par Abbink & *al.* (2002) permettent de préciser le contenu de telles campagnes. Ils montrent en effet que le comportement de corruption est indépendant de la conscience qu'ont les participants des dégâts infligés à l'économie par son développement.[14] Si l'éducation peut participer à promouvoir des valeurs morales défavorables à la corruption, les campagnes

[13] Voir Swamy & *al.* (2001, Section 5), pour une discussion détaillée des sources possibles de pareille hétérogénéité.

[14] Les effets destructeurs de la corruption sur l'activité économique (croissance, investissement) et le développement sont des faits largement documentés (Mauro, 1995 ; Bardhan, 1997). La corruption engendre de ce fait de très importants coûts économiques (Dreher & Herzfeld, 2005).

d'information autour des effets néfastes de la corruption semblent, quant à elles, destinées à un succès très mitigé. Paldam (2001) propose par ailleurs une désagrégation de l'influence de la culture en se concentrant, notamment, sur le rôle de la religion. Les zones géographiques fortement imprégnées par les religions nées après la Réforme (Protestantisme, Anglicanisme) s'avèrent moins touchées par la corruption. Ces résultats semblent robustes tant à diverses spécifications qu'à l'inclusion d'un grand nombre de variables économiques (Paldam, 2002), et sont confirmés par différentes études (Treisman, 2000 ; Serra, 2004).

(ii) La relation principal – corrupteur ?

Conditionnellement à l'existence d'un contrat de délégation, une relation de corruption pourra se nouer s'il existe un troisième joueur, appelé "corrupteur", qui est affecté par l'usage que l'agent fait de son pouvoir et dont les intérêts sont en conflit avec ceux du principal. Pour cette raison, le corrupteur souhaite établir une relation parallèle, un "pacte de corruption", par laquelle il espère obtenir une décision favorable de l'agent. Les dispositions du pacte de corruption sont destinées à faire converger les intérêts de l'agent vers les siens. En ce sens, le corrupteur se met donc dans la position d'un second principal.

Les relations qui lient un agent à plusieurs principaux ont fait l'objet de plusieurs développements dans la littérature consacrée aux incitations. La particularité de la position du corrupteur par rapport à l'agent et, surtout, au principal explique pourquoi l'analyse de la corruption s'en est peu inspiré, et permet de préciser la définition "contractuelle" de la corruption.

a) Corruption et relations multi-principaux

L'existence de deux principaux qui tentent simultanément d'influencer les décisions d'un même agent correspond au cadre de base des modèles d'agence commune (Bernheim & Whinston, 1986).[15] Cette structure d'interaction recouvre la définition de la corruption que nous avons énoncée plus haut. La corruption apparaît cependant comme un cas – très – singulier d'agence commune, pour au moins trois raisons. La première tient à l'horizon temporel dans lequel les accords sont conclus. Comme nous l'avons souligné, l'existence d'un contrat de délégation est la condition préalable de l'intervention du corrupteur : c'est par lui que le corrupteur se trouve affecté par les décisions de l'agent. Par définition, le contrat de délégation préexiste donc au pacte de corruption. C'est par conséquent avec les modèles d'agence commune séquentielle que la corruption partage le plus de propriétés. Ces versions introduisent d'importants changement dans les résultats et la méthode d'analyse (Prat & Rustichini, 1998 ; 2003). Dans le cas de la corruption, cette propriété implique en particulier que le corrupteur conçoit le pacte de corruption *conditionnellement* au contrat de délégation, tandis que le principal doit *anticiper* les conditions du pacte. Une deuxième différence en termes de structure informationnelle s'ajoute à celle-ci. Bien que la corruption pose effectivement un problème d'agence pour le principal, qui cherche à contrôler un effort inobservable, le corrupteur se trouve, quant à lui, en situation d'information parfaite. Contrairement au principal, le corrupteur ne cherche pas à contrôler les moyens mis en œuvre par l'agent (typiquement, l'effort) mais la décision finale qui sera la sienne : du point de vue du corrupteur, le comportement de l'agent n'est pertinent qu'à travers la décision binaire qui consiste

[15] De façon plus précise, ces modèles étudient les contrats d'équilibre, les paiements qui en résultent et leur optimalité, de situations dans lesquelles *«an individual (the agent) decides upon an action affecting his or her well-being as well as the well-being of n other individuals (the principals), each of them offering a menu of payments contingent on the action chosen. Specifically, the principals simultaneously offer nonnegative contingent payments to the agent, who subsequently chooses an action. The primitives of the common agency game are simply the set of feasible actions for the agent and the utilities derived by the agent and the principals for the different actions»*, Laussel & Le Breton (2001, p.94).

à lui donner satisfaction ou pas. Le corrupteur n'est donc pas confronté à un problème d'asymétrie d'information. Il doit, en revanche, résoudre les problèmes de mise en œuvre liés à l'illégalité de l'accord qu'il instaure avec l'agent. Les conséquences de cette troisième différence sur les propriétés des relations de corruption seront présentées plus en détail dans la Section *(iii)*. Il en résulte que les analyses de la corruption étudient le croisement d'un *contrat* de délégation et d'un *pacte* de corruption, tandis que les modèles d'agence commune sont consacrés à l'imbrication de contrats concurrents.

Cette imbrication d'un contrat légal et d'un pacte illégal est la caractéristique que la corruption partage avec les modèles de collusion (Tirole, 1986). Ces modèles s'intéressent en effet aux situations dans lesquelles (Tirole, 1992, p.154)

> *« A member of an organization, agent 1, uses the discretion conferred on her by the organizationnal design to help another member, agent 2. This discretion may take the form of a task allocation, the choice of compensation or penalities, or reports to superiors. Its foundation is the information held by agent 1, but not the center. In exchange for the favor, agent 2 offers a side transfer or else uses his own discretion in the organization to benefit agent 1.»*

Le détournement de pouvoir discrétionnaire et le versement de compensations parallèles (*side transfers*) sont donc autant de points communs qui s'y ajoutent et expliquent que collusion et corruption aient été rapprochées dans ces premiers travaux. Ces deux champs d'analyse ont cependant connu par la suite des développements largement autonomes, en raison notamment de l'absence de relation contractuelle entre le principal et le corrupteur.[16]Dans la mesure où ils considèrent les transactions illégales réalisées entre

[16]Aux frontières de ces deux champs, quelques travaux ont exploré la complémentarité qu'entretiennent la collusion au sein de l'organisation (ou corruption interne, Bac, 1996b) et la corruption (externe) des agents qui en font partie. Ces modèles adoptent donc une structure à quatre joueurs, dans laquelle un surveillant s'intercale entre le principal et l'agent dans la relation de délégation. L'agent corrompu devient alors à son tour corrupteur (interne) en transmettant le pot-de-vin à son supérieur hiérarchique pour en obtenir l'indulgence. Ce phénomène produit une dilution du contrôle

les membres d'une même organisation, les modèles de collusion étudient la conclusion d'un pacte entre deux ou plusieurs membres de l'organisation qui sont chacun liés par un contrat de délégation avec un principal unique. Le principal, en tant que créancier résiduel de l'organisation, souhaite minimiser l'ampleur de la collusion et s'appuie sur chacun de ces contrats de délégation. La littérature consacrée aux problèmes de collusion repose alors sur une conception qui : *«views organizationnal behavior as a game* ***among members, the rule*** *of which are* ***defined by the initial contract»***[17] (Tirole, 1988, p.464). A l'inverse, la corruption naît de l'externalité imposée au corrupteur par la relation de délégation, *contre et même en dehors de la volonté du principal*. En particulier, le corrupteur n'entretient en général aucune relation contractuelle avec le principal. Si, dans le cas de la collusion, le principal dispose d'autant d'instruments qu'il y a d'agents déviants, la corruption ne lui en laisse ainsi qu'un seul, le contrat de délégation passé avec l'agent, parce qu'il ne contrôle pas les modalités du contrat par lequel le corrupteur est affecté par les décisions de l'agent.

Au total, le corrupteur apparaît donc vis-à-vis de l'agent comme un second principal, proposant un accord illégal à la suite du contrat de délégation, pour obtenir la faveur d'une décision parfaitement observable ; mais qui n'entretient aucune autre relation avec le principal que celle qui transite par l'agent. En raison de ces propriétés, les analyses qui exploitent la relation entre le principal et le corrupteur sont essentiellement consacrées à mettre en évidence le lien entre la corruption et les autres secteurs de l'économie, à travers les activités que peut exercer le corrupteur.

au sein de l'organisation et permet de mettre en lumière des résultats originaux quant à l'efficacité du contrôle hiérarchique (Bac, 1996a ; Bag, 1997 ; Bac & Bag, 2000), du salaire et de la promotion interne (Carrillo, 2000a) pour lutter contre la corruption externe ou au choix de la meilleure structure d'organisation (Mishra, 2002).

[17] Les emphases ont été ajoutées.

b) De la corruption des lois...

Une première tradition fait suite au travail de Becker & Stigler (1974), en étudiant la capacité d'une économie – ou d'une organisation, voir Note (16) – à garantir l'application des lois (*law enforcement*) en présence de corruption. En cas de détection, les contrevenants deviennent en effet des corrupteurs potentiels, puisqu'ils peuvent éviter les sanctions qui les menacent en offrant un pot-de-vin au représentant de la loi. Le surplus attendu de la corruption correspond alors à la perte de bien-être (privé) évitée grâce à cet accord. La corruption tend, par là, à transformer les sanctions en pot-de-vin et dilue l'efficacité du contrôle (Polinsky & Shavell, 2001). Cette dilution peut même aller jusqu'à l'impossibilité de sanctionner les comportements illégaux si la probabilité de les confondre dépend de l'effort des représentants de l'ordre. Sous cette hypothèse, leur stratégie d'équilibre consiste en effet à choisir le niveau d'effort qui assure qu'un taux suffisant de crimes seront commis, de façon à préserver les bénéfices de la corruption (Marjit & Shi, 1998).

De plus, ce mécanisme renverse les résultats classiques en matière d'élaboration des lois optimales. Puisque la détection est socialement coûteuse (salaires des représentants de l'ordre, équipements, ...) le système de sanctions devrait être conçu de façon à minimiser la probabilité de détection et à infliger les sanctions maximales (Becker, 1968 ; Stigler, 1970). En présence de corruption, cependant, ces sanctions sont transformées en pot-de-vin et leur accroissement tend donc à renforcer la corruption. Cet effet ajoute un nouvel argument à l'arbitrage entre détection et sanction, d'où il résulte que la probabilité maximale n'est plus nécessairement optimale (Bowles & Garoupa, 1997 ; Chang, Lai & Yang, 2000). Par ailleurs, cet effet agit quelle que soit la nature de la sanction. Il tend donc à rendre caduque la distinction classique entre peines monétaires et non-monétaires (emprisonnement, travaux d'intérêt général, ...) puisque toutes se réduisent à des transactions monétaires dans le passage par la corruption (Garoupa & Klerman, 2004).

Lorsque le système en charge de l'application des lois en est lui-même l'objet, la corruption entretient naturellement une très forte complémentarité avec les activités illégales (Celik & Sayan, 2005). Ainsi, la corruption des surveillants qui en sont chargés peut rendre inopérant le contrôle de la pollution (Mookherjee & Png, 1995). De la même façon, lorsque les juges sont corruptibles, la corruption apparaît comme un complément stratégique du crime organisé : la corruption se développe en proportion des sanctions infligées au crime et en accentue par conséquent la propagation (Kugler, Verdier & Zenou, 2005). Un mécanisme semblable, enfin, explique que la corruption puisse encourager le développement de l'économie souterraine (Cule & Fulton, 2005 ; Choi & Thum, 2005) et permet de comprendre la forte corrélation observée empiriquement entre ces activités (Johnson, Kaufmann, McMillan & *al.*, 2000 ; Friedman, Johnson, Kaufmann & *al.*, 2000). Une importante littérature s'est en particulier attachée à étudier l'influence de la corruption sur la collecte des taxes et les revenus qui en découlent. Bien que l'instauration d'importantes marges discrétionnaires puisse jouer un rôle important de révélation de l'information (Franzoni, 2004), elles permettent dans le même temps aux contribuables d'asseoir l'évasion fiscale sur la corruption. Il se peut alors qu'un accroissement du taux de taxes conduise à une diminution des revenus fiscaux, en raison de la diffusion induite de la corruption parmi les percepteurs (Chander & Wilde, 1992 ; Toye & Moore, 1998).[18]

Partant de la collecte des taxes, une seconde tradition élargit cette perspective et étudie l'influence de la corruption sur les activités économiques où les marges discrétionnaires sont étendues. Dans le cas de la politique industrielle, Ades & Di Tella (1997) montrent ainsi que les subventions à l'investissement ont non seulement l'effet positif attendu, mais également un effet négatif dû à la mutation des aides en pot-de-vin. Plutôt que d'affecter l'efficacité de ses politiques, la corruption peut aussi peser sur les

[18] C'est également la situation qu'étudient Besley & McLaren (1993), quoique leur argument porte plus spécifiquement sur les modes de rémunération des percepteurs. Marjit, Mukherjee & Mukherjee (2000, 2003) et Saha (2003) élargissent l'analyse au cas où les percepteurs extraient des pots-de-vin en menaçant les contribuables de leur prélever des sommes indues.

finances du gouvernement. Dans les marchés attribués par appels d'offre le pot-de-vin s'ajoute ainsi aux coûts sur lesquels s'appuient les firmes pour calculer leur prix. Bien que la probabilité que la meilleure firme soit sélectionnée ne soit que faiblement affectée – même si elle devient inférieure à l'unité – (Burguet & Che, 2004), il en résulte un accroissement du prix moyen d'attribution du marché (Compte, Lambert-Mogiliansky & Verdier, 2005). Plus généralement, le détournement qu'organise la corruption est susceptible de changer l'orientation de la plupart des activités qui ont recours à une délégation de pouvoir discrétionnaire – politique environnementale (Damania, Fredriksson and List, 2003), composition des dépenses publiques (Mauro, 1998 ; Hines, 1995), ... – et d'en modifier les bénéficiaires.

Pour organiser ce détournement, le pacte de corruption se superpose au contrat de délégation. Sa conclusion repose sur la rencontre entre l'agent et le corrupteur, qui décident de ses modalités.

(iii) Pacte de corruption : la relation agent – corrupteur

Les motivations de l'agent ont été largement décrites dans la Section *(i)*. La position du corrupteur par rapport au principal qui a décidé de la délégation, présentée dans la section précédente, éclaire par ailleurs les motivations du corrupteur à instaurer une relation de corruption. Avant d'établir les propriétés du pacte de corruption qui résultent de leur rencontre, nous décrivons les déterminants du comportement du corrupteur dérivés de cette motivation.

a) Comportement du corrupteur

On suppose en général que l'agent possède un monopole sur le pouvoir discrétionnaire dont il est chargé (Klitgaard, 1988).[19] Par définition, le corrupteur se trouve en outre dans l'impossibilité d'obtenir le service attendu par les voies légales.[20] En raison de ces deux propriétés, la corruption, à travers le versement du pot-de-vin (b), est donc le seul moyen pour le corrupteur d'obtenir le bénéfice du service, noté $\pi(b)$. De plus, l'hypothèse de monopole de l'offre conduit à ce que le corrupteur assume seul les coûts – moraux, réels et monétaires, discutés ci-dessous – liés à l'instauration de la relation de corruption, notés q. Lorsqu'un corrupteur met en œuvre ces démarches, son entreprise ne pourra aboutir que s'il trouve un agent ayant décidé d'être corrompu compte tenu des incitations qui lui sont offertes.[21] Nous avons vu (Section *(i)*) que ce cas se produit avec la probabilité y. Le nombre de tentatives de soudoiement nécessaires à ce que la

[19] S'il existe plusieurs agents concurrents dans la fourniture du service, le pacte de corruption d'équilibre dépend de la structure de l'offre (collusion, concurrence monopolistique, concurrence parfaite). Cette question dépasse le champ de cette présentation qui se concentre sur les relations entre les trois joueurs au coeur de la relation de corruption. Ces configurations sont étudiées par Shleifer & Vishny (1993), qui montrent que l'effet de la structure de l'offre sur les propriétés du pacte est en tout point conforme aux résultats traditionnels de la théorie de l'organisation industrielle. Plus récemment, Sanyal (2004) compare l'efficacité de la concurrence et du contrôle dans la lutte contre la corruption.

[20] Pour reprendre les termes de R. Posner, la définition d'un corrupteur «*involve[s] giving money [...] in the hope of obtaining some favor to which the donor would not otherwise be entitled.*»(*The Becker-Posner Blog*, 3 Septembre 2005). Quoique nécessaire, cette propriété n'est pas pas suffisante et doit être complétée par les éléments de définition déjà évoqués. Elle permet cependant de distinguer la corruption d'autre transactions parallèles par lesquelles un individu verse une somme monétaire qui, certes, échappe aux transactions légales, mais ne constitue pas un détournement de pouvoir discrétionnaire (fraude fiscale liée aux pourboires, ...).

[21] Un certain nombre d'auteurs s'intéressent à la compétition entre les corrupteurs pour l'obtention du service (attribution d'une licence unique, par exemple). Dans ce cas, la victoire dans le tournoi ainsi instauré doit être ajoutée à cette condition. Cette littérature s'intéresse principalement à l'efficience allocative qui résulte du tournoi de corruption (Lui, 1985). Clark & Riis (2000) proposent une synthèse critique de ces résultats, permettant de relier l'efficience de l'allocation aux propriétés du tournoi (hétérogénéité des corrupteurs notamment).

relation de corruption soit établie est donc l'espérance d'une distribution géométrique, soit : $\frac{1}{y}$. Au total, la fonction de profit du corrupteur, Π_C, s'écrit donc :

$$\Pi_C = \pi(b) - \frac{q}{y} \tag{1.2}$$

Un corrupteur décide de s'engager dans une relation de corruption si le bénéfice attendu en est positif. Dans cette expression, la probabilité de réussite (y) et le montant du pot-de-vin (b) résultent de l'interaction entre le corrupteur et l'agent. Ils constituent les termes du pacte de corruption et feront l'objet de la prochaine section. Nous nous concentrons ici sur le coût d'initiation de la relation (q) et le profit de corruption ($\pi(.)$), qui déterminent le comportement isolé du corrupteur, et décrivent par conséquent la demande de corruption.

Le coût d'initiation de la relation regroupe l'ensemble des coûts liés à l'instauration de la relation de corruption avec l'agent. Il s'agit, d'abord, des coûts moraux subis par le corrupteur lorsqu'il s'engage dans une relation de corruption. Ce premier type de coût relie donc le comportement de soudoiement aux caractéristiques individuelles – comme le sexe, voir Section *(i)* – connues pour leur influence sur le bénéfice de l'honnêteté. Une seconde catégorie regroupe, d'autre part, l'ensemble des coûts réels et monétaires engagés pour permettre cette rencontre (multiplication des contacts, persuasion, etc.).

Toutes choses égales par ailleurs, ces coûts devraient faire diminuer la probabilité que le corrupteur s'engage dans une relation de corruption, puisque : $\frac{\partial \Pi_C}{\partial q} = -\frac{1}{y} < 0$. Bac (2001) montre en particulier que la transparence des transactions permet au corrupteur d'identifier avec plus de facilité l'agent en charge du pouvoir discrétionnaire. La réduction du coût d'instauration de la relation encourage alors la corruption. Cet effet tend à nuancer l'effet bénéfique généralement attendu de la transparence (voir Section *(i)*), puisque l'augmentation de la probabilité de détection, p, est accompagnée d'une diminution du coût q.

Qu'ils soient moraux ou réels, ces coûts sont difficilement observables. Les applications empiriques qui cherchent à en évaluer l'importance recourent donc en général à des variables instrumentales, dont le choix est inspiré par la nature de ces coûts. Bien qu'elle s'intéresse au degré d'exposition à la corruption, plutôt qu'aux contrats effectivement signés, l'étude de Mocan (2004) en constitue un exemple, qui confirme l'influence de chacune de ces deux catégories sur le comportement du corrupteur. Plus précisément, l'auteur étudie l'impact des coûts d'initiation sur la probabilité que les individus de l'échantillon se trouvent en situation de corrupteur dans leurs relations avec l'administration. Les données micro-économiques utilisées proviennent du croisement de plusieurs enquêtes, réalisées dans une trentaine de pays. Le premier type de coût – coût moral – est essentiellement pris en compte par une variable de sexe. Les coût réels et monétaires sont intégrés, quant à eux, par le biais des variables qui influencent, d'une part, la fréquence des contacts avec l'administration publique – qui constituent autant d'occasions d'évaluer les possibilités de corruption – telles que l'âge, la richesse, la participation au marché du travail et le niveau d'éducation ; et, d'autre part, la familiarité des individus avec les fonctionnaires qu'ils rencontrent, à travers la taille de la ville. Il semble que cette stratégie permette de capturer l'effet du coût, puisque l'ensemble des variables citées influencent significativement, et dans le sens attendu, la probabilité de se trouver en situation de corrupteur.

Ces résultats sont en outre corroborés par l'étude de Hunt & Laszlo (2005), qui porte sur les pots-de-vin effectivement versés aux agents publics afin d'alléger les démarches administratives. L'analyse se concentre plus spécifiquement sur le rôle de la richesse, x, qui est présumée influencer le comportement du corrupteur par deux canaux. D'une part, à l'instar de Mocan (2004), les auteurs considèrent la richesse comme un indicateur de la fréquence des contacts avec l'administration, participant par conséquent à diminuer le coût d'instauration de la relation ($\frac{\partial q}{\partial x} < 0$). Dans le contexte étudié, différents mécanismes (coût d'opportunité du temps plus élevé, utilité marginale du revenu décroissante) créent d'autre part une relation positive entre la richesse et le profit de corruption ($\frac{\partial \pi(b)}{\partial x} > 0$). La richesse augmenterait alors la propension à verser un pot-de-

vin. L'application empirique, qui s'appuie sur les données issues d'une enquête réalisée auprès de ménages péruviens, fait apparaître une élasticité-revenu du pot-de-vin proche de 0.3. Cet effet est attribuable pour moitié à chacun des deux effets décrits.

L'ensemble de ces études confirme donc l'importance du coût comme du profit espéré pour comprendre le comportement du corrupteur. Les termes d'équilibre du pacte de corruption résultent de son interaction avec l'agent, dont le comportement conditionnel au contrat de délégation a fait l'objet de la Section *(i)*.

b) Pot-de-vin d'équilibre[22]

Au niveau agrégé, la demande de corruption est proportionnelle au nombre d'individus pour lesquels le profit de corruption est positif. Une simple normalisation de cette quantité par la taille du bassin d'agents – potentiellement corrompus – permet alors de définir la demande de corruption adressée à chaque agent comme une fonction de la proportion d'agents corrompus (y) et du montant du pot-de-vin (b). D'après la définition du profit de corruption (1.2), la demande est décroissante du pot-de-vin et croissante de la proportion d'agents corrompus. Le premier constitue en effet le coût direct subit par le corrupteur lorsque l'agent accepte le contrat de corruption ($\frac{\partial \Pi_C}{\partial b} = \frac{\partial \pi(b)}{\partial b} < 0$) ; la seconde est la probabilité que les démarches visant à instaurer une relation de corruption débouchent sur la conclusion d'un pacte, et détermine par conséquent le bénéfice espéré de ces démarches ($\frac{\partial \Pi_C}{\partial y} = \frac{q}{y^2} > 0$). La demande de corruption est donc monotone dans ses arguments : $D = \underset{+}{D}(y, \underset{-}{b})$.

A l'équilibre, le pot-de-vin est choisit de façon à ce que l'offre et la demande de corruption s'égalisent : $y = D(y, b)$. Le pot-de-vin d'équilibre, b^*, qui réalise cette condition s'écrit donc comme une fonction implicite de la proportion d'agents corrompus :

[22] Cette section s'appuie sur une extension du modèle initialement développé par Andvig & Moene (1990).

$b^* = f_C(y)$. Cette fonction décrit les ajustements subis par le pot-de-vin d'équilibre lorsque les agents modifient leur décision d'être corrompus (en réaction, par exemple, à un changement dans les incitations offertes par le principal). Elle peut donc s'interpréter comme la fonction de réaction du corrupteur.

Si le pot-de-vin pouvait s'interpréter comme le prix qui réalise l'équilibre sur le marché de la corruption, chacun s'attendrait à ce qu'il soit décroissant de la proportion d'agents corrompus. Comme le montrent Andvig & Moene (1990), pourtant, rien n'assure que la meilleure réponse à une augmentation de la proportion d'agents corrompus soit de diminuer le pot-de-vin proposé. Il est par conséquent possible que la corruption soit d'autant plus rentable pour les agents qu'elle est largement répandue.

Preuve La fonction de réaction du corrupteur est définie par la condition d'équilibre du marché : $y = D(y, b^*)$. La sensibilité de b^* aux variations de y se déduit de la différentielle totale de cette expression (on note D'_x la dérivée partielle première de la fonction D par rapport à la variable x) :

$$\frac{\partial y}{\partial b} = D'_y \frac{\partial y}{\partial b} + D'_b$$

Après manipulations, on obtient : $\dfrac{\partial b}{\partial y} = \dfrac{1 - D'_y}{D'_b}$. La fonction de profit du corrupteur (1.2), nous a permis d'établir que $D'_b < 0$ et $D'_y > 0$. On a donc dans le cas général $D'_y \lesseqgtr 1$. Pour tous les marchés tels que $D'_y < 1$, la meilleure réponse à la diffusion de la corruption est alors d'augmenter le pot-de-vin : $\dfrac{\partial b^*}{\partial y} > 0$. ■

Dans ces conditions, la corruption constitue un processus auto-entretenu : une augmentation de la proportion d'agents corrompus tendrait à accroître le pot-de-vin proposé ; mais cette amélioration des termes du pacte de corruption peut, en retour, élargir le bassin des agents qui renoncent à l'honnêteté.

c) Propriétés du pacte de corruption

La condition qui détermine le comportement de l'agent, (1.1), est en effet croissante du montant du pot-de-vin. Au niveau agrégé, nous avons vu que cette expression définit la proportion d'agents corrompus – l'offre de corruption – comme une fonction des termes du contrat de délégation et du montant du pot-de-vin.[23] Cette relation décrit donc la réaction optimale des agents à une variation du pot-de-vin proposé, et peut ainsi s'interpréter comme la fonction de réaction de l'agent : $y^* = f_A(\underset{-}{p}, \underset{-}{w}, \underset{+}{b})$.

Les termes d'équilibre du pacte de corruption réalisent l'intersection des meilleures réponses. Ils sont donc décrits par le système :

$$\{b^* = f_C\left[f_A(p, w, b^*)\right]; y^* = f_A\left[p, w, f_C(y^*)\right]\}$$

Alors que les résultats présentés dans la Section *(i)* isolaient les décisions de corruption prises par l'agent, cette expression synthétise la résultante de l'interaction entre le corrupteur et l'agent. Elle apporte un éclairage nouveau sur la sensibilité des termes du pacte de corruption aux instruments d'incitation et contrôle mis en œuvre dans le cadre du contrat de délégation. L'inter-dépendance entre le niveau du pot-de-vin et la proportion d'agents corrompus crée en effet une relation indirecte entre le comportement du corrupteur et les conditions de délégation offertes à l'agent. Cet effet indirect s'ajoute à l'effet direct étudié dans la Section *(i)*. Formellement, la réaction à l'équilibre des termes du pacte de corruption aux conditions de la délégation ($\beta = p, w$) est donnée par :

$$\frac{\partial y^*}{\partial \beta} = \frac{\partial f_A}{\partial \beta} + \frac{\partial b^*}{\partial \beta}\frac{\partial f_A}{\partial b}; \frac{\partial b^*}{\partial \beta} = f'_C \frac{\partial y^*}{\partial \beta}$$

[23] La Section *(i)* est consacrée aux déterminants du comportement dérivés de la relation entre le principal et l'agent. Le pot-de-vin est une variable exogène à cette relation. Il est considéré comme une variable endogène dans la présente section, afin de prendre en compte l'interaction entre le corrupteur et l'agent.

Tableau 1.1 – Effet des instruments de lutte contre la corruption

	f'_C	$\left(1 - f'_C \frac{\partial f_A}{\partial b}\right)$	$\frac{\partial f_A}{\partial p}$	$\frac{\partial y^*}{\partial \beta}$	$\frac{\partial b^*}{\partial \beta}$
(i)	+	+	−	−	−
(ii)	+	−	−	+	+
(iii)	−	+	−	−	+

Dans cette expression, le premier terme reflète l'effet direct, c'est à dire l'ajustement lié à l'incitation à l'honnêteté pour l'agent. Le second terme correspond à l'effet indirect, issu de la variation induite du pot-de-vin. Comme le souligne Carrillo (2000a), ces deux effets peuvent jouer en sens inverse (cas (i) et (ii), Tableau 1.1) et nuancer l'efficacité communément admise des instruments de lutte contre la corruption. Dans ce cas, la tendance de la proportion d'agents corrompus à diminuer – en raison de l'accroissement de l'incitation à l'honnêteté – est compensée par l'accroissement du pot-de-vin. Sous certaines conditions, le second effet peut dominer le premier, de sorte que la proportion d'agents corrompus comme le montant du pot-de-vin deviennent croissants de la probabilité de détection et du salaire (cas (ii)).

Preuve Pour chaque instrument $\beta, \beta \in \{p, w\}$, on a :

$$\frac{\partial b^*}{\partial \beta} = f'_C \frac{\partial f_A(p, w, b^*)}{\partial \beta} \qquad \frac{\partial y^*}{\partial \beta} = \frac{\partial f_A\left[p, w, f_C(y^*)\right]}{\partial \beta}$$

$$\frac{\partial b^*}{\partial \beta} = f'_C \left[\frac{\partial f_A}{\partial b}\frac{\partial b^*}{\partial \beta} + \frac{\partial f_A}{\partial \beta}\right] \qquad \frac{\partial y^*}{\partial \beta} = \frac{\partial f_A}{\partial \beta} + \frac{\partial f_A}{\partial b} f'_C \frac{\partial y^*}{\partial \beta}$$

$$\frac{\partial b^*}{\partial \beta}\left[1 - f'_C \frac{\partial f_A}{\partial b^*}\right] = f'_C \frac{\partial f_A}{\partial w} \qquad \frac{\partial y^*}{\partial \beta}\left[1 - f'_C \frac{\partial f_A}{\partial b}\right] = \frac{\partial f_A}{\partial \beta}$$

Après manipulations, les signes peuvent alors être déduits de :

$$\frac{\partial b^*}{\partial \beta} = f'_C \frac{\partial y^*}{\partial \beta} = f'_C \frac{\dfrac{\partial f_A}{\partial \beta}}{1 - f'_C \dfrac{\partial f_A}{\partial b}} \tag{1.3}$$

La fonction $f_A()$ est monotone dans ses arguments : $\frac{\partial f_A}{\partial b} > 0, \frac{\partial f_A}{\partial p} < 0$ et $\frac{\partial f_A}{\partial w} < 0$. D'après les résultats de la section précédente, le pot-de-vin est croissant de la proportion d'agents corrompus ($f'_C > 0$) si $D'_y < 1$ et décroissant dans le cas contraire. Dans le premier cas, le numérateur de (1.3) est toujours négatif ; le dénominateur est positif (respectivement négatif) si $f'_C \frac{\partial f_A}{\partial b} < 1$ (resp.

$f'_C \frac{\partial f_A}{\partial b} > 1$). Dans le second cas, le dénominateur est toujours positif et le numérateur est positif dans l'expression qui concerne b^*, négatif pour y^*.

Ces éléments sont résumés dans le Tableau 1.1, présentant les signes déduits de (1.3) dans les différents cas envisageables. ■

Cette analyse du pacte de corruption d'équilibre suppose résolue la question de sa mise en œuvre effective (*implementation*) : lorsqu'un pacte de corruption est conclut, nous avons supposé jusqu'à présent que l'agent comme le corrupteur s'y conforment, en détournant le pouvoir qui lui est confié pour le premier ; en versant un pot-de-vin pour le second. Il convient cependant de souligner que ces stratégies ne sont individuellement rationnelles que lorsque la relation de corruption est répétée selon un horizon infini. Dans ce cas, la perspective de renouveler les bénéfices de la relation incite les joueurs à respecter leurs engagements. Pourtant, si l'horizon est trop court pour que ce mécanisme fasse effet (Rosenthal, 1981), les relations de corruption sont sujettes aux difficultés de mise en œuvre inhérentes à leur illégalité (Garoupa, 1999). Cette propriété interdit en effet le recours au système juridique pour garantir l'application des termes de l'échange tels qu'ils ont été convenus. Contrairement aux contrats légaux, chaque partie (agent et corrupteur ici) se trouve, en conséquence, dans l'impossibilité de se protéger de l'opportunisme de l'autre. Ainsi, une fois le pot-de-vin reçu par l'agent (ou, de façon équivalente, une fois le service obtenu par le corrupteur) trahir l'accord – ne pas offrir le service ou ne pas verser le pot-de-vin promis – devient une stratégie dominante, puisqu'elle permet d'éviter les risques de sanction tout en tirant tout le bénéfice de la relation (Boycko, Shleifer & Vishny, 1996).

Privée de la protection du système judiciaire, la mise en œuvre des contrats de corruption doit donc faire appel à des institutions alternatives. La littérature s'est intéressée à deux types d'institutions capables de jouer ce rôle : les premières sont des institutions parallèles, qui préexistent à la relation ; les secondes sont individuelles, produites par l'interaction entre les joueurs.

D'abord, un certain nombre de relations de corruption, et plus généralement d'activités illégales voire criminelles, se font sous la protection d'institutions qui jouent le rôle de système juridique. C'est le cas très connu de la mafia, et plus généralement des organisations criminelles.[24] On peut en effet les considérer comme des organisations collectives qui garantissent la mise en œuvre des contrats illégaux, à partir d'un principe d'action qui repose sur la violence (Konrad & Skaperdas 1997).[25] Ainsi, par exemple, la mafia sicilienne a initialement émergé pour pallier les défaillances du système public de protection des propriétaires terriens (Bandiera, 2003).[26] Prenant acte de cette origine, Anderson & Bandiera (2000) établissent que ce type d'institution peut de fait émerger spontanément à l'équilibre d'une économie en transition. A mesure du développement de l'économie, et du système de protection des droits de propriété, les organisations criminelles tendent cependant à se spécialiser dans la mise en œuvre des activités non protégées par le système juridique (Akerlof & Yellen, 1994). Elles servent alors d'institution de mise en œuvre des activités illégales, de la corruption en particulier (Tanzi, 1995). Bien que peu de travaux empiriques soient consacrés à établir entre elles une relation de causalité directe, la très large coexistence entre la participation à des activités de corruption et l'appartenance à une organisation mafieuse[27] témoigne, à tout le

[24]Un autre exemple fortement lié à ce phénomène est la pression exercée au sein d'un réseau par ses membres. La menace d'exclusion du réseau se substitue alors à la violence comme instrument de mise en œuvre (Lambert-Mogiliansky, 2002). La pression sociale en est un autre, dont Garicano, Palacios & Prendergast (2005) fournissent une illustration. Les auteurs étudient la sensibilité des arbitres des rencontres de football au lieu où se déroule la rencontre, et montrent qu'ils favorisent significativement l'équipe locale sous la pression exercée par le public.

[25]Konrad & Skaperdas (1998) analysent les problèmes de crédibilité liés l'utilisation de la violence comme instrument de mise en œuvre ; Marjit, Mukherjee & Mukherjee (2000, 2003) et Saha (2003) en proposent une application au cas particulier de la corruption, dans laquelle la violence prend la forme de "harcèlement" au sens où les percepteurs utilisent la menace d'augmentation des prélèvements pour extorquer des pots-de-vin.

[26]Le cas de la Russie en fournit également un exemple, plus contemporain. Voir à ce sujet : E. Duflo "Perspectives – Les désordres de la transition en Russie : Quand la mafia évince l'Etat", *Le Monde* (6 Septembre 1994).

[27]Au point que cette concomitance fait l'objet d'un champ de recherche spécifique, comme en atteste l'existence d'un centre d'étude consacré à ce problème : Nathanson Centre for the Study of

moins, de la facilité avec laquelle la corruption est mise en œuvre lorsqu'elle bénéficie du soutien de ce type d'institution.

Les exemples abondent, cependant, de relations de corruption qui se déroulent – et sont honorées – en dehors du recourt au crime organisé.[28] Outre la protection d'institutions collectives, la littérature a également mis l'accent, pour en expliquer l'existence, sur deux propriétés produites par l'interaction entre le corrupteur et l'agent.

Le premier mécanisme tient à la structure d'information des joueurs. Comme nous l'avons vu, le comportement de corruption repose sur un certain nombre de caractéristiques individuelles, souvent inobservables, telles que le coût moral. Le corrupteur et l'agent peuvent donc se trouver en situation d'information incomplète et la mise en œuvre de la corruption fait alors intervenir un mécanisme de réputation (Kreps, Milgrom, Roberts & *al.*, 1982). Dans ce cadre, renoncer à la trahison peut être un investissement rationnel pour un joueur si cette décision entretient la confiance que son partenaire place en lui. Pour le corrupteur comme pour l'agent, la volonté de pérenniser la relation de corruption peut alors être suffisante à garantir le respect des engagements pris (Klochko & Ordeshook, 2003).

L'investissement en réputation des individus a, en outre, une influence sur la réputation collective du groupe auquel ils appartiennent (organisation, catégorie professionnelle, ...) : les croyances du corrupteur quant à la probabilité que l'agent se conforme à ses engagements est fortement influencée par le comportement passé des membres du groupe dont il est issu. Comme le montre Tirole (1996), la réputation du groupe devient alors un bien public, au sens où elle facilite la mise en œuvre de tous les contrats de corruption instaurés par ses membres. Surtout, cette réputation collective crée une inertie qui entretient le développement de la corruption : en raison du stigma lié à l'apparte-

Organized Crime and Corruption, *York University*.

[28] Ainsi, la plupart des scandales de corruption politique que la France a connu dans les années 90 n'a été rattaché à l'intervention d'aucune mafia.

nance au groupe, les individus perdent le bénéfice de l'honnêteté indépendamment de leur comportement, y compris par conséquent lorsqu'ils refusent la corruption. Lorsque cet effet domine trop fortement celui de la réputation individuelle (Andrianova, 2001) – ou lorsque, en l'absence de réputation individuelle, il devient trop important (Tirole, 1996) – la réputation collective peut alors conduire des agents, qui dans d'autres circonstances mais face aux mêmes incitations auraient choisi l'honnêteté, à accepter la corruption. Ainsi, lorsqu'une économie est fortement ancrée dans la corruption, c'est non seulement les comportements contemporains mais également la réputation collective que les mesures de lutte contre la corruption doivent combattre (Lui, 1986). Cet argument milite donc en faveur de thérapies de choc et peut participer à expliquer le succès mitigé de nombreux plans de lutte contre la corruption dans les économies où elle est très largement répandue (Steves & Rousso, 2003).[29]

Le second mécanisme exploite les développements récents de la littérature sur les fondations psychologiques des comportements économiques (Fehr & Schmidt, 2002 ; Tirole, 2002) et, en particulier, sur le comportement de réciprocité (Fehr & Gächter, 2000b). Dans une étude consacrée au motivations de la coopération, Cooper, DeJong, Forsythe & *al.* (1996) montrent en effet que les arguments fondés sur la réputation sont insuffisants pour expliquer à eux seuls les comportements observés dans les jeux destinés à tester la coopération[30]. L'une des raisons principales en est que ce mécanisme ne peut pas prendre en compte la coopération dans les jeux où la relation est ponctuelle (*one-shot*). Comme le souligne Abbink (2004), le régime de corruption généralisée qui a souvent été dénoncé dans la presse, au cours de la présidence du CIO de Juan-Antonio Samaranche, est un exemple frappant de ce que la corruption peut être mise en œuvre y compris lorsque la relation n'est pas répétée[31]. L'observation des relations de corruption

[29] Honk-Kong est le cas le plus connu du succès de ce type de thérapies de choc (Skidmore, 1996).

[30] Dilemme du Prisonnier simultané dans le cas évoqué ici. Le *Gift Exchange Game* de Berg, Dickhaut & McCabe (1995) ou la version séquentielle du Dilemme du prisonnier sont également très fréquemment utilisés.

[31] Voir, par exemple, M. Dalloni "Le mouvement olympique ébranlé par une affaire de corruption" , *Le Monde* (15 Décembre 1998).

fait en outre ressortir une forte implication des motivations fondées sur la confiance et la réciprocité :

> *« Paradoxically, a deeply corrupt regime usually operates with a high degree of reciprocal, affect-based trust. Because bribers and bribees are operating outside the law, they need to trust each other in order to maintain their relationships. They may design schemes that minimize the possibilities of betrayal, such as making payments only when corrupt services are delivered, or that limit the costs of betrayal, such as the use of middlemen. Nevertheless, the risks that one side will betray the other can be substantial so that links based on kinship or friendship can be important ways to lower the risk. The corrupt official is an untrustworthy and dishonest agent of the public interest but a trustworthy friend and relative.»*[32]

De fait, l'existence de préférences "sociales" (*other-regarding preferences*)[33] est reconnue pour faciliter la mise en œuvre des contrats et élargir par là l'éventail des arrangements auto-exécutoires (Fehr, Gächter & Kirchsteiger, 1997). Il semble en outre que la réciprocité[34] l'emporte sur d'autres types d'explication – l'altruisme en particulier – pour comprendre les motivations sous-jacentes à la coopération (Clark & Sefton, 2001). Dans le cas de la corruption, on peut donc s'attendre à ce qu'une relation de réciprocité entre l'agent et le corrupteur facilite la mise en œuvre effective du pacte conclut. C'est ce que confirment les résultats expérimentaux de Abbink, Irlenbusch & Renner (2002). Ce mécanisme semble également robuste à l'absence de répétition de la relation. Dans un travail consacré au travail illégal plutôt qu'à la corruption, mais s'appuyant sur un jeu extrêmement proche du précédent, Abbink, Irlenbusch & Renner (2000) montrent en effet qu'un nombre significatif de participants parvient à mettre en œuvre une relation illégale, pourtant ponctuelle.

[32] Rose-Ackerman (2001), p.18.

[33] Le terme "préférences tournées vers les autres" sera également utilisé de façon équivalente.

[34] Conformément à un assez large consensus dans la littérature expérimentale, le terme "réciprocité" désigne ici ce que Rabin (1998) appelle *altruisme réciproque* (*reciprocal altruism*).

Une nouvelle expérience fondée sur le jeu de corruption corrobore ce résultat (Abbink, 2004). La comparaison entre les niveaux de corruption selon que la relation entre le corrupteur et l'agent est, ou non, répétée, met cependant en évidence que la mise en œuvre des pactes de corruption est d'autant moins probable que l'interaction est courte. Au regard de ces résultats, les mesures qui tendent à abréger la relation entre le corrupteur et l'agent (typiquement, la rotation du personnel) apparaissent donc de nature à lutter contre l'instauration de relations de corruption. Plus généralement, des efforts importants ont été consacrés récemment à étudier des mécanismes susceptibles de rompre la relation de réciprocité établie entre l'agent et le corrupteur. Encourager la dénonciation du pacte de corruption en est l'exemple le plus étudié. Outre qu'elle consiste par définition à rompre la relation de réciprocité, la dénonciation présente également l'avantage de faire porter les coûts de la détection sur les contrevenants eux mêmes (Kaplow & Shavell, 1994). L'instauration de clauses de clémence, issues du droit de la concurrence et consistant à réduire la peine infligée en cas de dénonciation, est par conséquent un instrument fréquemment évoqué pour lutter contre la corruption.[35] Cooter & Garupa (2000) formalisent cette idée dans un modèle où la dénonciation est rémunérée – au sens où elle permet au dénonciateur d'obtenir non seulement une annulation de l'amende mais également une récompense – et en confirment l'efficacité. Cette conclusion dépend cependant de façon importante du type de programme de clémence mis en place.[36]

[35] Le recourt à ce type de mesure pose cependant d'importants problèmes politiques et la République Tchèque est, à notre connaissance, le seul pays a y être parvenu.

[36] Voir Chapitre 3 pour une discussion des analyses de l'efficacité des clauses de clémence et une application de leur revers potentiel. Buccirossi & Spagnolo (2001) en proposent une application au cas de la corruption.

(iv) Principal – agent – corrupteur : l'effet de délégation

La réciprocité constitue donc l'un des instruments privilégiés de mise en œuvre des pactes de corruption. Par ce biais, le corrupteur peut obtenir de l'agent une décision qui lui est favorable, mais qui est, par nature, contraire aux intérêts du principal (Section *(ii)*). Pour l'en dissuader, le principal recourt dans le cadre du contrat de délégation à des instruments d'incitation et de contrôle destinés à instaurer une relation d'efficience avec l'agent (Section *(i)*). L'analyse expérimentale proposée dans ce chapitre met en évidence les conséquences sur le comportement de corruption de cette position particulière de l'agent, à l'intersection de deux engagements contradictoires.

Les théories du salaire d'efficience ont récemment bénéficié d'un important regain d'intérêt, lié au développement conjoint des concepts de préférences tournées vers les autres et de la méthode expérimentale. Un vaste courant de recherche tend en effet à fonder le salaire d'efficience non sur un niveau de satisfaction attendue (Section *(i)*), mais sur une relation de réciprocité entre le principal et l'agent.[37] Revenant à la formulation originelle d'un échange de dons et contre-dons, ces résultats établissent une relation positive entre la générosité du salaire choisit par le principal et l'intensité de l'effort exercé par l'agent. Plutôt que la perception qu'en a l'agent, ce sont donc les intentions du principal et, partant, la nature de la relation établie qui influencent le comportement de l'agent (Fehr, Falk & Fischbacher, 2000).[38]

Lorsqu'un échange de dons caractérise la relation de délégation, la décision de cor-

[37] Les résultats expérimentaux de Fehr, Kirchsteiger & Riedl (1993) confirment l'instauration d'une relation de réciprocité entre un principal et un agent par l'intermédiaire du salaire versé. Les expériences de Hennig-Schmidt, Rockenbach & Sadrieh (2005) montrent par ailleurs l'importance, dans l'instauration de cette relation, de la capacité de l'agent à apprécier la bienveillance du principal qui se manifeste dans le choix du salaire. Voir également Danthine & Kurmann (2005) pour une version théorique du salaire d'efficience explicitement formulée en termes de réciprocité.

[38] Akerlof & Yellen (1990) proposent une évaluation de l'importance empirique de ce type de relation de délégation.

ruption est donc guidée par des incitations à la réciprocité contradictoires entre elles : l'agent se trouve dans l'impossibilité d'honorer simultanément les relations de réciprocité instaurées respectivement par le principal – via le salaire d'efficience – et par le corrupteur – par le versement d'un pot-de-vin. Si son importance est confirmée, ce *conflit de réciprocités* est à l'origine d'un *effet de délégation* sur le comportement de corruption, au sens où il dépend de la nature de la relation établie par le principal, plutôt que des dispositions du contrat. Si le principal est parvenu à instaurer une relation de réciprocité avec l'agent, l'effet de délégation devrait donc tendre, toutes choses égales par ailleurs, à diminuer l'inclination de l'agent à être corrompu. L'un des instruments privilégié pour y parvenir est le salaire choisit par le principal. Le niveau du salaire versé détermine alors l'acuité du conflit de réciprocités auquel fait face l'agent lorsqu'il reçoit une proposition de corruption. L'effet de délégation pourrait donc constituer un élément supplémentaire d'explication quant à l'influence du salaire sur le comportement de corruption.

Afin d'évaluer la pertinence empirique de l'effet de délégation, nous introduisons un jeu de corruption à trois joueurs qui s'inspire des traits caractéristiques décrits dans les sections précédentes. Une exception importante est que nous excluons le risque de détection afin d'isoler l'effet de délégation d'autres déterminants du comportement de corruption. Il convient cependant de remarquer que tant le risque de détection que le coût d'opportunité que constitue le salaire devraient réduire l'incitation à participer à un contrat de corruption (Section *(i)*). Dans la mesure où c'est également le résultat attendu de l'effet de délégation, ces mécanismes devraient jouer dans le même sens. Ce choix ne devrait pas, par conséquent, altérer la pertinence qualitative de nos résultats.

Ce jeu de corruption permet de soumettre l'effet de délégation à réfutation expérimentale. Le plan expérimental comporte deux expériences, conduites sous plusieurs traitements. La première expérience reproduit un jeu de corruption à deux joueurs : un agent reçoit un salaire exogène pour prendre une décision coûteuse, qui affecte le bien-être d'un corrupteur. Avant de prendre cette décision, l'agent peut accepter ou

refuser le pot-de-vin que ce dernier lui propose. Dans la seconde expérience, une étape préliminaire est ajoutée, au cours de laquelle un troisième joueur, le principal, choisit le salaire versé à l'agent. Les comparaisons inter-expériences permettent donc de mettre en évidence l'influence, sur le comportement de corruption, de la relation qui lie le principal et l'agent. Chaque expérience est par ailleurs conduite sous différents niveaux de salaire, qui constituent autant de traitements. Les comparaisons intra-expérience reflètent par conséquent le rôle joué par le niveau de salaire.

Le jeu de corruption ainsi que son analyse théorique sont présentés dans la Section 1.1. Il forme la base de l'expérience que nous avons conduite, dont les protocole et modalité sont présentés à la Section 1.2. Les comportements observés tendent à confirmer l'importance de l'effet de délégation sur le comportement de corruption. Le rôle corollaire du niveau de salaire est également confirmé par les observations. Ces résultats, présentés dans la Section 1.3, apportent un nouvel éclairage sur la corruption bureaucratique et les instruments de son contrôle (Section 1.4).

1.1 Jeu de corruption à trois joueurs

Etudier l'effet de délégation nécessite de prendre en compte le comportement des trois joueurs impliqués dans une relation de corruption. Pour ce faire, nous proposons une extension du jeu à deux joueurs utilisé dans les travaux expérimentaux précédents. Malgré un certain nombre de caractéristiques originales, l'équilibre du jeu correspond au résultat traditionnel, selon lequel l'illégalité du pacte de corruption constitue un obstacle à sa mise en oeuvre. La réciprocité est un l'un des moyens de le surmonter et l'effet de délégation est issu de cette déviation par rapport à l'équilibre. Nous proposons donc, ensuite, une description formelle des prédictions théoriques fondées sur la réciprocité et l'effet de délégation.

1.1.1 Description du jeu

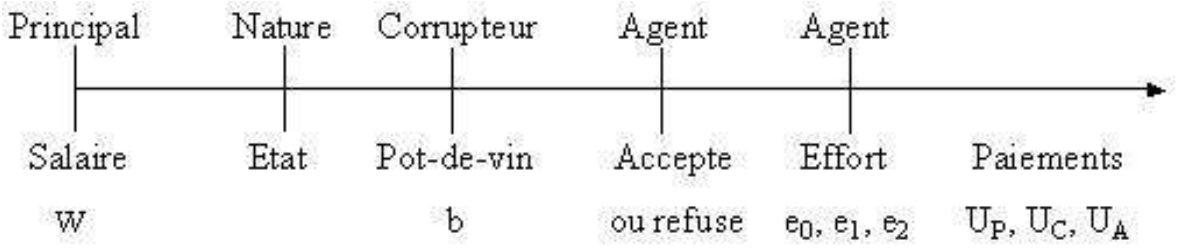
Nous considérons un jeu de corruption à trois joueurs : un Agent (A), un Principal (P) et un Corrupteur (C). L'Agent obtient du Principal une rémunération fixe, notée w, pour assumer la responsabilité du pouvoir qui lui est délégué. On note e l'effort exercé par l'Agent dans la réalisation de cette tâche, au coût $v(e)$. Si l'effort était supposé vérifiable par le Principal, celui-ci pourrait immédiatement repérer et sanctionner la corruption. L'hypothèse d'information imparfaite sur le niveau d'effort exercé est par conséquent une condition nécessaire pour que s'instaure une relation de corruption. Nous considérons donc une fonction de production aléatoire, au sens où l'effort détermine la production à un terme d'erreur près, noté ϵ. Quelle que soit la réalisation de cet état de la nature, l'effort détermine simultanément les gains du principal et ceux du corrupteur, respectivement notés $f_P(e,\epsilon)$ et $f_C(e,\epsilon)$. Pour l'un comme pour l'autre, les gains sont croissants du niveau d'effort de sorte que : $\frac{\partial f_P(e,\epsilon)}{\partial e} \geq 0$ et $\frac{\partial f_C(e,\epsilon)}{\partial e} \geq 0$. Il existe cependant un conflit d'intérêt entre le Principal et le Corrupteur quant à la nature de l'effort exercé : la création de valeur qui en découle peut être relativement plus favorable à l'un, l'autre ou aucun d'entre eux.

Par souci de simplification, nous considérons en effet trois niveaux d'effort : $e \in \{e_1; e_2; e_3\}$. Le premier niveau correspond à un effort de tire-au-flanc, qui n'assure au Principal comme au Corrupteur qu'un gain minimum. Cet effort est supposé être exercé sans coût par l'Agent : $v(e_0) = 0$. En exerçant l'un des deux autres niveaux d'effort, l'Agent accroît le niveau de production. Dans la terminologie classique des modèles d'agence, ces deux niveaux d'effort correspondent donc à un effort *positif*. L'agent peut cependant choisir de procurer un gain plus important :

- soit au Principal, en choisissant le niveau d'effort e_1 – appelé pour cette raison effort *productif* – tel que $f_P(e_1,\epsilon) > f_C(e_1,\epsilon), \forall\epsilon$;
- soit au Corrupteur en choisissant l'effort e_2 – effort *corrompu* – tel que $f_P(e_2,\epsilon) < f_C(e_2,\epsilon), \forall\epsilon$.

Pour éviter de créer par hypothèse une distorsion en faveur de l'un de ces efforts, nous

Graphique 1.1 – Forme séquentielle du jeu de corruption



supposons que leur coût est identique pour l'agent : $v(e_1) = v(e_2) = c > 0$. Etant donné un niveau d'effort positif (*i.e.* coûteux), choisir l'effort productif ou l'effort corrompu modifie donc uniquement la répartition du surplus ainsi créé entre le Principal et le Corrupteur. Plutôt que par leur intensité, ces deux niveaux d'effort se distinguent donc uniquement par leurs conséquences sur les parties avec lesquelles l'Agent prend un engagement.

Le premier engagement est pris dans le cadre d'une relation légale, mise en oeuvre grâce au salaire versé par le Principal. Le Corrupteur peut en proposer un second, illégal, en offrant un pot-de-vin, noté b, destiné à inciter l'Agent à choisir l'effort corrompu. L'Agent peut, s'il le souhaite, refuser ce contrat tacite. On note b_a le montant du transfert accepté par l'agent, $b_a = b$ si l'offre est acceptée, $b_a = 0$ si elle est refusée. Les bénéfices que le Corrupteur peut attendre de l'Agent sont conditionnés par l'existence d'un pouvoir discrétionnaire hérité du principal. La première relation préexiste donc à la seconde. La structure d'information du jeu (dont la forme séquentielle est présentée dans le Graphique 1.1) reproduit ces caractéristiques essentielles :

1. *Le Principal choisit le niveau de salaire de l'Agent, w ;*
2. *L'état de la nature, ϵ, est révélé au Corrupteur et à l'Agent ;*
3. *Le corrupteur offre un pot-de-vin, b ;*
4. *L'Agent accepte ou refuse le transfert, b_a, puis choisit son effort, $e \in \{e_1; e_2; e_3;\}$;*
5. *L'Agent, le Principal et le Corrupteur reçoivent les paiements associés à ces décisions.*

Puisque le risque de détection généralement associé à la corruption est ici exclu de l'analyse, nous retenons l'hypothèse de neutralité au risque. Les fonctions de paiements (U) du jeu peuvent donc être résumées de la façon suivante :

$$\begin{cases} U_P = f_P(e_i, \epsilon) - w \\ U_C = f_C(e_i, \epsilon) - b_a \\ U_A = w + b_a - v(e_i) \end{cases} \quad \text{avec} \quad \begin{cases} v(e_i) = c \ \forall i = 1, 2 \, ; \, c > 0 \\ v(e_0) = 0 \end{cases}$$

Par souci de réalisme, nous considérons une version répétée à horizon fini du jeu de corruption. L'équilibre de Nash parfait en sous-jeu concorde alors avec celui du jeu instantané. Bien que la minimisation du coût de l'effort le conduise, en dernière étape, à choisir l'effort de tire-au-flanc ($e^* = e_0$), l'Agent accepte donc tout pot-de-vin qui lui est offert ($b^* = b$). Anticipant cette décision, le Corrupteur renonce, à l'étape précédente, à proposer un pacte de corruption ($b^* = 0$). De la même façon, le Principal choisit le niveau de salaire le plus faible possible.

Proposition 1.1. *L'équilibre de Nash parfait en sous-jeu du jeu de corruption à 3 joueurs élimine la corruption.*

Ce résultat théorique est le propre des pactes de corruption, en raison de leur illégalité (Section *(iii)*). La réciprocité est cependant l'un des arguments avancés pour réconcilier cette prédiction avec les faits observés. L'existence d'un effet de délégation s'appuie sur cette déviation de l'équilibre.

1.1.2 Hypothèses de travail

Bien que la corruption ne constitue pas un équilibre du jeu que nous utilisons (Proposition 1.1), un certain nombre d'études expérimentales (Abbink, Irlenbush et Renner,

2002 ; Abbink, 2004 ; 2002) suggèrent en effet que la réciprocité entre l'Agent et le Corrupteur favorise la mise en oeuvre des pactes de corruption. Dans la mesure où le jeu de corruption que nous proposons diffère substantiellement des jeux existants[39], il convient de vérifier la robustesse de ce résultat aux changements que nous introduisons.

***Hypothèse* 1.1.** *La réciprocité permet de mettre en oeuvre les pactes de corruption.*

Si la réciprocité permet de mettre en oeuvre les pactes de corruption, nous avons vu que l'introduction de la relation de délégation entre le Principal et l'Agent peut faire naître un conflit de réciprocités. Si le Principal et le Corrupteur font simultanément appel à la loyauté de l'Agent – respectivement en lui offrant un pot-de-vin et en lui versant un salaire élevé – celui-ci devra en effet choisir de trahir l'une ou l'autre de ces offres. Lorsque le Principal offre un salaire élevé à l'Agent, nous nous attendons donc à ce qu'un effet de délégation tende à diminuer le niveau de corruption.

***Hypothèse* 1.2.** *Lorsque l'Agent est confronté à un conflit de réciprocités, un effet de délégation tend à dissuader la corruption.*

L'existence d'un effet de délégation repose sur le niveau du salaire choisit par le Principal. Si l'Hypothèse 1.2 est confirmée, l'effet de délégation devrait donc être à l'origine d'une corrélation négative entre le niveau de salaire et l'incitation à être corrompu. Dans la mesure où le jeu exclut tout risque de détection, cette corrélation devrait disparaître lorsque la relation de délégation est absente.

***Hypothèse* 1.3.** *L'effet de délégation crée une relation négative entre le niveau de salaire et la mise en oeuvre des pactes de corruption.*

[39]Entre autres différences, Abbink et *al.* (2002) et Abbink (2004, 2002) imposent des utilités marginales du revenu différentes entre le corrupteur et l'agent en multipliant le transfert reçu par un coefficient fixe. La question du fondement de cette différence comme du choix le plus judicieux de son ampleur restant largement ouverte, nous avons préféré écarter cette possibilité. Ces travaux considèrent par ailleurs un jeu de corruption à deux joueurs et se limitent par conséquent à deux niveaux d'effort, qui ne se distinguent que par leur effet sur les paiements du corrupteur.

Prises dans leur ensemble, ces trois hypothèses évaluent l'existence, et tirent les conséquences, d'un effet de délégation sur le comportement de corruption. Leur validité est appréciée grâce aux données obtenues dans le cadre d'expériences en laboratoire, fondées sur le jeu de corruption décrit plus haut.

1.2 Protocole de l'expérience

La méthode expérimentale est particulièrement adaptée pour tester l'influence des incitations sur les comportements puisqu'elle consiste à "*créer, dans le laboratoire, un environnement micro-économique contrôlé, où une mesure précise des variables pertinentes est garantie*"[40]. Dans le cas particulier de la corruption, l'économie expérimentale permet en outre de contourner les difficultés d'observation inhérentes à la discrétion qui entoure les activités illégales. L'"environnement micro-économique" regroupe l'ensemble des incitations offertes aux participants et des modalités selon lesquelles ils interagissent. Ces caractéristiques constituent le protocole de l'expérience, présenté dans la prochaine section.

Contrairement aux données naturelles, les données expérimentales sont recueillies par l'expérimentaliste lui-même, auprès de participants dont le comportement est contraint par le protocole et le déroulement de l'expérience. On peut donc s'attendre à ce que les résultats observés dépendent de façon importante des procédures utilisées. Nous consacrons donc la Section 1.2.2 à une description détaillée des modalités pratique de collecte des données.

[40] Wilde (1980), *in* The Philosophy of Economics, cité par Smith (1982).

1.2.1 Plan d'expériences

Pour qu'il permette de tester l'influence d'un effet de délégation, le plan d'expériences doit se distinguer des travaux existants selon deux aspects importants. Premièrement, l'existence d'un Principal est une condition nécessaire pour qu'apparaisse un effet de délégation. Cet aspect nécessite de mettre en oeuvre un jeu de corruption à trois joueurs ; ce qui implique, en particulier, d'introduire explicitement le conflit d'intérêt qui oppose le Corrupteur et le Principal. Deuxièmement, la condition suffisante est que le Principal choisisse les conditions de la délégation qui confère le pouvoir discrétionnaire à l'Agent. Il convient donc que les conditions de délégation – ici, le salaire – soient endogènes à l'expérience, et décidées par le Principal. Le protocole de l'expérience est conçu de façon à pouvoir séparer l'influence respective de ces deux changements.

Nous réalisons en effet deux expériences successives. La première, appelée *Expérience de Corruption* (EC), est destinée à assurer la comparabilité avec les résultats expérimentaux existants en isolant la relation de corruption à deux joueurs entre le Corrupteur et l'Agent. La seconde introduit les décisions du principal quant à la rémunération de l'Agent. Cette *Expérience de Délégation Explicite* permet donc, par comparaison avec la première, d'observer l'impact de la relation avec le Principal sur le comportement de corruption. A travers un certain nombre de traitements, nous introduisons, en outre, différentes variations de salaire qui permettent de tester l'effet de délégation.

a) Expérience de Corruption (EC)

La première expérience inclut deux joueurs : l'Agent et le Corrupteur. Elle permet donc d'isoler la relation de corruption. Le jeu mis en œuvre dans le cadre de cette première expérience se déroule en trois étapes.

L'Agent reçoit d'abord un salaire exogène (W, précisé ci-dessous) dont le montant est de connaissance commune. Ce salaire, versé à chaque période, est irrémédiablement acquis pour l'Agent, quelles que soient ses décisions futures. L'état de la nature est ensuite tiré aléatoirement dans sa distribution et annoncé à l'Agent comme au Corrupteur. Le Corrupteur doit alors décider du montant du pot-de-vin offert à l'agent, b.

L'Agent prend ensuite deux décisions successives. Il choisit d'abord d'accepter ou de refuser le pot-de-vin proposé, puis il choisit son niveau d'effort pour la période parmi l'un des trois niveaux mutuellement exclusifs : e_0, e_1 ou e_2.

Enfin, le Corrupteur et l'Agent sont informés de l'ensemble des décisions prises au cours de la période ainsi que des gains pour la période (en ECU[41]) qui en résultent.

b) Expérience de Délégation Explicite (EDE)

Cette seconde expérience réintègre les conditions qui président à la délégation de pouvoir. Une étape préalable est donc ajoutée à l'Expérience de Corruption, au cours de laquelle un Principal choisit, et verse, le salaire perçu par l'Agent.

Afin que le choix du Principal soit clairement interprété par les participants comme ou égoïste ou bienveillant (*fair*), ce salaire est choisi entre deux niveaux possibles. Le salaire faible (W_F) est proche du salaire de réserve de l'Agent dans la mesure où il compense la désutilité d'un effort positif. Le salaire élevé (W_E) correspond, quant à lui, à un partage de rente entre l'Agent et le Principal. La décision du Principal au cours de cette étape préliminaire correspond donc à : $w = \{W_L, W_E\}\,; W_L < W_E$.

Avant de participer aux trois étapes suivantes, correspondant au jeu utilisé dans

[41] *Experimental Currency Unit*, unité de compte utilisée dans l'expérience avant conversion en Euro.

l'Expérience de Corruption, l'Agent et le Corrupteur sont informés du choix du Principal. Pendant l'ensemble du déroulement de l'expérience, ce dernier ne connaît que ses gains pour la période (en ECU) et ses propres décisions. Il n'est en aucun cas informé des décisions de l'Agent.

c) Traitements

Afin d'évaluer l'impact du niveau de salaire sur le comportement de corruption, chaque expérience est conduite en utilisant trois paramétrages pour le salaire. Dans un Traitement de Salaire Modéré (TSM), la dotation exogène de l'Agent dans l'Expérience de Corruption ainsi que le salaire élevé dans l'Expérience de Délégation Explicite sont fixés à $W_E = 30$. L'un et l'autre sont portés à $W_E = 40$ au cours d'un Traitement de Salaire Fort (TSF). De plus, le Principal doit, dans l'Expérience de Délégation Explicite, choisir entre un salaire élevé et le salaire faible W_L. Nous observons donc le comportement de corruption associé à ce niveau de salaire dans l'Expérience de Délégation Explicite aussi souvent que le Principal le choisit. Un troisième traitement est par conséquent ajouté à l'Expérience de Corruption, dans lequel la dotation exogène est fixée à W_L.

L'ensemble des combinaisons de ces expériences et traitements constituent autant de conditions (EC-15, EDE-15, EC-30, etc ...) sous lesquelles le comportement de corruption est observé. Elles permettent, en particulier, d'étudier la sensibilité du comportement de corruption au niveau de salaire selon que celui-ci est versé dans le cadre d'une relation contractuelle à deux ou à trois joueurs.

Tableau 1.2 – Fonction de production

Effort	Principal	Corrupteur
e_0	70	70
e_1	100	70
e_2	70	100

1.2a : Etat favorable ($\overline{\epsilon}$)

Effort	Principal	Corrupteur
e_0	40	40
e_1	70	40
e_2	40	70

1.2b : Etat défavorable ($\underline{\epsilon}$)

1.2.2 Déroulement des sessions

a) Paramètres

Les expériences proposées reproduisent la structure du jeu de corruption présenté à la Section 1.1. La fonction de production est donc une caractéristique centrale de la structure d'intérêts qu'entretiennent les joueurs. Ces propriétés ont présidé au choix de la fonction de production utilisée dans le cadre des expériences, présentée dans le Tableau 1.2. Le surplus créé par un effort positif est constant, fixé à 30 ECU, qu'il s'agisse d'un effort productif (e_1) ou corrompu (e_2). L'état de la nature, quant à lui, ne consiste qu'en un accroissement additif du surplus créé par un effort positif : $\epsilon \in \{\overline{\epsilon}; \underline{\epsilon}\}$. Ces états sont supposés équiprobables ($p = 0.5$). Ces hypothèses permettent d'éviter que les choix ne soient guidés par un souci de maximisation de l'efficience.[42] Les deux types d'effort positif se distinguent, en revanche, par leurs conséquences en termes de répartition du surplus. Nous avons choisi un cas extrême de cette propriété, puisque le surplus bénéficie intégralement au principal lorsque l'Agent choisit l'effort productif, et au Corrupteur lorsqu'il opte pour l'effort corrompu.[43]

[42] Ainsi, par exemple, des participants souhaitant optimiser la valeur créée dans l'économie auraient plus probablement choisi : un effort positif dans l'état de la nature favorable si celui-ci permettait un accroissement du surplus ; et un effort productif plutôt qu'un effort corrompu si celui-ci engendrait un surplus plus important.

[43] Ce paramétrage rend maximales les incitations à être le bénéficiaire de la bienveillance de l'Agent, et semble donc le plus approprié pour étudier l'impact du conflit de réciprocités.

La désutilité d'un effort positif pour l'Agent est fixée à $c = 10$. Le salaire faible, $W_L = 15$, est donc très proche du niveau de salaire qui rend l'Agent indifférent entre un effort positif ($v(e_1) = v(e_1) = c$) et un effort nul ($v(e_0) = 0$). Nous avons cependant choisi un niveau légèrement supérieur au salaire d'indifférence afin d'offrir une rente de participation minimale aux participants qui se voient attribuer le rôle d'Agent.

Compte tenu de ces niveaux de salaire, le surplus transféré par le Principal à l'Agent en optant pour un salaire élevé plutôt que pour le salaire faible (égal à $30 - 15 = 15$ pour TSM et $40 - 15 = 25$ pour TSF) est toujours inférieur au surplus créé par un effort positif. Si le Principal considère qu'un salaire généreux lui garantira la bienveillance de l'Agent, choisir le salaire élevé est donc une stratégie dominante dans tous les traitements. Afin de limiter le risque de gains négatifs, le montant du transfert proposé par le Corrupteur est plafonné respectivement à 70 ou 100, selon que l'état de la nature est favorable ou défavorable.[44] Dans chacun de ces deux cas, le surplus espéré lorsque l'effort corrompu est choisi est très supérieur à la borne imposée. Il est donc très improbable que cette restriction influence les comportements observés dans les expériences.

b) Conditions pratiques

A leur arrivée, les participants procèdent à un tirage au sort qui sélectionne l'ordinateur qui leur est attribué. Les groupes sont alors constitués, et les rôles attribués, de façon aléatoire par le serveur qui gère l'expérience. Conformément au jeu de corruption décrit plus haut, l'ensemble de l'expérience se déroule en groupes fixes (*partners*). Les rôles et les groupes constitués sont donc maintenus inchangés pendant toute l'expérience. Séparés par des cloisons, les participants ne peuvent en aucun cas identifier les membres de leur groupe. Une fois les participants installés, les instructions leur sont lues avant qu'ils n'apprennent leur rôle. Leur compréhension du jeu n'est donc pas in-

[44] Cette restriction constitue une tradition méthodologique. Il est communément admis que le protocole dot être conçu de façon à proscrire la possibilité de gains négatifs.

fluencée par le rôle qu'ils auront à tenir et nous nous assurons ainsi qu'ils comprennent l'ensemble des enjeux de l'expérience.

Les participants jouent successivement chacune des deux expériences (EC et EDE). L'Expérience de Corruption est d'abord proposée aux participants à qui sont attribués les rôles d'Agent et de Corrupteur. Chaque traitement est répété pendant cinq périodes. L'ensemble de l'Expérience de Corruption comprend donc 15 périodes, réparties entre les trois niveaux de salaire exogène : les cinq premières sous W_L puis les 5 suivantes sous TSM et, enfin, les cinq dernières sous TSF.

Les participants qui jouent le rôle de Principal sont placés dans une pièce connexe. Ils sont priés de patienter pendant le déroulement de l'Expérience de Corruption, et peuvent accéder à internet pour s'occuper. A la fin de cette première phase, de nouvelles instructions – correspondant à l'Expérience de Délégation Explicite – sont lues et distribuées dans chacune des deux salles. Les portes séparant les salles sont alors ouvertes afin que les procédures (en particulier l'introduction de principaux et l'existence d'agents et de corrupteurs) soient de connaissance commune. Chaque traitement de l'expérience de délégation explicite est répété pendant 5 périodes. Ils se succèdent par ordre de salaire croissant, TSM puis TSF. A la fin de l'expérience, un questionnaire est proposé au participants afin de recueillir des informations sur leurs caractéristiques individuelles.

Les instructions lues aux participants – reproduites dans l'Annexe, Section 1.B – éliminent toute référence au contexte de corruption afin de purger les comportements observés de la charge éthique généralement associée à ce type de relation.[45] Ainsi le pot-de-vin offert est-il présenté comme une proposition de *transfert* et l'effort comme

[45]Bien que ce choix méthodologique paraisse plus cohérent avec l'analyse théorique proposée, les résultats obtenus par Abbink & Hennig-Schmidt (2005) tendent à nuancer l'effet de la contextualisation sur le comportement de corruption. En tout état de cause, la neutralité des instructions permet d'éliminer les effets de contexte qui sont une source d'hétérogénéité inobservable.

une *décision*. Les rôles sont désignés par un ordre alphabétique neutre. L'Agent est appelé *participant X*, le Corrupteur *participant Y* et le Principal *participant Z*. Mimant leur position respective dans le Tableau 1.2, les états de la nature favorable et défavorable sont respectivement désignés comme le *tableau gauche* et le *tableau droit*. Afin de vérifier la bonne compréhension de ces instructions, les participants disposent d'un questionnaire pré-expérimental, reproduit dans l'Annexe (Section 1.C). Les réponses au questionnaire ainsi qu'à toute autre question privée sont fournies avant le début de l'expérience. L'ensemble de ces élément est de connaissance commune.

Les gains obtenus par chaque participant sont calculés à partir de la somme des ECU accumulés pendant l'ensemble des périodes de l'expérience, selon le taux de conversion de 1 Euro pour 100 ECU. Ces gains sont versés de façon privée, à la fin de l'expérience. Le taux horaire de rémunération atteint une moyenne de 10 Euros, ce qui constitue une rémunération très attractive compte tenu du taux du salaire minimum.

Quatre sessions ont été conduites dans le laboratoire d'économie expérimentale du GATE. Le script informatique du plan d'expériences a été développé en utilisant le logiciel Regate (Zeiliger, 2000), en collaboration avec Romain Zeiliger. Ces sessions ont réuni un total de 87 participants, constitués d'étudiants inscrits en premier cycle à l'ITECH (Institut TExtile et Chimique de Lyon), à l'EM Lyon (Ecole de Management de Lyon) et à l'Ecole Centrale de Lyon. L'appariement par groupe de trois étant constant nous disposons donc de 29 observations indépendantes.

1.3 Déterminants du comportement de corruption

Les hypothèses présentées dans la Section 1.1.2 décrivent les comportements théoriques prédits par l'existence d'un effet de délégation. Les observations obtenues dans le cadre des expériences sont comparées à ces prédictions pour en évaluer la pertinence

empirique. Pour ce faire, nous étudions les conditions de mise en œuvre des pactes de corruption dans l'ensemble des expériences (Hypothèse 1.1), puis la variation de comportement provoquée par l'intervention du Principal (Hypothèses 1.2 et 1.3).

1.3.1 Mise en œuvre des pactes de corruption

Par définition, une relation de corruption est établie si le versement d'un pot-de-vin assure le détournement du pouvoir discrétionnaire confié à l'Agent. Dans le cadre des expériences, nous considérons donc qu'un pacte de corruption a été mis en œuvre dès lors qu'un transfert accepté a conduit l'Agent à choisir l'effort corrompu. Malgré les prédictions théoriques, cette mesure confirme la présence de pactes de corruption dans les expériences.

Observation 1.1. *Un nombre significatif de pactes de corruption sont mis en œuvre.*

Résultats Le Tableau 1.3 présente les taux d'acceptation et de corruption observés sous chaque condition (combinaison expérience/traitement).

Tableau 1.3 – Taux d'acceptation et de corruption

Condition		**Acceptation**	**Corruption**
Traitement	**Salaire**	**du transfert**	
EC	15	73 %	23 %
EC	30	67 %	30 %
EC	40	68 %	33 %
EDE - TSM	15	59 %	27 %
EDE - TSM	30	44 %	18 %
EDE - TSF	15	49 %	14 %
EDE - TSF	40	51 %	13 %

D'une part, l'agent accepte fréquemment le transfert proposé, le taux d'acceptation variant de 44% à 73% selon les conditions. Cette acceptation ne conduit pas de façon systématique à la conclusion

d'un contrat implicite de corruption, puisque le taux de corruption atteint au maximum 33%. Malgrè cette différence, un nombre significatif de pactes de corruption sont mis en œuvre.

Pour tester l'importance de ce phénomène, nous avons construit un traitement artificiel correspondant au cas où le agents ne sont jamais corrompus. La comparaison entre les comportements observés et cette situation hypothétique met donc évidence la prégnance de la corruption dans les expériences. Pour tous les traitements, l'hypothèse que les comportements observés sont identiques à une situation où la corruption est absente est rejetée (par un test de Wilcoxon sur données appariées, à 1%).

Comme l'indique le Tableau 1.4, il semble que les corrupteurs anticipent cette réaction favorable aux transferts proposés. L'offre moyenne de transferts est en effet comprise entre 8.46 ECU et 14.06 ECU selon les conditions. L'hypothèse que les observations sont identiques à une situation où le transfert proposé serait nul à toutes les périodes est rejetée à plus de 95% (test de Wilcoxon pour données appariées). ■

La mise en œuvre des pactes de corruption résulte de deux déviations successives par rapport à la stratégie d'équilibre. D'une part, alors même qu'à la dernière étape le choix de l'Agent ne devrait se poser qu'en terme de minimisation du coût, il choisit fréquemment de subir la désutilité maximale afin d'exercer un effort positif. D'autre part, à l'étape précédente le Corrupteur tend à proposer un pot-de-vin positif, et ce malgré l'absence de garantie quant à la mise en œuvre de l'effort corrompu.

Ces deux déviations permettent conjointement l'émergence de relations de corruption. Pourtant, au-delà de leur mise en œuvre effective, ce sont les motivations sous-jacentes aux relations de corruption qui ouvrent la possibilité d'un conflit de réciprocités. Plus spécifiquement, les choix de corruption ne peuvent être influencés par un conflit de réciprocités qu'à condition que leur mise en œuvre repose sur une relation de réciprocité entre l'Agent et le Corrupteur. Il existe cependant un certain nombre de concepts concurrents de préférences tournées vers les autres qui pourraient expliquer les comportements décrits ci-dessus.[46] Dans la mesure où nous disposons d'observations répétées, le protocole de l'expérience permet de sélectionner l'hypothèse de comportement qui fournit l'explication la plus adéquate aux choix de corruption. Au regard

[46] Altruisme et aversion à l'inégalité en sont des illustrations. Rabin (1998) propose une synthèse très complète des enrichissements de la rationalité hérités des recherches en psychologie cognitive.

de leur dynamique, les comportements observés dans les expériences semblent bien, en effet, être motivés par des considérations de réciprocité.

Observation 1.2. *La mise en œuvre des pactes de corruption repose sur une relation de réciprocité entre l'Agent et le Corrupteur.*

Résultats La décision de l'agent d'exercer l'effort corrompu est positivement corrélée avec le montant du transfert proposé.

Le Tableau 1.4 présente les montants moyens proposés sous chaque condition, décomposés en fonction du comportement induit de l'agent. Un agent est considéré comme s'étant comporté de façon *honnête* s'il a refusé le transfert et n'a pas choisi l'effort corrompu. Ce comportement est principalement associé à des niveaux de transfert très faibles, compris en moyenne entre 0.02 ECU et 2.66 ECU. Une *trahison* correspond à un transfert accepté qui n'a pas été suivi de l'effort corrompu. Le transfert moyen conduisant à cette décision varie de 8.13 ECU à 15.06 ECU selon les traitements. Il correspond à des niveaux de transfert intermédiaires. Les transferts les plus élevés sont ceux qui s'avèrent capables d'inciter l'agent à conclure un contrat implicite de *corruption* : accepter le transfert et choisir l'effort corrompu.

TABLEAU 1.4 – TRANSFERT MOYEN EN FONCTION DU COMPORTEMENT DE L'AGENT

Condition		**Comportement de l'agent**			**Total**
Traitement	**Salaire**	**Honnêteté**	**Trahison**	**Corruption**	
EC	15	2.66	13.72	28.14	14.06
EC	30	0.26	15.06	25.26	13.33
EC	40	0.38	10.25	24.66	11.89
EDE - TSM	15.00	0.89	12.89	25.33	11.38
EDE - TSM	30.00	0.48	14.36	30.60	9.49
EDE - TSF	15.00	0.02	8.13	29.47	6.85
EDE - TSF	40.00	0.05	13.80	24.40	8.46

Ce premier aperçu est confirmé par le niveau de corrélation entre les décisions de corruption et le niveau du transfert proposé, résumé dans le Tableau 1.5. Tous les coefficients de corrélation sont positifs : une augmentation du transfert proposé accroît la probabilité et d'acceptation et de corruption. Pour toutes les conditions sauf une (EC sous un salaire de 40), le coefficient de corrélation entre l'acceptation et le transfert est plus faible que celle de la corruption avec le transfert. Ce résultat

reflète l'effet des trahisons, par lesquelles l'agent accepte le transfert proposé mais n'exerce pas l'effort corrompu. Pour l'acceptation comme pour la corruption, l'hypothèse nulle d'indépendance avec le transfert (*i.e.* corrélation non significativement différente de zéro) est rejetée à plus de 99% de confiance (test de correlation de Spearman).

Tableau 1.5 – Corrélation (Spearman) avec le transfert proposé

Condition		**Corruption**	**Acceptation**	**Corruption**
Traitement	**Salaire**		**du transfert**	**Passée**
EC	15	0.612***	0.595***	0.576***
EC	30	0.749***	0.596***	0.627***
EC	40	0.696***	0.719***	0.716***
EDE - TSM	15	0.758***	0.675***	0.469***
EDE - TSM	30	0.832***	0.680***	0.538***
EDE - TSF	15	0.708***	0.644***	0.569***
EDE - TSF	40	0.792***	0.486***	0.449***

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Les décisions du corrupteur semblent également réagir aux décisions de l'agent. La dernière colonne du Tableau 1.5 présente les coefficients de corrélation entre le transfert proposé à la période t et la décision de corruption à la période $t-1$. Un coefficient de corrélation positif indique donc que le corrupteur tend à accroître le transfert proposé en réaction à une acceptation suivie de l'effort corrompu. Pour toutes les conditions, les corrélations sont positives et significativement différentes de 0. Les observations permettent de désagréger la réaction du corrupteur. Si les transferts proposés varient considérablement selon les conditions, aucune différence n'apparaît entre les niveaux de transfert *positifs* : conditionnellement à la proposition d'un transfert non nul, le montant proposé est constant (le comportement de transfert dans EDE pour un salaire de 40 est la seule exception à cette observation). Les corrupteurs tendent donc à réagir positivement aux décisions de l'agent, mais en s'appuyant sur une stratégie binaire (transfert positif/nul) plutôt qu'en ajustant de façon continue le montant du transfert. ■

Le comportement du Corrupteur obéit à une stratégie de participation plutôt qu'à un choix d'intensité : les variations de l'environnement – décision de l'Agent notamment – conduisent le corrupteur à se retirer de la relation de corruption. Cette conclusion confirme les résultats obtenus, dans une expérience à deux joueurs, par Abbink, Irlen-

busch & Renner (2002). Elle suggère que le coût de participation au pacte de corruption présente une discontinuité qui déconnecte la décision de proposer un transfert de celle de son montant.

Surtout, il apparaît que la mise en œuvre de la corruption repose sur une réciprocité bilatérale entre l'Agent et le Corrupteur. Si l'Agent réagit positivement à la générosité du Corrupteur, en choisissant d'autant plus souvent l'effort corrompu que le transfert proposé est élevé ; le Corrupteur est, lui aussi, sensible à la loyauté manifestée par l'Agent, et propose d'autant plus fréquemment un transfert positif que l'Agent a, dans le passé, honoré le pacte de corruption.

Les Observations 1.1 et 1.2 confirment l'Hypothèse 1.1 : en dépit de son illégalité, la corruption trouve dans la réciprocité un mécanisme efficace de mise en œuvre. L'apparition de pactes de corruption dans nos expériences permet en outre d'étudier la sensibilité du comportement de corruption aux variations des conditions de la délégation.

1.3.2 Effet de délégation

Un pacte de corruption est mis en œuvre lorsque l'Agent récompense la générosité du Corrupteur en choisissant l'effort corrompu. Dans l'expérience de délégation explicite, c'est un Principal qui verse, à l'Agent, un salaire choisi entre deux niveaux possibles. Ce salaire est constant et versé quelles que soient les décisions ultérieures. Lorsque le salaire est élevé (W_E), ce choix manifeste donc la volonté du Principal d'opter pour un salaire d'efficience et d'établir par là une relation de réciprocité avec l'Agent. Du point de vue du Principal, choisir un salaire d'efficience n'est en effet rationnel que si l'Agent exerce en retour l'effort productif.

Dans la mesure où ils diffèrent exclusivement par la répartition du surplus créé par un effort positif, l'effort corrompu et l'effort productif sont, par nature, mutuellement

exclusifs. Ainsi, lorsque le Corrupteur et le Principal choisissent simultanément de faire appel à la réciprocité de l'Agent, celui-ci se trouve confronté à un conflit de réciprocités. Les comparaisons entre expériences (CE contre EDE) semblent confirmer son importance dans les choix de corruption.

Observation 1.3. *La corruption est d'autant moins probable que le conflit de réciprocités auquel fait face l'Agent est intense.*

Résultats Les résultats obtenus dans cette section proviennent de comparaisons croisées entre les différentes conditions. Elles consistent à tester la variation du comportement de corruption induite par le passage d'une condition de contrôle à une condition de traitement. Dans la mesure où l'expérience est conduite en *partners*, les tests réalisés sont des tests pour données appariées. Les Tableaux 1.6 et 1.7 présentent les résultats des tests de Wilcoxon sur des données appariées qui testent la différence de comportement entre la condition de contrôle, indiquée dans la première colonne, et la condition de traitement qui apparaît dans la deuxième. Les seuils critiques de rejet de l'hypothèse nulle selon, laquelle le comportement reste inchangé entre les conditions, sont également présentés.

Le signe de la statistique de test indique le sens de la variation de comportement. Il est positif si la corruption est plus élevée sous la condition de contrôle que sous la condition de traitement ; négatif dans le cas contraire. Les conditions de contrôle et de traitements sont définies de façon à ce que les hypothèses présentées dans la Section 1.1.2 prédisent toujours une diminution, ou une stabilité, du niveau corruption dans le passage d'une condition à l'autre. Nos prédictions théoriques correspondent donc toujours à un effet positif – ou, éventuellement, nul.

La Section 1.3.1 a montré qu'une relation de réciprocité est à l'origine de la mise en œuvre des pactes de corruption. Conditionnellement à cette relation, les résultats de cette section se concentrent sur le comportement de l'agent afin de mettre en évidence l'influence du conflit de réciprocités qu'engendre la délégation. Par ailleurs, la validation de l'Observation 1.2 a montré que la fonction de réaction du corrupteur est essentiellement binaire, consistant à renoncer ou non à inciter l'agent par un transfert positif. Le montant du transfert proposé est donc insensible aux variations de l'environnement induites par le changement de conditions. Afin d'isoler l'effet des incitations sur la réaction de l'agent de celle qu'engendrent les ajustements du corrupteur, nous nous limitons donc aux observations où le transfert proposé est positif.

Enfin, nous proposons une confirmation supplémentaire à nos résultats à travers des régressions sur les variables de décision, présentées dans l'Annexe de ce chapitre (Section 1.A). Bien qu'elles fournissent d'importantes indications sur l'effet marginal des variables exogènes, elles doivent être considérées avec la plus grande précaution. En particulier, le traitement économétrique employé ne tient pas compte des

problèmes d'endogénéité dus à l'inclusion de variables de décisions dans la liste des variables exogènes (niveau du transfert dans la régression sur le comportement de corruption, par exemple.)

L'effet de la relation de délégation sur le comportement de corruption est mis en évidence par les comparaisons entre expériences pour un niveau donné du salaire. Les résultats sont présentés dans le Tableau 1.6.

Tableau 1.6 – Comparaisons inter-expériences

Contrôle	**Traitement**	**Difference**	
(CE)	(EDE)	**Signe**	**Seuil**†
15	15	–	0.112
30	TSM - 30	+	0.096*
40	TSF - 40	+	0.008***

† Valeurs critiques, ∗ 10%, ∗∗ 5%, ∗ ∗ ∗ 1%.

Une différence marquée apparaît en fonction du niveau de salaire obtenu par l'agent. Si le niveau de corruption est – de façon non-significative – plus élevé dans EC que dans EDE lorsque le salaire est le plus faible, il décroît entre les expériences dès lors que le salaire atteint l'un des niveaux plus élevés (TSM ou TSF). Pour chacun de ces deux niveaux de salaire, la statistique de test est positive et l'hypothèse nulle de stabilité du comportement est rejetée à plus de 95% de confiance. ■

L'introduction explicite du Principal diminue l'incitation pour l'Agent à participer à la corruption. Si ce résultat confirme l'existence d'un effet de délégation – par lequel la relation entre le Principal et l'Agent constitue une barrière naturelle à la corruption – il faut, une fois encore, approfondir les motivations de ce comportement pour relier l'effet de délégation à un conflit de réciprocités.

En comparaison de l'expérience de corruption, l'expérience de délégation explicite introduit en effet deux changements dans l'environnement économique. D'une part, elle réintroduit le Principal à l'origine de la délégation de pouvoir. Ce premier changement conduit à intégrer dans l'expérience le conflit d'intérêt qui oppose le Corrupteur et le Principal quant aux décisions de l'Agent.[47] D'autre part, les modalités d'incitation qui

[47] Cet aspect se traduit pratiquement par l'ajout des gains du Principal dans la fonction de production

lient l'Agent et le Principal sont endogènes dans l'expérience de délégation explicite afin de mettre en évidence l'influence des conditions de la délégation sur le comportement de corruption. C'est par ce dernier aspect qu'un conflit de réciprocités peut apparaître.

Les conditions (combinaisons expérience/traitement) sous lesquelles nous observons le comportement de corruption permettent de distinguer ces deux effets. Les observations sous un salaire faible, selon que celui-ci est exogène ou choisi, permettent de mettre en évidence le rôle du conflit d'intérêt entre le corrupteur et l'Agent. Les motivations liées à un conflit de réciprocités sont en effet absentes de cette comparaison, puisque le choix d'un salaire faible par le Principal ne peut en aucun cas être interprété par l'Agent comme l'instauration d'une relation de réciprocité. Les observations dont nous disposons rejettent l'hypothèse que ce conflit d'intérêt entre le Corrupteur et le Principal influence le comportement de corruption. C'est au contraire lorsque le Principal choisit un salaire élevé que l'inclination de l'Agent à être corrompu diminue significativement. C'est donc bien le conflit de réciprocités rencontré par l'Agent qui sous-tend l'effet de délégation.

Les conséquences de ce résultat sont à l'origine de l'Hypothèse 1.3 tire les conséquences de ce résultat : un conflit de réciprocités survient lorsque le Principal offre un salaire d'efficience à l'Agent ; ce conflit est d'autant plus prégnant que le salaire versé par le Principal est élevé. Si ce conflit de réciprocités affecte la propension de l'Agent à être corrompu, il devrait donc créer une corrélation négative entre le niveau de salaire et le niveau de corruption. Ce conflit de réciprocités est absent lorsque l'interaction entre le Corrupteur et l'Agent est isolée des décisions du Principal. Le niveau de salaire devrait donc laisser inchangé le niveau de corruption observé dans ce contexte. Cette hypothèse est, en partie, confirmée par les comparaisons intra-expériences.

Observation 1.4. *Le niveau de corruption tend à être* ***croissant*** *du salaire reçu par l'Agent. Cette relation est inversée en présence d'un conflit de réciprocités.*

présentée aux participants. Voir les instructions de l'Expérience de Délégation Explicite, Section 1.B.

Résultats Le Tableau 1.7 présente les résultats des tests de la variation de comportement entre les traitements, au sein de chaque expérience. Les tests présentés mettent donc en évidence la sensibilité du comportement de corruption aux accroissements de salaire pour une relation de délégation donnée (anonyme pour EC, explicite pour EDE).

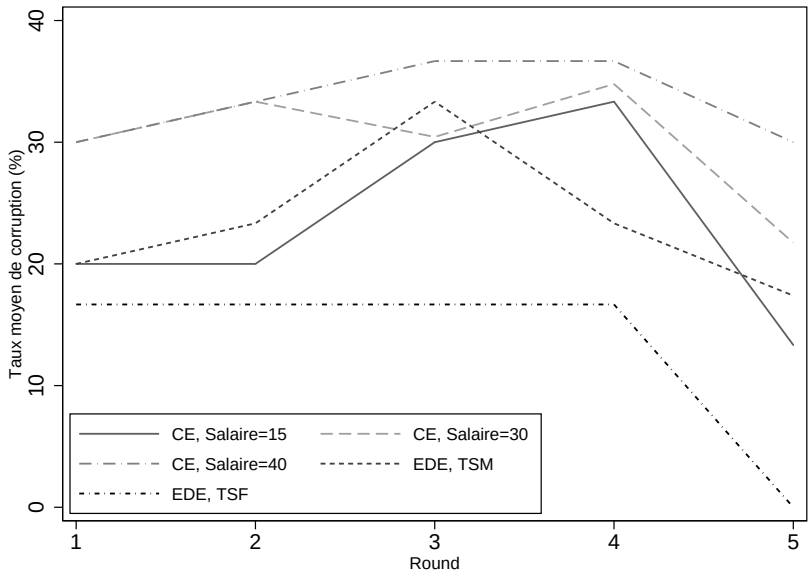
Tableau 1.7 – Comparaisons intra-expériences

Salaire		**EC**		**EDE**	
Contrôle	**Traitement**	**Signe**	**Seuil**†	**Signe**	**Seuil**†
15	30	-	0.090*	+	0.056*
15	40	-	0.016**	+	0.034**
30	40	-	0.491	+	0.083*

† Valeurs critiques, $*$ 10%, $**$ 5%, $***$ 1%.

Le première colonne résume les résultats obtenus pour les comparaisons intra CE. Toutes les variations sont négatives, indiquant un accroissement du niveau de corruption avec le salaire. Elles sont statistiquement significatives pour les deux premières comparaisons (salaires élevés contre salaire faible).

Graphique 1.2 – Evolution du niveau de corruption au sein des traitements



Les trois traitements dans EC sont joués successivement, par ordre de salaire croissant. L'observation précédente pourrait donc être attribuée à un processus d'apprentissage, plutôt qu'à l'effet

marginal des augmentations de salaire. Il faut noter que, dans ce cas, l'apprentissage devrait influencer le comportement non seulement entre les traitements, mais également au sein de chacun. Le Graphique 1.2 propose un premier aperçu du profil d'apprentissage au sein de chaque condition. Pour chaque traitement, l'évolution du niveau moyen de corruption entre les périodes est représenté. Les profils d'évolution sont très similaires entre les traitements et caractérisés par un fort effet de fin de jeu. Pour tester statistiquement l'influence de l'apprentissage, nous avons comparé les comportements d'une période à l'autre au sein de chaque traitement. Pour tous les traitements, l'hypothèse de stabilité du comportement au cours du temps est acceptée (test de Wilcoxon sur données appariées).

Les niveaux moyens de transfert présentés dans le Tableau 1.4 donnent une indication du comportement qui sous-tend cette observation. Le transfert moyen qui conduit les agents à être corrompus est en effet décroissant du salaire dans CE comme dans EDE. Il semble donc que l'effet marginal du salaire soit de diminuer le transfert minimum suffisant à ce que l'agent soit corrompu.

Dans EDE, la délégation s'ajoute à cet effet. A l'inverse des résultats précédents, les statistiques qu lui sont associées, présentées dans la seconde colonne, sont toutes positives. Le niveau de corruption tend donc à décroître lorsqu'un principal offre un salaire plus élevé. Toutes les variations sont significatives à plus de 10%.[48] ■

Le jeu de corruption auquel nous nous intéressons exclut par hypothèse les mécanismes traditionnellement considérés comme reliant le comportement de corruption au niveau de salaire.[49] Ce choix permet d'isoler le rôle de l'effet de délégation dans le lien que ces variables entretiennent : lorsque la relation de délégation est explicitement prise en compte, l'intensité du conflit de réciprocités est d'autant plus forte que le salaire est élevé. La propension de l'Agent à être corrompu s'en trouve diminuée.

En l'absence de conflit de réciprocités, dans l'Expérience de Corruption, ce protocole devrait en outre assurer la neutralité du salaire sur le comportement de corruption. Loin de confirmer cette hypothèse, nos résultats mettent en évidence un effet original : en

[48]Ces résultats sont confirmés par l'estimation fournie en Annexe, Tableau 1.A : l'effet marginal du salaire est un accroissement significatif de la probabilité lorsqu'il est mesuré en interaction avec la relation de délégation. La régression rejette également l'hypothèe d'un apprentissage, puisque contrôler l'effet de fin de jeu suffit à épurer l'effet du temps.

[49]Le risque de détection, les problèmes de sélection adverse ou le sentiment d'équité en sont autant d'exemples, décrits dans la Section *(i)*.

l'absence de tout autre mécanisme – risque de détection notamment – la propension à être corrompu tend à être croissante du niveau de salaire. Comme nous l'avons vu, cet aspect peut résulter de l'interaction entre le corrupteur et l'agent, lorsque l'effet du salaire sur l'honnêteté est plus que compensée par l'accroissement du pot-de-vin (cas (*ii*), Tableau 1.1). Les comportements observés suggèrent cependant une interprétation légèrement différente des mécanismes à l'œuvre. Plutôt au le niveau du pot-de-vin proposé, il semble en effet que le pot-de-vin proposé tende à diminuer lorsque le salaire augmente. Il apparaît en effet que les augmentations de salaire sont associées à une diminution du niveau moyen de transfert qui incite l'Agent à être corrompu. Le salaire reçu tend donc, toutes choses égales par ailleurs, à diminuer le transfert de réserve, c'est à dire le transfert nécessaire à ce que les Agents acceptent d'exercer l'effort corrompu.

1.4 Conclusion

Dans ce chapitre, nous avons montré que la corruption tire ses spécificités des relations qu'entretiennent les trois joueurs en présence. Une situation de corruption lie un principal, un agent et un corrupteur. Outre qu'elle consiste donc en une relation d'agence à trois joueurs, une relation de corruption se définit comme un accord par lequel le mandataire collabore avec un tiers à l'insu du mandant, dans une transaction illégale dont eux seuls tirent bénéfice.

Ainsi, tandis que le principal et l'agent sont liés par une relation de délégation avec aléa moral légale et, en ce sens, conforme à l'analyse traditionnelle des incitations ; le corrupteur apparaît, quant à lui, comme une tierce partie, affectée par les décisions prises dans le cadre de cette délégation et dont les intérêts divergent de ceux du principal. Ce conflit d'intérêt le conduit à instaurer une relation illégale avec l'agent, à ravers la conclusion d'un pacte de corruption. Quoique cette relation s'établisse à son détriment, le principal ne dispose que d'instruments de contrôle indirects sur le corrupteur dans

la mesure où ils appartiennent à des organisations distinctes. Tirer les conséquences de ces caractéristiques a permis à l'analyse microéconomique d'aboutir à une compréhension approfondie tant des déterminants que des conséquences du comportement de corruption.

La contribution de ce chapitre consiste à exploiter la position particulière qu'occupe l'agent à l'intersection de ces relations croisées. Plus précisément, un certain nombre de situations de corruption mettent l'agent en face d'un conflit de réciprocités : lorsque principal et corrupteur font simultanément appel à la réciprocité de l'agent, la décision prise doit nécessairement trahir l'une de ces propositions. La propension de l'agent à être corrompu s'en trouve donc diminuée, et la relation de délégation constitue, alors, une barrière naturelle à la corruption.

Le comportement observé dans les expériences que nous avons conduites confirme l'importance de ce mécanisme. La corruption est donc d'autant moins répandue qu'une relation de réciprocité préside à la délégation du pouvoir discrétionnaire. Ces résultats apportent un éclairage original sur la diffusion de la corruption et les instruments qui permettent de la contrôler. D'abord, ils pourraient participer à expliquer pourquoi les administrations publiques sont plus sujettes à la corruption que d'autres types d'organisation (Banfield, 1975 ; Banerjee, 1997). Outre qu'elles concentrent un certain nombre des conditions qui favorisent l'émergence de la corruption (Rose-Ackerman, 1975), elles reposent en effet sur une organisation hiérarchique anonyme et impersonnelle (Crozier, 1963). En ce sens, elles se montreraient donc plus vulnérables au développement de la corruption puisque celle si ne fait plus intervenir, alors, qu'une seule relation de réciprocité.

En raison de l'importance de la réciprocité dans sa mise en œuvre, Abbink (2004) propose, ensuite, de lutter contre la corruption en instaurant une rotation fréquente du personnel. Il montre en effet que la relation de réciprocité qui fonde la corruption est d'autant moins forte que les relations entre le corrupteur et l'agent sont de courte

durée. Nos résultats tendent, pourtant, à nuancer l'efficacité potentielle de cette mesure. Si la rotation du personnel permet de rompre les relations qu'entretiennent l'agent et le corrupteur, elle peut également amoindrir l'efficience de la relation de délégation. La question de savoir si la relation de délégation est maintenue tant que l'agent appartient à la même organisation reste largement ouverte ; dans le cas contraire, la rotation du personnel peut conduire à un affaiblissement de la relation entre le principal et l'agent, donc à un accroissement de la corruption.

Enfin, les résultats établissent également la capacité d'augmentations de salaires à lutter contre la corruption, en l'absence même de détection. Simultanément, nous avons observé que s'il n'existe pas d'effet de délégation le transfert de réserve tend à être décroissant du salaire et, partant, la corruption d'autant plus répandue que le salaire est élevé. S'ils confirment le rôle du niveau des salaires dans la lutte contre la corruption, nos résultats mettent donc en évidence l'importance de la nature de la relation de délégation par laquelle ils sont décidés : les augmentations de salaire peuvent donc s'avérer contre-productives, à moins qu'elles s'inscrivent dans une structure organisationnelle qui entretient une relation de proximité entre le principal et l'agent. La responsabilité des supérieurs hiérarchiques directs dans la détermination des conditions de la délégation apparaît donc, dans ce cadre, comme un élément clé du succès des incitations dans la lutte contre le corruption.

Annexes

1.A Regression

Le Tableau 1.A présente les résultats de régression (Probit) de la décision de corruption sur les variables décrivant l'environnement (*Délégation*, *Tableau gauche*, *Fin de jeu* et *Round*) ainsi que l'interaction entre les joueurs (*Salaire* et *Transfert proposé*). Les décisions retardées sont intégrées afin de contrôler partiellement d'éventuels problèmes d'endogénéité.

Tableau 1.A – Décision de corruption

Variable	**Coefficient**	**(Std. Err.)**
Probit sur la décision de corruption		
Corruption rétardée	0.089	(0.241)
Acceptation rétardée	-0.183	(0.266)
Transfert rétardé	-0.007	(0.011)
Transfert proposé	0.103***	(0.010)
Salaire	0.038***	(0.010)
Salaire sous Délégation	-0.059***	(0.019)
Délégation	1.253***	(0.479)
Tableau gauche ($\bar{\epsilon}$)	0.499**	(0.196)
Fin de jeu	-0.715**	(0.327)
Round	0.027	(0.113)
Constante	-3.399***	(0.554)
Distribution estimée		
$\hat{\sigma}$	2.440***	(0.945)

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Probit à effets aléatoires individuels. La variable endogène (*Corruption*) est égale à 1 lorsque le participant accepte le transfert (pot-de-vin) et choisit la décision B (effort corrompu). L'hétérogénéité individuelle inobservable est intégrée par un effet aléatoire propre aux participants (ayant le rôle d'Agent). Les variables de *Corruption retardée*, d'*Acceptation retardée* et de *Transfert retardé* mesurent respectivement la décision de corruption, la décision d'acceptation et le montant du transfert proposé, à la période précédente. La variable *Transfert* correspond au montant du transfert proposé par le corrupteur. La variable *Salaire* mesure le niveau de salaire versé à l'Agent, il est mesuré en interaction avec l'expérience (*Dénonciation* égale à 1 pour l'Expérience de Délégation Explicite) dans la variable *Salaire sous Délégation*. La variable *Tableau gauche* indique un tirage favorable de l'état de la nature ; La variable *Fin de jeu* indique la dernière période d'un traitement, la variable *Round* mesure le passage du temps au sein de chaque traitement (réinitialisée au début de chacun).

1.B Instructions de l'expérience

Les instructions présentées ci-dessous sont lues aux participants au début de l'expérience. La prise de décision au cours de l'expérience se fait à l'aide d'une interface graphique dont le Graphique 1.A propose une capture d'écran.

Vous participez à une expérimentation dans laquelle vous pouvez gagner de l'argent. La somme d'argent que vous gagnerez dépendra de vos décisions ainsi que des décisions des autres participants de votre groupe. Toutes les décisions que vous aurez à prendre et toutes les informations dont vous disposerez se feront par le terminal informatique qui vous a été attribué.

Cette expérimentation comprend **deux phases**. Les règles de la première phase sont décrites ci-dessous. A la fin de la première phase, de nouvelles règles vous seront distribuées.

Au début de cette expérimentation, des groupes de 2 personnes (vous-même et 1 autre participant) sont constitués au hasard. Chaque membre du groupe a un rôle différent qui lui est attribué au début de l'expérimentation, ce rôle est X ou Y. **Chaque participant conserve le même rôle et appartient au même groupe tout au long de l'expérimentation**.

REGLES DE LA PREMIERE PHASE

La première phase se décompose en **trois parties** de **cinq périodes** chacune.

Déroulement d 'une période

Au début de chaque période, le participant X reçoit une dotation, connue de tous les participants.

Au début de chaque période, un tirage au sort détermine le tableau selon lequel la période se déroule. Il s'agit du tableau *gauche* ou du tableau *droit* avec une probabilité 0.5 pour chaque tableau. X et Y savent quel est le tableau tiré au sort.

Décision de X	Coût pour X	Gain de Y
A	0	70
B	10	70
C	10	100

Transfert minimum	0
Transfert maximum	100

Tableau gauche (g)

Décision de X	Coût pour X	Gain de Y
A	0	40
B	10	40
C	10	70

Transfert minimum	0
Transfert maximum	70

Tableau droit (d)

Au cours de la période, X doit choisir entre trois décisions : A, B ou C. Ce choix a des effets sur le gain du participant X **et** sur celui du participant Y. Par exemple, si le tableau *gauche* est tiré au sort et que X choisi la décision B, cela lui coûte 10 unités qui seront retirées

de ses gains de la période et cela rapporte 70 à Y.

Avant que X prenne sa décision, Y peut lui proposer un transfert. X peut accepter ou refuser ce transfert. Si le transfert est accepté, son montant sera retiré des gains de Y et ajouté à ceux de X. Si le tableau *gauche* est tiré au sort, le transfert proposé par Y devra être compris entre 0 et 100. Si le tableau *droit* est tiré au sort, le transfert devra être compris entre 0 et 70.

A la fin de chaque période, le participant Y est informé de la décision de X.

En résumé, le déroulement d'une période est donc le suivant : :

1ere Etape : X reçoit une dotation.
2eme Etape : Un tableau est tiré au sort avec une probabilité 0.5 pour chaque tableau. X et Y sont informés du tableau sélectionné ainsi que de la dotation de X.
3eme Etape : Y propose un transfert à X compris entre 0 et 100 si la période se déroule selon le tableau *gauche*, entre 0 et 70 si la période se déroule selon le tableau *droit*.
4eme Etape : X décide d'accepter ou de refuser ce transfert.
5eme Etape : X choisit A, B, ou C dans le tableau tiré au sort.
6eme Etape : Y obtient le gain déterminé par la décision de X dans le tableau tiré au sort.
Le gain net de Y est égal à ce gain moins le transfert qu'il propose à X si celui-ci l'accepte.
Le gain net de X est égal à sa dotation plus le transfert proposé par Y s'il l'accepte, moins le coût de sa décision.
A la fin de la période, tous les participants sont informés de leur gain et de leur gain net pour la période.

Déroulement d'une partie

La première phase comporte trois parties. **Rien ne change entre les trois parties à part le montant de la dotation** reçue par X au début de chaque période. Ce montant est annoncé à l'ensemble des participants avant la première période de chaque partie.

A la fin des cinq périodes, tous les participants commencent une nouvelle partie. Les groupes et les rôles restent les mêmes pour les trois parties.

Comment prendrez-vous vos décisions ?

Sur votre écran d'ordinateur, trois zones apparaissent :

La première vous informe de votre rôle tout au long de l'expérimentation ainsi que des règles qui s'appliquent à la partie que vous êtes en train de jouer.
La deuxième vous permet de prendre vos décisions. Pour prendre une décision, cliquez sur l'un des boutons présents à l'écran.
La troisième vous rappelle les décisions et les gains des périodes précédentes.

Exemple : Supposons que le participant X reçoit une dotation de 30.

Tous les participants sont informés de la dotation que reçoit X. La première période commence.

Un tableau est choisi au hasard : par exemple le tableau droit. Le résultat du tirage est annoncé à X et Y.

Le participant Y propose un transfert au participant X. Comme la période se déroule selon le tableau droit, le transfert maximum est 70.

Supposons que le participant Y propose 15.

Le participant X est informé du montant du transfert proposé.

Il choisit par exemple d'accepter puis prend une décision :

S'IL CHOISIT A	**S'IL CHOISIT B**	**S'IL CHOISIT C**
Gain de X : $30 + 15$ $= \mathbf{45}$	Gain de X : $30 + 15 - 10$ $= \mathbf{35}$	Gain de X : $30 + 15 - 10$ $= \mathbf{35}$
Gain de Y : $40 - 15$ $= \mathbf{25}$	Gain de Y : $40 - 15$ $= \mathbf{25}$	Gain de Y : $70 - 15$ $= \mathbf{55}$

Paiement de vos gains

A la fin de l'expérimentation, nous calculerons le total de vos points gagnés au cours de l'expérimentation. Cette somme sera convertie en Euros sur la base de 100 points = 1 €. Cette somme vous sera payée en espèce de façon privée à la fin de l'expérimentation.

A cette somme sera ajoutée un forfait de 3 €.

Si vous avez à poser des questions, levez la main, une personne viendra y répondre. Il vous est demandé de ne pas parler au cours de cette expérimentation. Toute communication entraînera votre exclusion sans paiement des gains éventuels. Merci de suivre ces consignes.

Merci de votre participation.

REGLES DE LA DEUXIEME PHASE

La deuxième phase se décompose en **deux parties** de **cinq périodes** chacune. Les groupes et les rôles de tous les participants **restent les mêmes que pendant la première phase**.

Au début de la deuxième phase, un participant est ajouté à votre groupe, il s'agit du participant Z. Le participant Z ajouté à votre groupe **reste le même tout au long de la deuxième phase**.

Changements dans le déroulement d'une période

Au début de chaque période, **le participant Z choisit la dotation du participant X**. Pour ce faire, il choisit, à chaque période, entre deux gains possibles. Les 2 gains entre lesquels Z doit choisir sont annoncés à tous les participants.

Après la décision de Z, tous les participants sont immédiatement informés de son choix. Ensuite, le déroulement de la période est le même que pendant la première phase. Le tirage au sort détermine le tableau selon lequel la période se déroule, avec une probabilité 0.5 pour chaque tableau. **Seuls les participants Y et X savent quel est le tableau tiré au sort**.

Décision de X	Coût pour X	Gain de Z	Gain de Y
A	0	70	70
B	10	100	70
C	10	70	100

Transfert minimum	0
Transfert maximum	100

Tableau gauche (g)

Décision de X	Coût pour X	Gain de Z	Gain de Y
A	0	40	40
B	10	70	40
C	10	40	70

Transfert minimum	0
Transfert maximum	70

Tableau droit (d)

X choisi ensuite entre A, B et C. Ce choix a des effets sur le gain du participant X, sur le gain du participant Y **mais aussi sur le gain du participant Z**. Par exemple, si le tableau *gauche* est tiré au sort et que X choisit la décision B, cela lui coûte 10 unités qui seront retirées de ses gains de la période et cela rapporte 70 à Y et 100 à Z. **A chaque période, seul le participant Y est informé de la décision de X**.

En résumé, le déroulement d'une période de la deuxième phase est donc le suivant :

1ere Etape : Z choisit la dotation de X entre 2 gains possibles.

2eme Etape : Le tableau selon lequel la période se déroule est tiré au sort avec une probabilité 0.5 pour chaque tableau. X et Y sont informés du tableau sélectionné et du choix de Z.

3eme Etape : Y propose un transfert à X compris entre 0 et 100 si la période se déroule selon le tableau gauche, entre 0 et 70 si la période se déroule selon le tableau droit.

4eme Etape : X décide d'accepter ou de refuser ce transfert.

5eme Etape : X choisit A, B, ou C dans le tableau tiré au sort.

6eme Etape : Y et Z obtiennent le gain déterminé par la décision de X dans le tableau tiré au sort.

Le **gain net de Y** est égal à ce gain moins le transfert qu'il propose à X si celui-ci l'accepte.

Le **gain net de Z** est égal à ce gain moins le gain pour X que Z a choisi.

Le **gain net de X** est égal à sa dotation plus le transfert proposé par Y s'il l'accepte, moins le coût de sa décision.

A la fin de la période, tous les participants sont informés de leur gain et de leur gain net pour la période.

Déroulement d'une partie

La deuxième phase comporte deux parties. **Rien ne change entre les deux parties à part les deux gains pour X entre lesquels le participant Z peut choisir**. Ces règles sont annoncées à l'ensemble des participants avant la première période de chaque partie.

A la fin des cinq périodes, tous les participants commencent une nouvelle partie. Les groupes et les rôles restent les mêmes pour les deux parties.

Exemple : Supposons que le participant Z doit choisir entre 15 et 30.

Tous les participants sont informés des gains entre lesquels le participant Z doit choisir. La première période commence.

Un tableau est choisi au hasard : par exemple le tableau droit. Le résultat du tirage est annoncé à X et Y.

Supposons que Z choisisse 30. Les participants X et Y en sont tous deux informés.

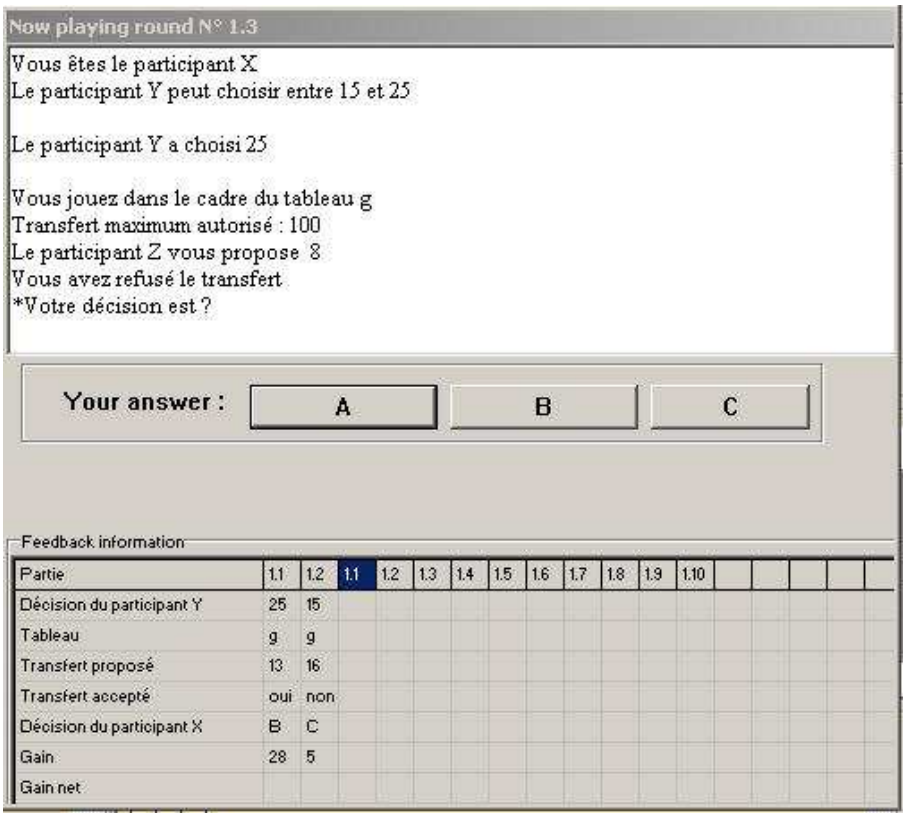
Le participant Y propose un transfert au participant X. Comme la période se déroule selon le tableau droit, le transfert maximum est 70.

Supposons que le participant Y propose 15.

Le participant X est informé du montant du transfert proposé. Il choisit par exemple d'accepter puis prend une décision :

S'IL CHOISIT A	S'IL CHOISIT B	S'IL CHOISIT C
Gain de X : $30 + 15$ $= \mathbf{45}$	Gain de X : $30 + 15 - 10$ $= \mathbf{35}$	Gain de X : $30 + 15 - 10$ $= \mathbf{35}$
Gain de Z : $40 - 30$ $= \mathbf{10}$	Gain de Z : $70 - 30$ $= \mathbf{40}$	Gain de Z : $40 - 30$ $= \mathbf{10}$
Gain de Y : $40 - 15$ $= \mathbf{25}$	Gain de Y : $40 - 15$ $= \mathbf{25}$	Gain de Y : $70 - 15$ $= \mathbf{55}$

Graphique 1.A – Ecran de contrôle de l'agent (EDE, LHW)



1.C Questionnaire pré-expérimental

1.C.1 Expérience de Corruption

Pour vous aider à comprendre les règles de la première phase, nous vous proposons de répondre aux questions suivantes :

Questions :

- Les groupes de 2 personnes sont reconstitués à chaque période 0 *Oui* 0 *Non*
- Le participant X peut refuser la dotation qui lui est attribuée 0 *Oui* 0 *Non*
- Les groupes de 2 personnes restent les mêmes entre les parties 0 *Oui* 0 *Non*
- Le participant Y est informé des décisions du participant X 0 *Oui* 0 *Non*
- Tous les participants connaissent la dotation du participant X 0 *Oui* 0 *Non*
- Chaque partie comprend ___ périodes.
- Le participant X reçoit une dotation à chaque période 0 *Oui* 0 *Non*

1^{er} Exemple : Le tableau gauche est tiré au sort. La dotation du participant X est 40. Le participant Y propose un transfert de 30.

Si le participant X refuse le transfert et choisit A, son gain net est : _____

Le gain net du participant Y est : _____

Si le participant X accepte le transfert et choisit C, son gain net est : _____

Le gain net du participant Y est : _____

2^{eme} Exemple : Le tableau droit est tiré au sort. La dotation du participant X est 30. Le participant Y propose un transfert de 25.

Si le participant X refuse le transfert et choisit A, son gain net est : _____

Le gain net du participant Y est : _____

Si le participant X accepte le transfert et choisit C, son gain net est : _____

Le gain net du participant Y est : _____

1.C.2 Expérience de Délégation Explicite

Pour vous aider à comprendre les règles de la deuxième phase, nous vous proposons de répondre aux questions suivantes :

Questions :

- Les groupes de 3 personnes sont reconstitués à chaque période 0 *Oui* 0 *Non*
- Le participant X peut refuser le gain que Z choisit pour lui 0 *Oui* 0 *Non*
- Le participant Y est informé des décisions du participant X 0 *Oui* 0 *Non*
- Le participant Z est informé des décisions du participant X 0 *Oui* 0 *Non*
- Tous les participants connaissent les gains entre lesquels Z choisi 0 *Oui* 0 *Non*
- Les participants X et Y ne changent pas de groupe 0 *Oui* 0 *Non*

1^{er} Exemple : Le tableau gauche est tiré au sort. Entre 15 et 40, le participant Z choisit 40. Le participant Y propose un transfert de 30.

Si le participant X refuse le transfert et choisit A, son gain net est : _____

Le gain net du participant Y est : _____

Le gain net du participant Z est : _____

Si le participant X accepte le transfert et choisit C, son gain net est : _____

Le gain net du participant Y est : _____

Le gain net du participant Z est : _____

2^{eme} Exemple : Le tableau droit est tiré au sort. Entre 15 et 30, le participant Z choisit 15. Le participant Y propose un transfert de 25.

Si le participant X accepte le transfert et choisit A, son gain net est : _____

Le gain net du participant Y est : _____

Le gain net du participant Z est : _____

Si le participant X refuse le transfert et choisit B, son gain net est : _____

Le gain net du participant Y est : _____

Le gain net du participant Z est : _____

Chapitre 2

Coût et qualité de l'offre de soins, quelle(s) rémunération(s) ? Une application au Québec[1]

> *«The re-engineering of health care will certainly require a reform in the way that medical providers are paid.[. . .] Fee-for-service payment promotes productivity but also encourages over-use ; conversely, paying doctors a straight salary in heavily regulated markets may lead to under performance, because it fails to reward productivity. Successful reform will involve remoulding payments systems so that they reward quality and performance. »*
>
> *"Keep taking the medicine"* , *The Economist* (15 Juillet 2004, p.17)

Dans le dernier classement qu'elle a consacré à la performance des systèmes de santé, en 2000, l'OMS plaçait la France au premier rang mondial.[2] La même année, la France

[1] Ce chapitre est inspiré d'un travail réalisé en collaboration avec Bernard Fortin et Bruce Shearer.

[2] Organisation Mondiale de la Santé (2000). *Rapport sur la santé dans le monde 2000*, Genève.

se situait parmi les systèmes de santé les plus chers de l'OCDE, se situant au 4° rang avec 9.4% de son PIB consacré aux dépenses de santé (à égalité avec le Canada, contre 10.3% pour l'Allemagne et 12.9% pour les Etats-Unis).[3] Le rapprochement de ces deux faits illustre parfaitement les termes contemporains dans lesquels se pose la question de la gestion des soins de santé. Après des décennies d'efforts dirigés presque exclusivement vers la maîtrise des coûts, c'est aujourd'hui vers la rentabilité de ces dépenses, en termes d'amélioration de la santé, que se déplace le débat (McGlynn, Asch, Adams & *al.*, 2003). L'accroissement tendanciel des ressources consacrées à la santé apparaît ainsi comme l'orientation naturelle d'une société développée[4] et l'objectif assigné à la politique de santé est dès lors formulé selon une exigence d'efficience plutôt que d'économies. Cette réorientation des objectifs nécessite un renouvellement profond des réflexions quant à l'efficacité des instruments de la politique de santé (Dudley, Miller, Korenbrot & *al.*, 1998 ; Malin & Keating, 2005), tant les spécificités du secteur médical et le souci de contrôle des coûts s'opposent à l'amélioration de la qualité (McNeil, 2001).

En raison de l'importance de son expertise, le médecin est au cœur du fonctionnement du système de santé (Arrow, 2001b). En termes de coûts, il représente à la fois un poste de dépenses important[5] et l'origine indirecte de la grande majorité des dépenses de santé, à travers son rôle de prescripteur. Cette préoccupation d'amélioration de la qualité s'est donc, en particulier, traduite par d'importantes innovations destinées à mieux contrôler la qualité de l'offre de soins, telle que l'instauration d'objectifs-cibles (Kiefe, Allison, Williams & *al.*, 2001).[6] En accord avec la majorité de la profession

[3]Sources : Direction de la Recherche, des Etudes, de l'Evaluation et des statistiques (Drees). Comparaison internationale des dépenses de santé, *Etudes et Résultats* n°175, 2002 ; Institut Canadien d'Information sur la santé (ICIS). Les soins de santé au Canada, 2000.

[4]Voir, par exemple, Le cercle des Economistes (2004). Economie de la santé : Une réforme ? Non, une révolution, *Cahier n°6.*

[5]Dans le cas du Canada, Deber, Narine, Baranek & *al.* (1998) évaluent à 15% la part des dépenses de santé inhérente à la rémunération des médecins.

[6]Voir également J. Carroll "Quality Counts : So Why Not Offer Physicians Bonuses ?", *California Health Consensus* (1 Janvier 2003), pour un survol d'initiatives prises récemment en ce sens aux Etats-unis et Rosenthal, Fernandopulle, Song & *al.* (2004) pour une revue critique.

(McKenna, 2002), le gouvernement britannique a récemment franchi un pas important dans cette direction, instituant, en 2002, un système de rémunération fondé sur un score de qualité (Shekelle, 2003) et susceptible d'affecter jusqu'à 18% de la rémunération totale des médecins généralistes (Smith & York, 2004). Le nouveau système s'appuie sur une évaluation large de la qualité de la pratique, puisque ce score est calculé à partir d'un tableau de bord composé de plus de 70 indicateurs. Le revers de cette innovation ambitieuse est d'avoir considérablement accru la complexité du système. Ce dernier effet semble l'avoir emporté, et explique en grande partie l'échec de cette réforme (Smith, 2003). Cette expérience milite donc en faveur de l'exploration des possibilités d'amélioration de la qualité offertes par les sytèmes de rémunération traditionnels, présentant l'avantage de la simplicité («*it may be more appropriate to pursue quality-oriented refinements of traditional payment approaches, rather than radical transformation.*». Cunningham, 2004, p.36). C'est la voie suivie dans ce chapitre, qui se propose d'évaluer la capacité des systèmes de rémunération fondés sur les mesures de performance traditionnelles à résoudre la contradiction entre amélioration de la qualité et contrôle des coûts.

La capacité des incitations financières traditionnelles à orienter la pratique professionnelle est pourtant moins évidente dans le cas des médecins que pour la plupart des autres professions. L'arbitrage entre consommation et loisir des médecins, en particulier, est fortement influencé par leur appartenance aux catégories de revenus les plus élevées.[7] Pour les individus appartenant à cette couche de la population, en effet, le niveau de rémunération est tel que l'effet revenu est susceptible de prendre l'avantage sur l'effet substitution (Feldstein, 1995). Dans ce cas, les médecins prendraient alors leurs décisions le long d'une fonction d'offre de travail à rebroussement, si bien que les variations de rémunération auraient un effet ambigu sur le nombre d'heures de travail ; positif ou négatif selon la position initiale le long de la fonction d'offre. Les premiers travaux consacrés à cette question ont confirmé cette crainte (Sloan, 1975). Feldstein

[7]Dans le cas des Etats-Unis, Showalter & Thurston (1997) indiquent ainsi que les médecins représentent 15% de la moitié la plus riche de la population.

(1970) estime ainsi à 60% la probabilité que l'offre de travail des médecins américains inclus dans son échantillon présente un rebroussement au salaire courant. Bien que l'effet revenu estimé reste important (Rizzo & Blumenthal, 1994), les travaux plus récents tendent cependant à renverser ces premières conclusions et obtiennent une élasticité positive de l'offre de travail au niveau de salaire offert (Showalter & Thurston, 1997). En raison de l'importance de l'effet revenu, cette élasticité tend à être inférieure au niveau généralement observé, oscillant entre 10% (Sæther 2003) et 30% (Baltagi, Bratberg & Holmas, 2005) en fonction des données et méthodes utilisées.

S'ils apparaissent comme un instrument utile de pilotage du comportement des médecins, les systèmes de rémunération doivent encore répondre à la double exigence de maîtrise des coûts et de promotion de la qualité des soins. Outre le volume de rémunération, et le volume d'heures de soins qui en résulte, c'est ainsi la nature de ces rémunérations et le type de pratiques qu'ils encouragent qui constituent la question centrale de la gestion des soins de santé.[8] En accord avec la préoccupation de Sloan (1975, p.554), pour qui «*no single variable may appropriately be used as an indication of physician supply behavior*», résoudre cette question implique en particulier de prendre en compte l'ensemble des dimensions de l'activité médicale. A cet égard, la réforme du mode de rémunération des médecins réalisée au Québec en 1999 constitue une innovation importante, par le type de rémunération instauré comme par les objectifs poursuivis.

Cette réforme consiste à offrir aux médecins spécialistes d'opter librement pour un mode de *rémunération mixte* alternatif aux modes de rémunération existants. En termes de politique de rémunération, ce nouveau système offre la particularité de combiner plusieurs instruments, puisqu'il associe un salaire forfaitaire, appelé *per diem*, à une diminution du taux de rémunération proportionnelle au nombre d'actes réalisés. L'adoption du nouveau mode de rémunération est de plus laissé à la discrétion des médecins,

[8] «*Empirical research as demonstrated financing methods as important tools in the management of health service. Knowledge of possible health effects for the patients as a consequence of financing methods seems limited.*»Aas (1995, p.205).

qui peuvent choisir de conserver le système antérieur. Pour reprendre les termes de ses concepteurs, la rémunération mixte a été introduite de façon à créer «*un mode équitable qui incite [les médecins] à avoir des comportements permettant à la fois d'améliorer les services à la population et d'être plus efficients.*»[9] Ces principes généraux se déclinent en trois objectifs. Les autorités souhaitent d'abord encourager la diversification des activités des médecins, en accordant le versement du *per diem* à toute activité médicale, du temps passé avec les patients aux heures de travail consacrées à l'administration des établissements ou à l'enseignement. La rémunération de ces activités constitue une nouveauté importante en comparaison du système le plus largement répandu jusqu'alors, la *rémunération à l'acte*, qui consiste uniquement en une rémunération proportionnelle aux actes délivrés et ignore le temps de travail qui n'est pas consacré aux patients. Cet élargissement des activités rémunérées devrait donc, ensuite, promouvoir l'équité des rémunérations, en offrant une compensation financière aux médecins qui acceptent de les exercer. En réduisant la rémunération proportionnelle aux actes réalisés, enfin, la rémunération mixte est également destinée à encourager les médecins à accroître le temps qu'ils consacrent à chaque patient et à améliorer ainsi la qualité des soins prodigués.

Par les instruments de rémunération qu'elle mobilise comme par ses résultats attendus, la rémunération mixte constitue donc une expérience originale, qui recouvre assez largement les préoccupations actuelles quant aux modalités de gestion de l'offre de soins de santé. Afin de mieux comprendre l'influence des incitations sur le coût et la qualité de l'offre de soins, ce chapitre propose une analyse des effets attendus et des résultats effectifs de l'introduction de la rémunération mixte. Dans ce but, l'analyse théorique et le modèle économétrique proposés intègrent explicitement les déterminants de l'arbitrage entre marges extensives (nombre d'actes, heures et semaines de travail) et marge intensive (temps consacré aux actes) de la pratique médicale. Dans ce cadre, la contrainte budgétaire qui gouverne les choix de pratique sous la rémunération mixte présente d'importantes non-linéarités. La méthode adoptée est destinée à résoudre les

[9] Conseil Médical du Québec (1997, p.13). Avis pour un mode mixte de rémunération des médecins de 2^{e} et 3^{e} lignes lié à leurs responsabilités, *Avis n° 97-03.*

problèmes analytiques posés par cette propriété.

Une première non-linéarité est due à l'endogénéité des prix lorsque le choix porte simultanément sur le nombre d'actes et le temps qui leur est consacré. Ainsi, par exemple, le prix qui rémunère une heure de travail consacrée aux patients dépend du nombre d'actes réalisés pendant ces heures, qui constitue lui-même une variable de choix. Cette première non-linéarité est analogue à celle que rencontrent les modèles d'arbitrage entre quantités et qualité dans les choix de consommation (Becker & Lewis, 1973). Le modèle théorique proposé emprunte par conséquent à cette littérature, en définissant des prix virtuels qui permettent une linéarisation locale de la contrainte budgétaire. La statique comparative du modèle permet de prédire les effets attendus du passage à la rémunération mixte. La combinaison des instruments – variation simultanée du salaire fixe et du taux de rémunération des actes – et l'arbitrage entre marges intensive et extensive rendent très ambigu l'impact des incitations sur les choix de pratique. Dans le cas général, rien ne garantit en particulier que la relation négative traditionnellement attendue entre le volume de soins (nombre d'actes et heures de travail) et l'affaiblissement des incitations financières soit respectée. Le modèle permet cependant d'isoler des conditions sur les préférences des médecins suffisantes à ce que ces résultats apparaissent.

L'analyse théorique est considérablement simplifiée par le caractère volontaire du passage à la rémunération mixte, qui permet d'interpréter son adoption comme une préférence révélée. En termes économétriques, en revanche, cet aspect se traduit par le risque que les estimations soient sujettes à un biais de sélection. Si les médecins qui choisissent la rémunération mixte se distinguent de ceux qui la refusent par des caractéristiques individuelles inobservables et corrélées avec les choix de pratique, les méthodes classiques de régression multiple échouent en effet à identifier l'effet des incitations sur ces choix. Ce problème est résolu en spécifiant un modèle économétrique structurel, dans lequel les paramètres estimés gouvernent les choix optimaux le long de la contrainte budgétaire.

L'endogénéité du schéma de rémunération affecte également la forme de cette contrainte budgétaire elle-même. En raison du libre choix entre les modes de rémunération alternatifs, la contrainte budgétaire est générée pour un choix de pratique donné par le mode de rémunération qui maximise le revenu. La contrainte budgétaire est donc linéaire par morceau, façonnée dans chaque espace de choix de pratique par le mode de rémunération optimal. Cette co-existence entre le mode de rémunération mixte et la rémunération à l'acte s'étend également aux choix de pratique conditionnels à son adoption. En raison des dispositions de sa mise en œuvre, les médecins qui ont choisi la rémunération mixte voient une partie de leurs activités rémunérées selon le mode de rémunération à l'acte. La contrainte budgétaire correspond alors à l'un ou l'autre des systèmes de rémunération, en fonction de l'adéquation de la pratique aux dispositions qui ouvrent droit à la rémunération mixte. Par définition, le passage de la rémunération à l'acte à la rémunération mixte provoque un changement simultané du salaire fixe reçu de la pratique – ordonnée à l'origine de la contrainte budgétaire – et du taux de rémunération des actes – pente dans le plan des actes. Les contraintes budgétaires associées à chaque mode de rémunération sont donc sécantes ; et chacune de ces deux propriétés est à l'origine de non-linéarités supplémentaires de la contrainte budgétaire le long de laquelle les médecins prennent leurs décisions.

Le choix de la méthode d'estimation du modèle répond au souci d'intégrer l'ensemble de ces non-linéarités. La solution la plus couramment adoptée consiste à utiliser l'algorithme d'Hausman, fondé sur un balayage de l'ensemble des segments linéaires de la contrainte (Burtless & Hausman, 1978 ; Hausman , 1979 ; 1980 ; 1985). Des travaux récents ont cependant montré que cette méthode restreint de façon importante les paramètres estimés. La convexité des préférences, en particulier, est imposée *a priori* par la méthode d'estimation (MaCurdy, Green & Paarsch, 1990), ce qui peut conduire à rejeter à tort les conditions de Slutsky (Meyer & Heim, 2003). Afin de laisser les comportements observés définir librement les préférences estimées, nous optons par conséquent pour une estimation par discrétisation de la contrainte budgétaire (Zabalza, Pissarides & Barton, 1980). Cette stratégie d'estimation n'impose aucune contrainte sur les para-

mètres estimés. La cohérence de la méthode requiert uniquement que l'utilité marginale du revenu soit positive (van Sœst, 1995).[10]

A travers l'évaluation de la rémunération mixte, ces outils permettent d'approfondir l'analyse théorique et empirique de la réponse optimale des choix de pratique aux variations des incitations. Un élément clé de cette approche, fondée sur la maximisation contrainte d'utilité, est la précision avec laquelle le niveau de consommation engendré par les choix de pratique est décrit. Dans notre cas, la contrainte budgétaire résulte des modalités institutionnelles qui gouvernent la rémunération des médecins du Québec avant et après la réforme (Section 2.1). Les traits essentiels de la rémunération mixte et les objectifs qui ont présidé à son instauration constituent les principaux ingrédients de l'analyse théorique (Section 2.2). Afin de lever les ambiguïtés qui en découlent, l'analyse économétrique intègre les dispositions précises de la rémunération mixte. Le modèle permet d'estimer les paramètres de préférences des médecins de l'échantillon, sous l'hypothèse que les choix observés maximisent l'utilité sous une contrainte de revenu discrétisée (Section 2.3).

L'identification empirique du modèle repose sur le comportement de pratique observé de l'ensemble des médecins du Québec entre 1996 et 2002. Les données trimestrielles utilisées couvrent donc, sous forme de panel, une période de 6 ans centrée sur l'année de la réforme. Surtout, ces données combinent des résultats d'enquête sur les semaines et les heures travaillées (ventilées selon le type d'activités) et des données administratives sur le volume d'actes délivrés et le revenu tiré de la pratique. Elles permettent donc

[10] Cette capacité à offrir un traitement adéquat des non-convexités de l'ensemble budgétaire a motivé de nombreuses applications de cette méthode en économétrie de l'offre de travail, au nombre desquels s'inscrivent, notamment, Hoynes (1996), Colombino (1998), Keane & Moffitt (1998), Euwals & van Soest (1999), Blundell, Duncan, McCrae & *al.* (2000) et van Sœst, Das & Gong (2002). Nyffeler (2005) propose une synthèse critique de l'adéquation de cette méthode à l'analyse de l'offre de travail. A notre connaissance, Sæther (2005) est le seul exemple d'application de cette méthode à l'économie de la santé, consacrée aux choix des médecins en termes d'établissements de pratique et d'heures de travail clinique.

d'associer le niveau réel de consommation à la pratique observée, lacune qui a constitué jusqu'à présent une barrière importante à l'analyse des choix de pratique des médecins.[11] A partir de ces observations, l'estimation structurelle des préférences des médecins et la modélisation de la contrainte budgétaire permettent d'évaluer l'effet propre de la rémunération mixte sur les comportements de pratique et de simuler l'effet potentiel de réformes alternatives (Section 2.4). Ces résultats montrent en particulier l'importance que revêt la liberté de choix du mode de rémunération, et participe par là aux réflexions contemporaines sur la capacité des incitations à améliorer la qualité des soins à un coût maîtrisé (Section 2.5).

2.1 Institutions : la contrainte budgétaire des médecins du Québec

Le gouvernement fédéral du Canada conditionne le financement des soins de santé à la conformité à un standard national. La politique publique de santé reste cependant une prérogative assez largement provinciale. Cette autonomie se traduit notamment par une grande diversité des politiques de santé au Canada. Cette section se limite au cas Québecois, d'où proviennent les données utilisées dans la partie empirique de notre analyse. L'introduction d'un mode de rémunération mixte, en 1999, constitue un changement profond dans les rémunérations des médecins du Québec, puisque la rémunération à l'acte était, jusqu'alors, le système le plus largement répandu.[12] Ces disposi-

[11] « *The greatest impediment to understanding physician behavior in Canada is the lack of data linking details of physician practice setting with individual and household physician income.* », Ferrall, Gregory & Tholl (1998, p.24).

[12] Emery, Auld & Lu (1999) proposent une synthèse très complète des différences institutionnelles entre les Etats de l'ensemble du Canada. On y trouvera également une discussion des motifs historiques et politiques qui ont conduit à la prédominance du mode de rémunération à l'acte, qui concerne 84% des médecins canadiens (Ferrall, Gregory & Tholl, 1998). Pour le cas Européen, voir, par exemple, la synthèse théorique d'Abel-Smith & Mossialos (1994) et la description des institutions du système de

tions institutionnelles gouvernent la contrainte budgétaire des médecins du Québec, le long de laquelle sont choisis les comportements de pratique optimaux. A titre d'analyse théorique préliminaire, la prochaine section propose un aperçu des effets attendus des instruments d'incitation sur ces choix.

2.1.1 Modes de rémunération et comportements de pratique : un survol

Comme nous l'avons souligné plus haut, une importante littérature empirique atteste de la sensibilité de la pratique des médecins au volume de rémunération. Ces travaux, qui empruntent pour la plupart aux méthodes d'analyse de l'offre de travail, reposent sur une simulation du salaire horaire basée sur le rapport entre le revenu total et le nombre d'heures travaillées. Les schémas de rémunération qui sont à l'origine de ce revenu total présentent cependant une très grande diversité. Ils sont traversés, en particulier, par la distinction importante établie par la littérature d'économie du travail (voir, par exemple, la synthèse de Lazear, 1995) entre rémunérations fixes et variables. Au-delà du montant du revenu, ces modes de rémunération influencent considérablement les choix de pratique. Cette section présente une courte synthèse des propriétés les plus connues de chacun d'entre eux.

Une rémunération variable consiste à verser un paiement proportionnel à une mesure de performance vérifiable. En matière médicale, les unités de mesure les plus couramment utilisées sont soit l'acte délivré – *rémunération à l'acte* – soit le nombre de patients traités – *capitation*. Bien que ces mesures de performances introduisent des différences importantes, ces modes de rémunération présentent donc tous deux les propriétés essentielles d'une rémunération à la pièce.[13] Ils en partagent par conséquent les qualités,

santé français fournie par M. Duriez (2000), *Le système de santé en France*, Rapport du Haut Comité de Santé Publique.

[13] A notre connaissance, le modèle de Selden (1990) est la seule analyse théorique explicitement

comme les défauts.

Les méthodes de rémunération à la pièce sont réputées pouvoir réconcilier les intérêts du principal (les autorités qui l'administrent dans le cas de la politique de santé) et ceux de l'agent (le médecin) par la dépendance de la rémunération sur la production (Prendergast, 1999). Lorsque la mesure de performance constitue une mesure adéquate des intérêts du principal et de l'activité de l'agent, ces schémas de rémunération permettent alors une amélioration de la performance (Lazear, 2000a). Dans le cas de la rémunération des médecins, l'étude de Hemenway, Killen, Cashman & *al.* (1990) établit ainsi que le passage à une rémunération à la performance, fondée en l'occurrence sur le revenu généré par l'activité du médecin pour son hôpital d'appartenance, permet d'augmenter significativement (12% ici) le nombre de patients traités. Au-delà des ces principes généraux, les spécificités de la pratique médicale peuvent en partie contrarier l'efficacité de ces modes de rémunération. Ces difficultés proviennent des asymétries d'information qu'engendrent les activités du médecin – avec l'organisme qui le rémunère comme avec le patient – et du caractère multi-dimensionnel de son activité.

Une première source d'asymétrie d'information provient de l'information cachée dont dispose le médecin dans sa relation avec le patient (Arrow, 1963). Le médecin est en effet le seul capable de juger à la fois de l'adéquation des soins aux affections dont souffre le patient et du diagnostic de ces affections elles-mêmes. Le médecin se trouve donc en position de manipuler la demande de soins en multipliant les prescriptions au-delà de ce qu'exige la préservation de la santé. En raison de cette "demande induite", la demande de soins qui s'adresse aux médecins est donc endogène (Evans, Parish & Sully, 1973). Dans ce cas, la mesure de performance sur laquelle est fondée la rémunération devient pour les médecins un instrument de maîtrise de leur propre revenu. Bien que ce mécanisme soit désormais théoriquement bien connu (voir, par exemple, De Jaegher

consacrée à la rémunération par capitation. Hutchison, Birch, Hurley & *al.* (1996) proposent une comparaison empirique entre ces deux types de rémunération à la pièce ; Gosden, Forland, Kristiansen & *al.* (2001) comparent l'efficacité de la capitation à celles de rémunérations fixes telles que le salaire.

& Jegers (2000) pour un modèle récent fondé sur le comportement du médecin), sa pertinence empirique a suscité d'intenses débats.[14]

Il semble cependant qu'un consensus se dégage pour admettre l'existence d'une demande induite,[15] compte tenu du faisceau convergent de confirmations empiriques qu'elle a reçues dans le cas du Canada (Schaafsma, 1994), de la France (Delattre & Dormont, 2003) et des Etats-Unis. Pour ce dernier cas, Gruber & Owings (1996) utilisent par exemple l'expérience naturelle offerte par le déclin de la fertilité pour évaluer la réaction des obstétriciens à une baisse exogène de la demande de soins. Les résultats confirment une induction de la demande de la part des médecins, puisque la baisse de la demande s'accompagne d'un déplacement des prescriptions vers des soins plus coûteux – donc plus rémunérateurs sous un régime de rémunération à la pièce – tels que les césariennes. La Norvège constitue à cet égard une exception persistante, puisque les travaux qui lui sont consacrés concluent systématiquement à l'absence d'induction de la demande (Carlsen & Grytten, 1998 ; 2000 ; Sørensen & Jostein, 1999 ; Grytten & Sørensen, 2001). Quoi qu'il en soit de son universalité, la demande induite constitue au minimum une éventualité que les médecins ajustent l'intensité de l'activité médicale

[14]Pour ne citer que les plus intenses, on pourra consulter à ce sujet les doutes émis par Feldman & Sloan (1988) et la réponse de Rice & Labelle (1989) ainsi que les débats qui opposent Labelle, Stoddart & Rice (1994a, 1994b) et Culyer & Evans (1996) à Pauly (1994a, 1994b).

[15]D'un point de vue méthodologique, le scepticisme le plus fondé ne peut qu'être ébranlé par la réaction de Fuchs (1986) à ces critiques, qui s'attend à ce que les «*economists will react to the study and the critique as they have in the past on this issue, with the fervant hope that maybe there is no inducement. This reaction has always reminded me of the story of the Frenchman who suspected that his wife was unfaithful. When he told his friend that the uncertainty was ruining his life, the friend suggested hiring a private detective to resolve the matter once and for all. He did so, and a few days later the detective came and gave his report : "One evening when you were out of town I saw your wife get dressed in a slinky black dress, put on perfume, and go down to the local bar. She had several drinks with the piano player and when the bar was closed they came back to your house. They sat in the living room, had a few more drinks, danced, and kissed." The Frenchman listened intently as the detective went on : "Then they went upstairs to the bedroom, they playfully undressed one another, and got into bed. Then they put out the light and I could see no more." The Frenchman sighed "Always that doubt, always that doubt".*

de façon à manipuler le revenu tiré de la pratique. Cette manipulation ne saurait avoir lieu si la rémunération était indépendante de l'intensité de l'activité, et seules les rémunérations à la pièce présentent par conséquent le risque d'y être sujettes (Grytten & Sørensen, 2001).

Outre le volume de soins délivrés, les médecins peuvent également manipuler les taux de rémunération à la pièce qui s'appliquent à ces soins. Dans la plupart des systèmes de santé, les taux de rémunération à la pièce varient en effet selon la nature des actes de façon à refléter leur difficulté de réalisation ainsi que leur efficacité en termes de santé. A cet égard, les médecins sont les seuls à connaître la nature des soins prodigués aux patients. La qualification des actes délivrés, qui sert de base à leur rémunération, est donc en général laissée à leur discrétion. Les médecins peuvent alors exploiter cette seconde information privée, dans la relation avec le principal qui décide de leur rémunération (Etat, direction du service, de l'hôpital,...), en falsifiant dans leurs déclarations les soins effectivement délivrés. Ce risque de sur-facturation (*billing-creep*) limite la possibilité de différencier les tarifs en fonction de la nature des actes (Evans, 1983). Il constitue donc une contrainte importante sur la capacité du principal à transmettre à l'agent, par l'intermédiaire des taux de rémunération, la hiérarchie de ses intérêts quant à l'importance des actes.

Compte tenu de la complexité de l'activité médicale, cette difficulté rend particulièrement délicate le choix des taux de rémunération à la pièce. La pratique médicale s'assimile en effet assez largement à une situation multitâche dont le diagnostic, la quantité de soins, la qualité des soins, le coût des soins et leur adéquation ou encore la gestion des établissements de santé sont autant de dimensions. Pour que la rémunération à la pièce soit efficace il convient alors que les taux de rémunérations relatifs reflètent la structure de priorités du principal (Holmstrom & Milgrom, 1991). Si des contraintes s'imposent à la différenciation des taux de rémunération, la rémunération à la pièce peut biaiser les choix de pratique dans un sens opposé à ses souhaits. Surtout, le système d'incitation encourage dans ce cadre l'abandon des activités pour lesquelles

il n'existe pas de mesure de performance vérifiable. Il est ainsi fondé théoriquement (Stiglitz, 1975) comme empiriquement (Paarsch & Shearer, 1999 ; Shearer & Paarsch, 2000) que les systèmes de rémunération à la pièce tendent à encourager la quantité au détriment de la qualité, plus difficilement mesurable. Bien que peu d'études empiriques soient consacrées à cette question, la pratique médicale ne semble pas faire exception à cette propriété (Jencks, Cuerdon, Burwen & *al.*, 2000).

En raison des nombreuses difficultés que posent la conception de rémunérations à la pièce, un principe de rémunération fixe leur est parfois préféré. Ce mode de rémunération consiste en général en un salaire constant, obtenu quel que soit le comportement de pratique dès lors que les termes du contrats (temps de présence dans l'établissement, participation aux réunions, ...) sont respectés. A l'inverse des rémunérations à la performance, ce schéma tend donc à déconnecter les choix de pratique de la rémunération. Elle laisse en conséquence les médecins libres de diversifier leurs activités, et rend non coûteux l'investissement en qualité. A partir de données d'enquête sur les médecins canadiens, Ferrall, Gregory & Tholl (1998) montrent ainsi que les médecins salariés consacrent 5.5 heures de plus par semaines à leur travail que les médecins rémunérés à la pièce, alors même qu'ils consacrent 5.9 heures de moins aux soins des patients.

A cette diversification s'ajoute un accroissement de l'attention consacrée aux soins, classiquement interprétée comme une mesure de qualité (Glazer & McGuire, 1993). La synthèse des résultats empiriques opérée par Gosden, Pedersen & Torgerson (1999) établit par exemple que la rémunération par un salaire est associée à des consultations plus longues ainsi qu'un nombre d'actes par patient et de patients par médecin moindre. L'envers de cette amélioration de la qualité reste cependant l'inévitable accroissement des coûts associé à un système de rémunération qui néglige la performance. Malgré les résultats de Gosden, Sibbald, Williams & *al.* (2003), qui montrent que le passage à une rémunération fixe a été sans effet sur la productivité des médecins généralistes en Angleterre, la très grande majorité des travaux empiriques soulignent la diminution dans le volume de soins associée à l'instauration d'un salaire fixe (Gosden, Forland,

Kristiansen & *al.*, 2001 ; Gaynor & Gertler, 1995).

Le choix entre rémunérations fixes et variables est donc gouverné par un arbitrage entre le volume de soins délivrés – éventuellement au-delà de ce que requiert l'amélioration de la santé – et la qualité de la pratique. Comme le soulignent Ma & McGuire (1997), cette tension entre les différents modes de rémunération peut s'interpréter comme une insuffisance du nombre d'instruments utilisés au regard du nombre d'objectifs poursuivis. Si l'efficience de la pratique médicale dépend à la fois de la quantité de soins et de leur qualité, il convient en effet que ce double objectif soit servi par au moins deux instruments. Un certain nombre de travaux se sont en conséquence tournés vers les modes de rémunération qui combinent des rémunérations fixe et variable.

Lorsque la demande de soins est excédentaire, Ma (1994) et Rogerson (1994) montrent ainsi que l'efficience, définie selon ces deux dimensions, nécessite que la rémunération soit une combinaison linéaire entre remboursement prospectif – enveloppe prévisionnelle, indépendante des soins effectifs – et remboursement des coûts. Bien qu'ils se dotent d'un instrument supplémentaire, ces modes de rémunération n'échappent pas aux difficultés liées aux asymétries d'information inhérentes à la pratique médicale. Même dans les cas où l'efficience l'exigerait, il est ainsi impossible d'utiliser une rémunération variable négative, au risque que le volume d'actes délivrés soit falsifié par le médecin – pour qui la rémunération variable devient un coût – en accord avec le patient – qui reste redevable de la partie des soins qui n'est pas couverte (Ma & McGuire, 1997). L'existence d'une partie variable maintient en outre la dépendance de la rémunération sur le volume de soins délivrés. Cette propriété perpétue par conséquent le risque de demande induite, bien que l'association à une rémunération fixe permette d'en diminuer l'importance (Levaggi & Rochaix, 2003).

Outre la combinaison de plusieurs instruments, offrir un menu de modes de rémunération alternatifs peut également permettre de renforcer l'efficacité des incitations par un effet de sélection (Encinosa, Gaynor & Rebitzer, 1997 ; Barro & Beaulieu, 2003).

La réponse des comportements de pratique aux incitations offertes dépend en effet de façon importante de caractéristiques individuelles inobservables, telles que la compétence (Dranove, 1988). Sous cette hypothèse d'hétérogénéité, offrir un menu de modes de rémunérations permet alors d'instaurer une auto-sélection des médecins (Demange & Geoffard, 2003) par laquelle le choix du mode de rémunération révèle ces caractéristiques inobservables.

La rémunération des médecins du Québec mobilise chacun des instruments décrits dans cette section. Les résultats théoriques présentés offrent donc un premier aperçu de leurs effets attendus et permettent, en particulier, d'évaluer l'adéquation du dispositif de rémunération mixte aux objectifs qui ont motivé son instauration.

2.1.2 Le règne de la rémunération à l'Acte

L'impulsion initiale de la rémunération mixte est née de la volonté de rééquilibrer les effets pervers de la rémunération à l'acte. La répartition des sources de revenu des médecins du Québec lève toute ambiguïté sur sa prédominance. La rémunération à l'acte représente en effet 80.57% des revenus des praticiens exerçant au Québec en 1996 et cette proportion reste stable jusqu'en 1999. Les 19.43% restants se répartissent entre les autres types de rémunération, très largement minoritaires, que sont les salaires (0.6%), les vacations (9.23%) – qui rémunérent des heures de travail ponctuelles dans un établissement – et les rémunérations provenant d'activités en laboratoire (9.6%).[16]

Contrairement au système américain où la rémunération des actes est un prix de marché qui varie selon les praticiens, celle-ci résulte, au Québec, de négociations entre le gouvernement provincial et les organisations professionnelles. Les prix sont donc exogènes du point de vue des médecins. Outre l'administration du mode de rémunération, le revenu des médecins du Québec a également fait l'objet de nombreuses mesures vi-

[16]La capitation n'est pas utilisée au Québec.

sant tant la réduction des coûts que l'amélioration des soins offerts à la population. Contrairement aux autres provinces, les ressortissants du Québec rencontrent en effet des barrières linguistiques et culturelles importantes à la mobilité. Cette caractéristique a permis la mise en oeuvre de mesures drastiques de maîtrise du revenu, qui expliquent en partie la faiblesse du revenu moyen des médecins Québecois en comparaison des autres provinces (Ferrall, Gregory & Tholl, 1998).

Le souci de maîtrise des coûts a en particulier conduit, dès 1976 et pour la première fois au Canada, à imposer un système de plafonnement des rémunérations. Une fois le montant mensuel du plafond atteint, ce système consiste à amputer les revenus de pratique de 75% de leur valeur. Le niveau des plafonds a fait l'objet d'ajustements constants, en réponse à l'évolution du pouvoir d'achat et aux spécificités des spécialités de pratique. Ainsi, les revenus de pratique en cabinet privé sont réduits de 35% (75% pour la radiologie diagnostique) avant application du plafond, afin de prendre en compte les charges liées aux frais professionnels. Pendant la période couverte par notre étude, les plafonds étaient fixés à 128 750$[17] par semestre pour toutes les spécialités à l'exception de la neurologie (142 000$), de l'endrocrinologie (103 500$), et de la pédiatrie (105 000$) jusqu'au premier trimestre 2001. Leur montant a ensuite été porté à 140 000$ pour toutes les spécialités à l'exception de la pédiatrie (115 000$). Pour l'ensemble des spécialités, les revenus provenant des services d'urgence étaient, jusqu'au premier trimestre 2000, exclus du revenu admissible au plafond. Cette mesure est étendue, depuis cette date, à l'ensemble de la pratique exercée en hôpital. Quoiqu'elles s'adaptent aux spécifités de pratique, ces dispositions peuvent être considérées comme très contraignantes. A titre de comparaison, les mesures de plafonnement instaurées en Ontario en 1993 réduisent seulement d'un tiers les revenus de pratique annuels qui dépassent 400 000$ (soit un plafond près de trois fois supérieur à celui qui s'applique à la plupart des spécialités).[18]

[17]Tous les montants monétaires sont exprimés en Dollars Canadiens.

[18]Ces dispositions de plafonnement des rémunérations sont peu utilisées ailleurs dans le monde. Les rares études qui leur sont consacrées tendent à montrer que l'effet des plafonds sur les choix des médecins s'apparente à celui qu'a une taxe sur le revenu des contribuables (Kralj, Kantarevic &

Au-delà de la maîtrise des coûts, les autorités québécoises ont également mené une politique active de réduction des inégalités régionales en termes de densité médicale. Dans ce but, un taux de rémunération différenciée déforme, depuis 1982, le prix payé pour les actes en fonction de différentes caractéristiques de pratique telles que la spécialité, la région administrative et la ville d'exercice. Cette mesure créé d'importants écarts de rémunération, puisque, s'il pénalise l'exercice dans les régions à forte densité médicale, le taux de rémunération différenciée accroît le prix payé dans les zones défavorisées. Ces distortions importantes dans les prix des services se sont avérées efficaces pour encourager l'installation dans les zones à faible densité médicale (Bolduc, Fortin & Fournier, 1996).

L'ensemble de ces mesures crée une déconnexion importante entre le revenu qui devrait résulter des choix de pratique – appelé *consommation potentielle* – et le revenu effectivement touché par les médecins, qui constitue leur *consommation effective*. Pour en tenir compte, notre approche consiste à modéliser l'ensemble de ces mesures. A ce titre, il faut noter que les mesures de plafonnement s'appliquent à l'ensemble du revenu après application du taux de rémunération différenciée. En notant $\widetilde{X}_i$ le revenu potentiel du médecin i[19] ; τ_i le taux de rémunération différenciée induit par ses caractéristiques individuelles ($\tau_i > 1$ dans les zones subventionnées, $\tau_i < 1$ sinon) et C_i le niveau du plafond au-delà duquel le revenu est diminué de 75%, la consommation effective, X_i, d'un médecin du Québec est donc :

$$X_i = min\left[\widetilde{X}_i, C_i\right] + max\left[0.25\ (\widetilde{X}_i - C_i), 0\right] + \tau_i \widetilde{X}_i \tag{2.1}$$

Ces mesures s'appliquent quelles que soient les dispositions qui président à la rémunération de la pratique. En particulier, elles s'appliquent dans les mêmes termes aux médecins rémunérés selon le mode de rémunération mixte, introduit en 1999.

Weinkauf, 2005). S'y ajoute cependant un effet de demande induite lié à la réduction du taux au-delà du plafond (Nassiri & Rochaix-Ranson, 2000).

[19] Afin de simplifier les expressions algébriques, nous omettons la dépendance des variables sur le temps aussi souvent que cela ne crée pas d'ambiguïté.

2.1.3 Le mode de Rémunération Mixte

Contrairement au système de rémunération à l'acte, qui se limite à rémunérer les soins délivrés selon un prix indexé sur la nature et la difficulté des actes, la rémunération mixte repose sur deux composantes : un taux partiel de rémunération à l'acte, qui rémunère les soins délivrés à un taux réduit (en comparaison du taux sous la rémunération à l'acte "pure", présentée ci-dessus) ; mais aussi un *per diem*, rémunération fixe assise sur un large éventail d'activités.

a) Objectifs et dispositions

L'instauration de la rémunération mixte, à compter du quatrième trimestre 1999, a fait l'objet d'une étroite collaboration entre les autorités provinciales et les organisations professionnelles. Les objectifs qui ont présidé à son instauration reflètent donc à la fois le souci de maîtrise des coûts et d'amélioration des soins et des exigences issues de l'expérience de pratique.

Cette réforme visait d'abord à donner aux médecins les moyens de consacrer une partie de leur temps aux charges administratives et à l'enseignement. Bien qu'elles constituent un élément essentiel du fonctionnement du système de santé – en termes de circulation de l'information sur les patients, de gestion des établissements et de transmission des connaissances aux nouvelles générations de médecins – ces activités sont en effet exercées à titre exclusivement bénévole sous la rémunération à l'acte. Un objectif connexe était donc de rétablir une certaine équité entre les médecins, qu'ils consacrent ou non une partie importante de leur temps à ces activités.

Pour tenir compte de cette hétérogénéité dans les choix de pratique, les autorités ont choisi d'assortir l'introduction de la rémunération mixte d'une assez grande flexibilité.

Plutôt qu'un nouveau mode de rémunération universel et obligatoire, la rémunération mixte est en effet une alternative à laquelle les praticiens peuvent librement adhérer. Ainsi, après 1999, les médecins dont l'activité est très largement consacrée aux activités cliniques peuvent choisir de conserver la rémunération à l'acte ; tandis que les praticiens qui privilégient les activités non cliniques peuvent librement opter pour la rémunération mixte.[20] Sous la rémunération mixte, ces activités sont en effet rémunérées par un salaire fixe, appelé *per diem*. Lorsque leurs heures de travail sont couvertes par un *per diem*, les médecins peuvent exercer diverses activités *admissibles* incluant l'enseignement, les activités administratives et les activités cliniques.[21]

Le *per diem* rémunère sans distinction toutes les activités admissibles, y compris les activités cliniques. Les activités cliniques présentent pourtant une très grande diversité, qui se manifeste, notamment, dans la difficulté des actes et dans le temps qui leur est consacré. Les organisations professionnelles ont donc fait valoir la nécessité d'une rémunération spécifique afin de garantir la continuité des soins. Dans ce but, une rémunération à l'acte s'ajoute au *per diem*.

Plutôt qu'une rémunération à taux plein, les autorités ont cependant choisit une rémunération à l'acte partielle, qui rémunère les actes pratiqués à un taux réduit en comparaison du taux en vigueur sous la rémunération à l'acte. Cette réduction du

[20]Plus précisément, l'adhésion à la rémunération mixte requiert l'unanimité des médecins appartenant à une unité médicale (en général un service). Nous sommes contraints d'ignorer cet aspect faute d'information sur l'établissement d'appartenance. Dans ce qui suit, nous faisons donc l'hypothèse que les médecins ont la possibilité de recourir au "vote par les pieds", et de changer de service en fonction de leurs préférences de rémunération. Cette hypothèse semble assez conforme à la pratique. Le vice-président du CMQ soulignait ainsi en Novembre 2000 que «*those specialties which depends on physicians spending large amounts of time with their patients and especially the university hospitals and pediatric hospitals are keen to adopt the new system*». (cité par S. Benady "Mixed payment a go in Quebec", *The Medical Post*, 7 Novembre 2000).

[21]La recherche, exclue des activités admissibles, constitue une exception importante. Elles sont considérées comme directement rémunérées par les établissements (en général les hôpitaux) où elles sont exercées.

taux des actes est destinée à améliorer la qualité des soins en accroissant le temps consacré à chacun. La combinaison du *per diem* à ce taux réduit devrait en outre participer à atténuer le risque qu'une induction de la demande accompagne cette baisse de rémunération.

En raison de ces dispositions, la rémunération mixte a attiré diversement les médecins en fonction de leur spécialité. Le Tableau 2.1 présente les taux d'adhésion à la rémunération mixte et les taux de rémunération des actes par spécialité en 2002, qui est la dernière année post-réforme de notre échantillon. Le taux de rémunération des actes est en moyenne diminué de 50% avec le passage à la rémunération mixte. La variété des types de pratique entre spécialités conduit cependant à une assez grande variabilité du taux de rémunération, oscillant entre 30% et 90%. L'adhésion à la rémunération mixte partage la population en deux sous-ensembles de tailles sensiblement égales (taux d'adhésion supérieur à 40% dans l'ensemble de la population), variant entre les spécialités de moins de 3% à près de 90%.

b) Consommation potentielle sous la rémunération mixte

Bien que la combinaison d'un *per diem* et d'un taux de rémunération à l'acte réduit soit la caractéristique essentielle de la rémunération mixte, ses modalités pratiques font intervenir de très nombreux critères qui complexifient considérablement son application. Cette section est consacrée à une description détaillée des dispositions adoptées, destinée à modéliser la consommation potentielle des médecins qui choisissent la rémunération mixte.

Lorsqu'un médecin choisit la rémunération mixte, un demi *per diem*, d'un montant D, lui est versé pour chaque tranche de $\overline{d} = 3.5$ heures de travail fournies. Le nombre maximum de demis *per diems* dont peut bénéficier un médecin est cependant limité à 28 par période de deux semaines, soit un *per diem* par jour ouvrable. Cette notion

Tableau 2.1 – Statistiques descriptives de la rémunération mixte

Spécialité	Taux d'adhésion (%)	Nombre total d'observations	Taux de rémunération moyen
Anesthésie-réanimation	55.0	122	0.5
Cardiologie	5.1	11	0.5
Chirurgie générale	36.5	121	0.6
Chirurgie orthopédique	13.4	84	0.7
Chirurgie plastique	3.2	21	0.6
Chirurgie thoracique	0.0	2	.
Dermatologie	28.6	44	0.5
Gastro-entérologie	5.9	34	0.8
Obstétrique-gynécologie	12.9	76	0.5
Pneumologie	20.3	24	0.8
Médecine interne	27.7	71	0.7
Physiatrie	64.6	16	0.5
Neuro-chirurgie	89.7	7	0.0
Neurologie	33.3	28	0.3
Ophtalmologie	6.9	40	0.6
Oto-rhino-laryngologie	21.3	37	0.6
Pédiatrie	65.2	120	0.3
Radiologie diagnostique	0.0	5	.
Radio-oncologie	79.3	13	0.8
Urologie	18.1	34	0.7
Chirurgie cardio-vasculaire	2.9	9	0.6
Néphrologie	18.7	19	0.3
Endocrinologie	42.7	28	0.3
Rhumatologie	62.8	19	0.3
Autres	69.6	235	0.4
Total	**40.9**	**1220**	**0.5**

Note. Le *taux d'adhésion* est mesuré par la proportion des individus d'une spécialité qui ont obtenu une partie de leur revenu sous la rémunération mixte. Le *taux de rémunération* correspond au rapport entre le taux de rémunération des actes (mesuré par l'indice de prix, voir Section 2.4.2) sous la rémunération mixte et le taux sous la rémunération à l'acte en 2002. La spécialité "Autres" regroupe des champs de pratique non reconnus par la Corporation professionnelle des médecins du Québec (CPMQ) telles que l'allérgie, l'immunologie clinique, l'anatomo-pathologie,... Les taux de rémunération sont manquants pour toute spécialité dont aucun professionnel n'a choisi la rémunération mixte.

est d'ailleurs au coeur du système, puisque seules les plages horaires d'une semaine traditionnellement ouvrée (du lundi au vendredi, de 7h à 12h et de 14h à 19h) sont admissibles au *per diem*.

En notant h le nombre d'heures de travail hebdomadaires d'un médecin qui choisit la rémunération mixte, le nombre de demis *per diems* versés par semaine est donc :[22]

$$N = \frac{\min\left\{floor\left(\frac{2.h}{d}\right), 28\right\}}{2} \tag{2.2}$$

Le montant versé dans le cadre d'un demi *per diem* est resté constant et fixé à $D = 300\$$ pendant l'ensemble de la période qui retient notre intérêt. Par la suite, il a été porté à $D = 308\$$ en Avril 2003, puis $335\$$ en Juillet 2003.

Outre le versement d'un salaire, la réclamation d'un *per diem* a par ailleurs des conséquences sur la rémunération des actes réalisés. Les actes sont en effet distingués selon qu'ils ont été pratiqués pendant des heures de travail couvertes par un *per diem*. Les actes réalisés en dehors d'un *per diem* (i.e. pendant des heures de travail non couvertes par un *per diem*) sont en effet rémunérés selon les conditions qui prévalent sous la rémunération à l'acte. Seuls les actes délivrés sous un *per diem* sont donc, à l'inverse, rémunérés à taux réduit. Ainsi, en notant P le prix versé pour un acte représentatif sous la rémunération à l'acte "pure", cet acte est rémunéré au prix P s'il est réalisé en dehors d'un *per diem* et au taux $(1-\alpha)P$, $\alpha < 1$, sinon. Le taux de réduction dans la rémunération des actes, α, dépend non seulement de la spécialité de pratique mais également de la nature des actes eux-même. Une nomenclature associe ainsi à chaque code d'acte un prix de rémunération à taux plein et un taux de réduction applicable si l'acte est réalisé sous un *per diem*.

[22] La fonction *floor* transforme un nombre décimal en sa partie entière. A titre d'illustration, on a : $floor(\frac{7}{2}) = floor(3.5) = 3$

Si la plupart des actes se voit attribuer un taux de réduction partiel, un certain nombre d'entre eux sont considérés comme étant rémunérés directement par le *per diem* et ne sont donc assortis d'aucune rémunération spécifique (on a donc $\alpha \in [0, 1]$). Cette caractéristique établit de fait une distinction importante entre les actes rémunérés sous la rémunération mixte – que nous appellerons *actes facturables*, notés AF – et les actes *non facturables*, notés ANF, pour lesquels les médecins n'ont aucune incitation sous la rémunération mixte.

Compte tenu de l'ensemble de ces mesures, la consommation potentielle d'un médecin varie donc considérablement selon le mode de rémunération choisi. On note d_i la variable binaire indiquant le mode de rémunération, $d_i = 1$ lorsque le médecin i a choisi la rémunération mixte. Pour tenir compte des différences de rémunération engendrées par le versement d'un *per diem*, nous distinguons les variables de pratique, $V = \{AF, ANF\}$, selon la période pendant laquelle elles ont été exercées. On note ainsi V^{RM} la variable de pratique V lorsqu'elle est réalisée sous un *per diem* et V^{RA} lorsqu'elle est réalisée en dehors d'un *per diem*. La consommation potentielle d'un médecin du Québec qui travaille W_i semaines par an s'écrit donc :

$$\begin{aligned} \widetilde{X}_i = d_i[W_i\ N_i\ D + (1-\alpha)P\ AF_i^{RM} + P\ (AF_i^{RA} + ANF_i^{RA})] \\ + (1-d_i)\ P\ (AF_i^{RA} + ANF_i^{RM}) \end{aligned} \tag{2.3}$$

La consommation réelle d'un médecin – c'est à dire le revenu de pratique qui lui est effectivement versé – dépend des mesures de maîtrise des revenus qui s'appliquent à lui. La contrainte budgétaire des médecins du Québec correspond donc à l'ensemble des équations (2.1) à (2.3). Le Tableau 2.2 propose une synthèse des dispositions décrites dans cette section.

Tableau 2.2 – Rémunération des médecins du Québec considérés dans l'analyse

<table>
<tr><th>RA</th><th colspan="2">RM</th></tr>
<tr><td>Pas de rémunération fixe
Heures non cliniques (h^o)
non rémunérées</td><td>Demi
Per diem :</td><td>- Rémunère chaque tranche de 3.5 h en établissement
- Toutes les heures de pratique sont admissibles (h^c, h^o)
- Plafonné à 28 toutes les 2 semaines</td></tr>
<tr><td rowspan="4">Actes
rémunérés
au prix P</td><td>Actes
Facturables :</td><td>- Rémunérés à $(1-\alpha)P$ pendant les heures *per diem*
- Rémunérés au prix P en dehors des heures *per diem*</td></tr>
<tr><td>Actes
Non facturables :</td><td>- Non rémunérés pendant les heures *per diem*
- Rémunérés au prix P en dehors des heures *per diem*</td></tr>
<tr><td colspan="2">Rémunération différenciée sur critères géographiques</td></tr>
<tr><td colspan="2">Plafonnement des rémunérations†</td></tr>
</table>

†A l'exception des activités en urgence jusqu'en 2001, et de toutes les activités en hôpital depuis. Voir Section 2.4.2.

c) Réalisations : un premier aperçu

La rémunération mixte repose sur un salaire fixe et un taux réduit de rémunération des actes, dont la combinaison est destinée à encourager une augmentation des heures de travail non clinique et une diminution du nombre d'actes réalisés par heure. Dans la mesure où l'adoption de ce nouveau mode de rémunération est un choix volontaire des médecins, un effet de sélection peut cependant amplifier cet effet d'incitation. Si le passage à la rémunération mixte repose sur une différence systématique de préférence entre les médecins, on peut en effet s'attendre à ce que cette différence influence également leur réaction aux incitations fournies sous la rémunération mixte. Par exemple, il est raisonnable de penser que des médecins qui ont de fortes préférences pour les activités non cliniques réagissent à une baisse du taux de rémunération des actes par une forte diminution des activités cliniques.[23]

[23]Nous proposons une illustration graphique de cette intuition dans la Section 2.2.2.

Tableau 2.3 – Statistiques descriptives de l'effet de la réforme

	Médecins ayant choisi la RM					
	Sous la RA		Sous la RM		Total	
	Moyenne	Ecart type	Moyenne	Ecart type	Moyenne	Ecart type
Heures hebdomadaires totales (h)	49.17	12.8	46.6	11.79	48.38	12.55
_____ Cliniques (h^c)	41.39	13.5	40.03	12.73	40.98	13.28
_____ Non-cliniques (h^{nc})	7.77	8.25	6.57	8.46	7.4	8.34
Semaines de travail (W)	45.55	4.62	45.43	4	45.52	4.44
Actes[a]	122.87	70.70	101.81	64.59	116.41	69.56
_____ Non Facturables (ANF)	28.42	462.94	19.33	41.17	25.64	44.98
_____ Facturales[b] (AF)	94.44	60.18	81.72	52.45	90.54	58.21
Revenu annuel[a] (X)	130.80	73.40	188.26	71.77	148.41	77.56

	Médecins constamment sous la RA					
	Avant la réforme		Après la réforme		Total	
	Moyenne	Ecart type	Moyenne	Ecart type	Moyenne	Ecart type
Heures hebdomadaires totales (h)	49.23	14.66	48.57	13.26	49.11	14.41
_____ Cliniques (h^c)	41.97	15.22	43.33	14.03	42.22	15.01
_____ Non-cliniques (h^{nc})	7.26	8.95	5.24	7.88	6.89	8.8
Semaines de travail (W)	45.15	5.29	45.18	3.88	45.16	5.06
Actes [a]	151.59	116.10	168.58	106.66	154.71	114.61
_____ Non Facturables (ANF)	54.09	64.07	58.97	77.79	54.98	66.83
_____ Facturables (AF)	97.50	104.23	110.67	92.74	99.92	102.34
Revenu annuel[a] (X)	165.42	96.05	224.02	117.71	176.18	102.91

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

[b]Inobservables pour les médecins ayant choisit la RM lorsqu'ils sont pratiqués pendant des heures couvertes par un *per diem*. Voir la Section (2.3.2).

Note. Profil de pratique moyen des médecins du Québec sur la période 1996-1998 et 2002. *Moitié supérieure* : Médecins ayant obtenu une partie de leurs revenus sous la rémunération mixte pendant la période d'observation, avant (partie gauche) et après (partie droite) l'avoir adoptée. *Moitié inférieure* : Médecins dont 100% du revenu provient de la rémunération à l'acte, avant (partie gauche) et après (partie droite) l'introdcution de la réforme.

Le Tableau 2.3 propose un premier aperçu de l'effet de la rémunération mixte sur les comportements de pratique.[24] Ces derniers sont décrits par le nombre d'heures hebdomadaires, consacrées respectivement aux activités cliniques (notées h^c) et non cliniques (h^{nc}), le nombre annuel de semaines de travail, le nombre d'actes pratiqués, distingués selon qu'ils sont facturables ou non sous la rémunération mixte, et, enfin, le revenu annuel. La moyenne et l'écart-type de chacune de ces variables sont calculés pour chacun des deux groupes de médecins créés par la réforme. La partie supérieure du tableau résume en effet le comportement de pratique des médecins qui sont passés à la rémunération mixte pendant notre période d'étude (1996-2002) ; la partie inférieure celui des médecins qui n'ont jamais abandonné la rémunération à l'acte pendant cette période. Avant l'introduction de la réforme, la comparaison entre les parties haute et basse du tableau permet donc d'apprécier l'ampleur de l'effet de sélection en comparant les choix des médecins selon leur groupe d'appartenance. A cet effet, nous nous intéressons, pour chaque groupe de médecins, au comportement de pratique sous la rémunération à l'acte (pendant la période qui précède l'introduction de la réforme pour les médecins qui sont restés à la rémunération à l'acte), sous la rémunération mixte (après la réforme pour ces mêmes médecins) et sur l'ensemble de la période. L'effet d'incitation émerge ainsi des comparaisons entre les deux premières colonnes du tableau.

En ce qui concerne l'effet de sélection, il semble que les médecins des deux groupes se distinguent moins par les heures de travail choisies que par la quantité d'actes pratiqués. Le nombre moyen d'heures hebdomadaires de travail (49.17 pour les médecins ayant choisit la rémunération mixte contre 49.23 pour ceux qui restent à la rémunération à l'acte) comme la répartition de ces heures entre les activités cliniques (41.39 contre 41.97) et non-cliniques (7.77 contre 7.26) sont en effet très similaires. A l'inverse, les médecins qui choisiront la rémunération mixte réalisent beaucoup moins d'actes que ceux qui restent à la rémunération à l'acte (122869 pour les premiers et 151591 pour les seconds). Alors que les actes facturables pratiqués sont très proches d'un groupe à l'autre (94440 contre 97503), cette différence se manifeste principalement dans les actes

[24] Les données utilisées pour construire ce tableau sont décrites dans la Section 2.4.1.

non facturables réalisés (28429 contre 54087). Il en résulte une importante différence de revenu, les médecins qui passeront à la rémunération mixte bénéficiant d'un revenu annuel significativement inférieur (130795$ contre 165422$).

Au total, les médecins qui choisiront la rémunération mixte se distinguent principalement par une faible quantités d'actes pratiqués, réalisant 18% d'actes de moins que les médecins qui conservent la rémunération à l'acte. Cette différence est presque entièrement imputable à un important écart de comportement en termes d'actes non-facturables. Sous les mêmes conditions d'incitation – la rémunération à l'acte – les médecins qui manifesteront leur préférence pour la rémunération mixte choisissent en effet une quantité d'actes facturables de 48% inférieure à celle des médecins qui la refusent. Ces résultats confirment l'existence d'un effet de sélection dans le passage à la rémunération mixte, fondé principalement sur des préférences divergentes vis-à-vis des actes pratiqués.

En utilisant les comportements de pratique moyens au sein de chaque groupe, un premier aperçu de l'effet d'incitation peut être obtenu par un estimateur de Différences en Différence (DD). Cet estimateur utilise le groupe de médecins qui ne sont pas affectés par la réforme comme un *groupe contrôle* des médecins qui passent à la rémunération mixte (*groupe traitement*). Sa validité repose donc sur l'hypothèse que le comportement des médecins du groupe contrôle reflète celui qu'auraient adopté les médecins du groupe traitement en l'absence de réforme (Heckman & Smith, 1995). L'effet de la réforme estimé correspond alors à la variation dans les différences de comportement, entre les médecins du groupe traitement et ceux du groupe contrôle, induite par la réforme. La dernière colonne du Tableau 2.4 présente l'estimateur DD et le t de Student associé pour chaque variable de pratique. Il correspond à la différence entre les deux premières colonnes du tableau, dans lesquelles sont calculées les différences entre les médecins des groupes contrôle et traitement avant et après la réforme.

La réforme semble rencontrer un succès mitigé en termes d'heures de travail. L'écart

Tableau 2.4 – Estimateurs de Différence en Différence

	Avant	**Après**	**DD**
Heures hebdomadaires totales	-.06 (-.231)	-1.97 (-4.838)	-1.9 (-3.66)
——— cliniques (h^c)	-.58 (-2.016)	-3.3 (-7.621)	-2.72 (-5.00)
——— non cliniques (h^{nc}	.51 (3.028)	1.33 (5.168)	.82 (2.53)
Semaines (W)	.4 (4.033)	.26 (2.071)	-.14 (-0.78)
Actes[a] totaux	-28721.63 (-13.796)	-66769.94 (-22.272)	-38048.32 (-9.72)
—— non facturables (ANF)	-25658.49 (-21.928)	-39631.22 (-18.485)	-13972.73 (-7.45)
——— facturables (AF)	-3063.13 (-1.646)	-28946.78 (-11.223)	-25883.65 (-6.04)
Revenu annuel[a] (X)	-34626.44 (-19.599)	-35756.15 (-10.796)	-1129.72 (-0.32)

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. *Deux premières colonnes* : Différence dans les profils de pratique moyens entre les médecins qui ont choisi la rémunération mixte (moitié supérieure du Tableau 2.3) et ceux qui sont restés à la rémunération à l'acte (moitié inférieure du Tableau 2.3), avant et après la réforme. *Dernière colonne* : Différence en Différence, *i.e.* différence entre les deux premières colonnes. Entre parenthèses, t de Student des différences de moyennes.

de comportement entre les médecins du groupe traitement et ceux du groupe contrôle se creuse après la réforme en termes d'heures de travail tant clinique que non-clinique. Pourtant, si cet écart se creuse positivement pour les heures non-cliniques, suggérant un effet positif de la réforme (qui incite les médecins à augmenter le temps consacré à ces activités de 0.8 heures), il est en effet négatif pour les heures de travail clinique. La réforme tend à engendrer une diminution de 2.7 heures du temps de travail consacré à cette activité. Il en résulte un effet net, sur les heures de travail totales, négatif, puisque les heures de travail clinique diminuent plus que les heures non-cliniques n'augmentent suite à la réforme.

Il n'est pas possible, à ce stade de l'analyse, de tirer des conclusions des observations

sur les choix d'actes non-facturables et, par conséquent, sur les choix d'actes totaux. Pendant les heures de travail couvertes par un *per diem*, les actes non-facturables sont en effet inobservables par définition. Les niveaux d'actes non-facturables présentés dans les tableaux proviennent donc, sous la RM, de la pratique exercée en dehors d'un *per diem* – sous la rémunération à l'acte avant la réforme, pendant des heures de travail rémunérées à l'acte après – et constituent, par conséquent, une borne inférieure des actes effectivement pratiqués. Au regard des actes facturables, observés en toutes circonstances, la réforme semble cependant avoir un fort impact négatif puisqu'elle provoque une diminution de 27% des actes pratiqués. Combinée à une diminution moins que proportionnelle des heures cliniques (6%), cette variation suggère un accroissement du temps consacré à chaque acte.

En résumé, le passage à la rémunération mixte s'appuie sur un effet de sélection important mais limité aux actes facturables. A heures de travail clinique identiques, les médecins qui choisissent la rémunération mixte pratiquent systématiquement moins de ces actes que ceux qui conservent la rémunération à l'acte. Si l'on interprète le temps consacré à chaque acte comme un critère de qualité des soins prodigués, la rémunération mixte tend donc à attirer des médecins qui valorisent relativement plus la qualité des soins. A cet effet de sélection s'ajoute un effet d'incitation, plus particulièrement marqué quant aux heures de travail et aux actes facturables. Les actes facturables connaissent une diminution importante suite au passage à la rémunération mixte. En termes d'heures de travail, enfin, la réforme influence à la fois la quantité d'heures travaillées et leur répartition entre les types d'activités. Si la rémunération mixte parvient à encourager une augmentation modérée des heures non-cliniques, elle se traduit simultanément par une diminution importante des heures de travail clinique. Il en résulte une diminution non négligeable des heures de travail totales.

Si les conditions qui assurent la validité de l'estimateur de Différence en Différence sont respectées,[25] ces résultats correspondent à l'effet de la réforme sur les médecins qui

[25]L'hypothése fondamentale est que seule la réforme est susceptible de modifier les différences de

ont choisi d'y adhérer, ou encore à l'effet de *traitement sur les traités* (Heckman, 1997). Si l'effet de la réforme est hétérogène – au sens où les caractéristiques inobservables des médecins influencent leur sensibilité à la réforme – ils offrent donc une compréhension assez restreinte de ses effets en se limitant à un sous-ensemble de la population. Pour y remédier, nous proposons une estimation structurelle des préférences des médecins, permettant de prédire la réaction de l'ensemble de la population aux variations des incitations. Ce modèle intègre en particulier les possibilités d'arbitrage entre la qualité et les quantités de soins délivrés. La prochaine section présente une version simplifiée du modèle (décrit en détail dans la Section 2.3) qui met en évidence les conséquences de cet arbitrage.

2.2 Analyse théorique du passage à la Rémunération Mixte

Comme l'a fait apparaître la synthèse proposée dans la Section 2.1.1, le choix entre une rémunération fixe et une rémunération variable se résume assez largement à un arbitrage entre l'intensité de l'activité médicale (qui assure une allocation efficace des ressources consacrées aux soins de santé) et la qualité des soins délivrés (qui détermine l'amélioration de la santé permise par un volume donné de moyens). Le recours à des modes de rémunération mixtes tente de tirer parti de chacun de ces avantages. Le modèle présenté dans cette section en évalue la possibilité en intégrant dans l'analyse l'arbitrage que réalisent les médecins entre marges intensive (qualité) et extensive (quantité) en réaction aux variations dans les incitations.

comportement entre les groupes contrôle et traitement. Voir Bertrand, Duflo & Mullainathan (2004) pour une présentation critique, qui montre en outre que les écart-types estimés par cette méthode sont non-convergents en présence d'auto-corrélation.

2.2.1 Modélisation du comportement des médecins

L'hypothèse retenue quant aux déterminants du comportement des médecins oriente de façon cruciale l'analyse théorique de la sensibilité des choix de pratique aux incitations offertes. Le choix de la fonction objectif peut en effet déterminer entièrement le changement dans les comportements de pratique induit par une variation des incitations. [26] Ainsi, comme le soulignent McGuire & Pauly (1991), l'hypothèse traditionnelle de maximisation du profit prédit de façon certaine une réduction des actes pratiqués en réponse à une diminution de leur taux de rémunération. A l'inverse, l'hypothèse de revenu-cible – selon laquelle les médecins ajustent leurs choix de pratique de façon à maintenir leur revenu à un niveau désiré – implique un accroissement du nombre d'actes suite à une réduction du taux de rémunération (Rice, 1983). Au regard de ces prédictions, la maximisation d'utilité constitue une hypothèse très générale puisqu'elle se réduit soit à la maximisation du profit, soit à la recherche d'un revenu-cible, en fonction de l'importance relative des effets revenu et substitution (McGuire & Pauly, 1991). Nous nous en remettons donc aux comportements observés pour discriminer entre ces hypothèses de comportement, et nous analysons les comportements de pratique qui résultent de la maximisation d'utilité des médecins.

Afin de simplifier l'analyse, nous nous limitons aux choix de pratique journaliers. Les activités cliniques sont décrites par le nombre d'heures qui leur sont consacrées par jour, h^c,[27] et le nombre d'actes, A, pratiqués.[28] Le nombre d'actes par heure, $e = A/h^c$, est donc une variable endogène du modèle, qui peut être interprétée comme une forme d'effort. A nombre d'heures cliniques donné, accroître le nombre d'actes

[26] McGuire (2000) propose une discussion très complète des résultats existants quant aux motivations des médecins.

[27] A la différence des autres sections, toutes les notations font référence, ici, aux choix de pratique journaliers.

[28] L'objectif de l'analyse est de mettre en évidence l'influence sur les choix de pratique de l'arbitrage entre marges intensive et extensive. Nous négligeons donc la distinction entre actes facturables et non-facturables, qui concerne les profils de substitution au sein de la marge intensive.

pratiqués requiert en effet une plus grande rapidité d'exécution et correspond donc à une augmentation de l'intensité des heures de travail. Simultanément, cette augmentation conduit également à une diminution du temps consacré à chaque acte ainsi qu'à une augmentation des soins prodigués à la population. Ces multiples effets sont résumés par la fonction de production du médecin, que nous supposons être le niveau de santé atteint par la population, s. Interprétant le temps consacré à chaque acte comme un critère de qualité des soins, nous supposons en effet que la santé est une fonction décroissante de l'effort. A effort donné, le nombre d'actes est un indicateur de la quantité de soins prodigués et est donc supposé accroître la santé. Au total, la fonction de production des médecins est donc : $s = s(\underset{+}{A}, \underset{-}{e})$.

Adoptant une hypothèse devenue classique dans la littérature (Dranove, 1988 ; Rochaix, 1989), nous supposons que les activités cliniques affectent le bien-être des médecins par l'intermédiaire de cette fonction de production. Cette hypothèse peut s'interpréter comme une norme éthique, par laquelle les médecins internalisent l'objectif d'amélioration de la santé dans la population (Arrow, 1963 ; Evans, 1974).

Le temps journalier laissé libre par les activités cliniques, $T - h^c$, est réparti entre le loisir pur, l, et les heure de travail non-clinique, h^{nc}. Ces dernières recouvrent l'ensemble des activités, telles que l'enseignement ou les tâches administratives, qui ne sont pas rémunérées sous la rémunération à l'acte. Malgré cette absence d'incitations, les médecins consacrent de fait une partie de leur temps aux activités non-cliniques sous la rémunération à l'acte (voir le Tableau 2.3). Cette observation révèle donc un goût pour les activités non-cliniques, au sens où celles-ci accroissent l'utilité (en diminuant la pression des pairs, en accroissant la satisfaction au travail, ...) alors même qu'elles laissent la consommation, X, inchangée. Cet aspect est pris en compte en modélisant les les heures de travail non-clinique comme une forme particulière de loisir.

Au total, la fonction d'utilité des médecins s'écrit donc : $U = U(X, l, h^{nc}, s)$. Au Québec – comme dans de nombreux pays industrialisés – les soins de santé sont très

largement pris en charge par les institutions de mutualisation des risques telles que l'assurance sociale, les mutuelles, etc... Dans ce contexte, la demande de soins de santé est déconnectée des variations de prix, et ne résulte que des besoins réels de la population. Nous supposons par conséquent que la sensibilité des choix de pratique aux variations de rémunération résultent exclusivement de la maximisation d'utilité sous contrainte budgétaire, et que les médecins peuvent donc, en particulier, allouer librement leur temps entre les différents types d'activités.

La contrainte budgétaire dépend à la fois des variables de pratique et du mode de rémunération adopté. On note ainsi y le revenu obtenu indépendamment du nombre d'actes pratiqués, correspondant donc au *per diem* sous la rémunération mixte[29], $y = D$, et fixé à $y = 0$ sous la rémunération à l'acte. Le revenu hors-travail – non affecté par la réforme – est supposé être nul. La contrainte budgétaire journalière d'un médecin peut donc être schématiquement résumée par $X = fA + y$, où f désigne le taux de rémunération des actes, égal à P sous la rémunération à l'acte et $(1-\alpha)P$ sous la rémunération mixte, $\alpha \in [0,1]$. Les variables décrivant les comportements de pratique – en ignorant les solutions de coin – sont donc les solutions du programme :

$$\begin{aligned} \underset{\{X,l,h^o,A,e\}}{Max} \quad & U = U\left(X, l, h^{nc}, s\left(A, e\right)\right) \\ s.c. \quad & (i) \quad T = h^{nc} + l + h^c \\ & (ii) \quad A = eh^c \\ & (iii) \quad X = fA + y \end{aligned}$$

En remplaçant les heures cliniques et le nombre d'actes par les valeurs imposées respectivement par les contraintes (i) et (ii), la fonction d'utilité des médecins s'écrit encore $U\left(X, l, h^o, s\left((T - h^o - l)e, e\right)\right)$ soit, en forme réduite : $\widetilde{U}\left(X, l, h^o, e\right)$. Les variables de pratique laissées libres par les contraintes techniques qui s'imposent aux choix sont alors

[29] Nous supposons que les heures de travail journalières respectent la condition $h^c + h^{nc} \geq \bar{h}$, où $\bar{h}$ est le nombre d'heures ouvrant droit à un *per diem* – 7h dans le dispositif actuel, voir Section 2.1.3.

solution de :

$$\begin{array}{ll} \underset{\{X,l,h^{nc},e\}}{Max} & \widetilde{U}\left(X,l,h^{nc},e\right) \\ s.c. & X - fe\left[T - h^{nc} - l\right] = y \end{array} \tag{2.4}$$

Compte tenu de la description des comportements de pratique retenue, l'arbitrage traditionnel entre consommation et loisir est donc généralisé pour inclure deux types de loisir (l, h^{nc}) ainsi que l'intensité des heures de travail (e). En raison de cette dernière propriété, la contrainte budgétaire que nous étudions est non-linéaire dans les variables endogènes. Plus précisément, les prix des variables de pratique sont eux-mêmes endogènes puisque, par exemple, le prix qui rémunère les heures de travail clinique (fe) dépend du niveau d'effort exercé pendant ces heures. Cette caractéristique nécessite de mettre en oeuvre des outils d'analyse spécifiques, qui permettent notamment d'étudier l'influence du mode de rémunération sur l'arbitrage entre qualité (effort) et quantités (heures et nombre d'actes). Cette approche sera présentée dans la Section 2.2.3. Nous nous limitons dans un premier temps à analyser l'ajustement des marges extensives aux modes de rémunération, en supposant un effort constant. Bien qu'il nécessite d'être réévalué pour intégrer l'endogénéité des choix d'effort, ce modèle apporte en effet une première compréhension des effets de la réforme et des mécanismes à l'oeuvre dans la réponse optimale aux changements de rémunération.

2.2.2 Quantités optimales : analyse du modèle à effort exogène

Si l'effort est supposé constant (et normalisé, $e = 1$) le programme d'optimisation (2.4) ne fait plus intervenir que les heures de travail :

$$\begin{array}{ll} \underset{\{X,l,h^{nc}\}}{Max} & \widetilde{U}\left(X,l,h^{nc}\right) \\ s.c. & X - f\left[T - h^{nc} - l\right] = y \end{array} \tag{2.5}$$

et les choix optimaux qui en résultent s'écrivent comme des fonctions des paramètres

du mode de rémunération : $h^{nc} = h^{nc}(P^c, y)$, $l = l(P^c, y)$.[30]

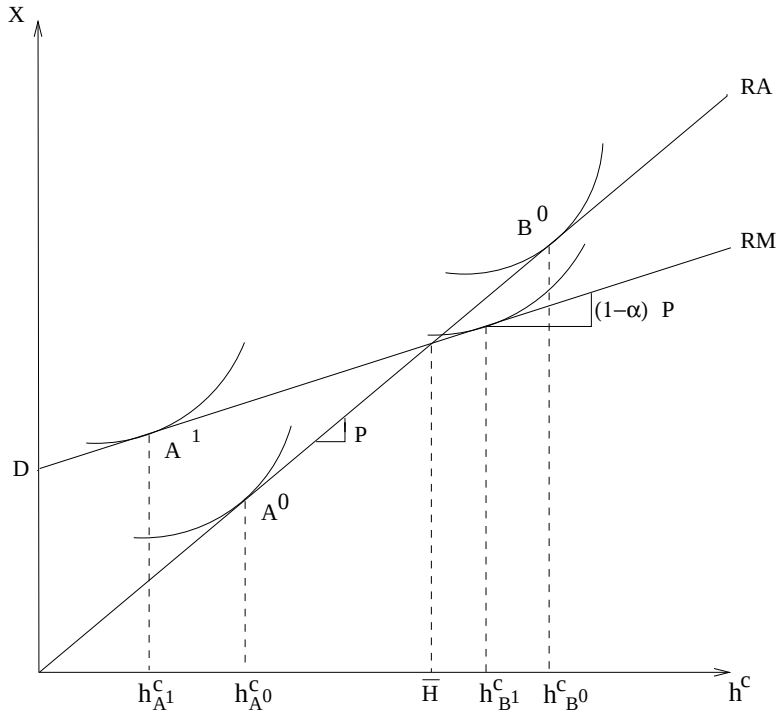
Cette hypothèse simplificatrice permet donc d'éliminer les problèmes de non-linéarité de la contrainte budgétaire liés à l'endogénéité des prix. Elle permet également d'obtenir une illustration graphique simplifiée des déterminants du passage à la rémunération mixte, fournie dans le Graphique 2.1. Lorsque seule la rémunération mixte existe, la contrainte budgétaire est la droite $\boldsymbol{RA}$, de pente P et passant par l'origine. La réforme introduit une seconde contrainte budgétaire – droite $\boldsymbol{RM}$ – dont l'ordonnée à l'origine correspond au *per diem*, D. En raison de la réduction de taux qui s'applique à la rémunération des actes (α), la pente de cette seconde contrainte est inférieure à la première. Les courbes se croisent donc pour un niveau donné d'heures de travail, noté $\overline{H}$.[31] Pour tout niveau des heures de travail en deçà de $\overline{H}$, le revenu sous la rémunération mixte domine donc strictement celui qui aurait résulté du mode de rémunération à l'acte. A l'inverse, le revenu est systématiquement supérieur sous la rémunération à l'acte pour tous les choix de pratique par lesquels les heures de travail excèdent $\overline{H}$. Comme l'adhésion à la rémunération mixte est un choix volontaire des médecins, ces contraintes budgétaires ne sont pas mutuellement exclusives. Les deux segments que nous venons de décrire – droite $\boldsymbol{RM}$ avant $\overline{H}$, $\boldsymbol{RA}$ au delà – constituent donc, ensemble, la contrainte budgétaire efficiente des médecins après la réforme. Cette combinaison de deux droites de pentes différentes est à l'origine d'une seconde non-linéarité dans la contrainte budgétaire.

Ces segments déterminent aussi, et surtout, la décision de passer à la rémunération mixte après la réforme. Les médecins qui décident d'adopter la rémunération mixte sont en effet ceux dont les choix optimaux, après la réforme, se situent sur le premier segment – droite $\boldsymbol{RM}$, avant $\overline{H}$ – de la contrainte budgétaire efficiente. Etant donné que les choix optimaux dépendent des préférences, puisqu'ils résultent d'un programme

[30] Les conditions du premier ordre (CPO) du programme à effort exogène sont omises dans cette section. Elles sont identiques aux CPO du programme à effort endogène, présentées en (2.10), après substitution de la normalisation $e = 1$.

[31] Formellement, ces heures de travail sont telles que le revenu sous la rémunération à l'acte et le revenu sous la rémunération mixte concordent, c'est à dire : $D = \alpha\ P\ \overline{H}$.

Graphique 2.1 – Passage à la rémunération mixte



de maximisation de l'utilité, le passage à la rémunération mixte repose donc sur une auto-sélection des médecins. A titre d'illustration, les préférences de deux médecins-types sont représentées sur le Graphique 2.1. Avant l'introduction de la réforme, le médecin ***A*** comme le médecin ***B*** choisissent les heures de travail qui maximisent leur utilité le long de la contrainte budgétaire ***RA***, matérialisées respectivement par les points $\mathbf{A^0}$ et $\mathbf{B^0}$. Leurs préférences sont telles, cependant, que le premier choisit des heures de travail inférieures à $\overline{H}$, tandis que le second se trouve au delà de $\overline{H}$. Après la réforme, les médecins choisissent leurs heures de travail le long de la contrainte budgétaire efficiente. Pour le médecin ***A***, la possibilité de choisir la rémunération mixte permet une amélioration de bien-être, qui le conduit du point $\mathbf{A^0}$ au point $\mathbf{A^1}$. Il fera donc partie des médecins qui décident d'adopter la rémunération mixte, et une réduction des heures de travail clinique accompagne cette adoption. A l'inverse, le passage du point $\mathbf{B^0}$ au point $\mathbf{B^1}$ correspondrait, pour le médecin ***B***, à une diminution d'utilité. On s'attend donc à ce que ses choix de pratique restent inchangés après la réforme.

Comme le suggère cet exemple graphique, l'adoption de la rémunération mixte par un médecin révèle donc en partie ses préférences à l'égard de la pratique médicale. En notant $\Delta V|_{RM}$ la variation de la variable V engendrée par le passage à la rémunération mixte, on a donc pour tous les médecins dont la pratique est affectée par la réforme : $\Delta\widetilde{U}\Big|_{RM} > 0$. Cette propriété simplifie considérablement l'analyse théorique de l'effet du passage à la rémunération mixte sur les choix de pratique. En notant $\widetilde{V}$ la demande Hicksienne de la variable V et $E(f, \widetilde{U})$ la fonction de dépense, les variations respectives des heures de travail totales et des heures de travail non cliniques peuvent en effet être déduites des expressions suivantes :

$$\Delta h|_{RM} \approx \frac{\partial \widetilde{h}}{\partial f}(-\alpha\ P) + \frac{\partial h}{\partial y} E_u\ \Delta\widetilde{U}\Big|_{RM} \tag{2.6}$$

$$\Delta h^{nc}|_{RM} \approx \frac{\partial \widetilde{h}^{nc}}{\partial f}(-\alpha\ P) + \frac{\partial h^{nc}}{\partial y} E_u\ \Delta\widetilde{U}\Big|_{RM} \tag{2.7}$$

Preuve Les demandes d'heures de travail non clinique et de loisir s'écrivent en fonction des paramètres de rémunération : $l(f,y), h^{nc}(f,y)$. En utilisant la contrainte d'allocation du temps (i), on a donc : $h^c = T - l(f,y) - h^{nc}(f,y) = h^c(f,y)$. Les heures de travail totales correspondent à la somme du temps consacré au travail clinique et non-clinique et on obtient de la même façon : $h = h(f,y)$. Par définition de la fonction de dépense, la demande d'heures de travail s'écrit encore : $h(f,y) = h(f, E(f,\widetilde{U}))$. Soit $\widetilde{U}_C$ l'utilité optimale atteinte sous le mode de rémunération C. La variation induite par le passage à la rémunération mixte correspond alors à la différence : $\Delta h|_{RM} = h\left[(1-\alpha)\ P\ ,\ E\left((1-\alpha)\ P\ ,\ \widetilde{U}_{RM}\right)\right] - h\left[P\ ,\ E\left(P\ ,\ \widetilde{U}_{RA}\right)\right]$. Autour de l'équilibre, cette quantité peut être approximée par l'expression :

$$\Delta h|_{RM} \approx \frac{\partial h}{\partial f}\Delta f + \frac{\partial h}{\partial y}\left[E_f \Delta f + E_{\widetilde{U}}\right] \Delta\ \widetilde{U}\Big|_{RM}$$

où E_i indique la dérivée première de la fonction de dépense par rapport à l'argument i. Par ailleurs, la décomposition de Slutsky met en évidence la combinaison des effets revenu et substitution selon l'expression : $\dfrac{\partial h}{\partial f} = \dfrac{\partial \widetilde{h}}{\partial f} - \dfrac{\partial h}{\partial y}\ f$. On sait en outre, par le lemme de Shephard, que : $E_f = f$. Par ailleurs, la variation du taux de rémunération des actes dans le passage à la rémunération mixte correspond au taux de réduction, soit : $\Delta f = (1-\alpha)\ P - P = -\alpha\ P$. Par substitution, l'expression de la variation des heures de travail s'écrit donc :

$$\Delta h|_{RM} \approx \frac{\partial \widetilde{h}}{\partial f}(-\alpha\ P) + \frac{\partial h}{\partial y} E_u \Delta\ \widetilde{U}\Big|_{RM}$$

L'approximation de la variation des heures non-cliniques s'obtient de la même façon. ■

Sachant que la variation d'utilité est positive, l'hypothèse de normalité des loisirs est alors suffisante pour prédire l'effet de la rémunération mixte sur les heures de travail totales. Si l'on suppose, en outre, que consommation et loisir sont des substituts nets, le modèle à effort exogène permet de prédire sans ambiguïté l'effet de la rémunération mixte sur les choix de pratique, résumé dans la Proposition 2.1.

Proposition 2.1. *Si l'effort (nombre d'actes par heure) est constant, le passage à la rémunération mixte devrait :*

- *Diminuer les heures de travail totales, si les loisirs sont des biens normaux ;*
- *Augmenter les heures de travail non clinique et diminuer les heures de travail clinique, si loisir et consommation sont des substituts nets.*

Preuve Le signe de la variation des heures totales s'obtient directement à partir de l'expression (2.6). Si le loisir est un bien normal, on a : $\frac{\partial h}{\partial y} < 0$. Par définition de la fonction de dépense et sachant que l'utilité s'accroît dans le passage à la rémunération mixte, le second terme du membre de droite est donc négatif. Par définition, la demande Hicksienne est décroissante de son prix. La rémunération des actes étant le prix du loisir, on a donc : $\frac{\partial \widetilde{h}}{\partial f} > 0$. Puisque le prix et le taux de réduction sont positifs, le premier terme du membre de droite est par conséquent négatif et : $\Delta h|_{RM} < 0$.

Les fonctions de demande Hicksiennes sont homogènes de degré 0. L'équation d'Euler de la demande Hicksienne d'heures de travail non clinique est donc :

$$\frac{\partial \widetilde{h}^{nc}}{\partial P_{nc}}.P_{nc} + \frac{\partial \widetilde{h}^{nc}}{\partial P_l}.P_l + \frac{\partial \widetilde{h}^{nc}}{\partial P_X}.P_X = 0$$

L'hypothèse que loisir et consommation sont des substituts nets se traduit alors par : $\frac{\partial \widetilde{h}^{nc}}{\partial f} < 0$. Dans l'expression (2.7), le premier terme est donc positif. Si les heures non cliniques – modélisées comme une forme de loisir – sont un bien normal, le second terme est également positif et $\Delta h^{nc}|_{RM} > 0$.

Les heures totales de travail incluent les heures cliniques et non cliniques. La contrainte d'allocation du temps journalier peut donc s'écrire en termes d'heures totales comme : $T = h + l$, et donc : $\Delta l|_{RM} = -\Delta h|_{RM} > 0$. Ainsi, la décomposition de la contrainte de temps selon les différents types de temps de travail ($T = h^c + h^{nc} + l$) permet de prédire le signe des heures cliniques : $\Delta h^c|_{RM} = -\Delta l|_{RM} - \Delta h^{nc}|_{RM} < 0$. ■

Si l'effort est, comme c'est en général le cas dans la littérature théorique, supposé constant quel que soit le mode de rémunération, l'analyse prédit donc un succès mitigé à la réforme. En particulier, la diminution des heures totales de travail constitue un grave effet pervers au regard des problèmes de liste d'attente rencontrés au Québec. Comme le montre le Tableau 2.3 le nombre d'actes par heures a cependant subi d'importantes variations suite à la réforme. La prochaine section propose donc une extension du modèle, capable d'intégrer les ajustements en termes de marge intensive.

2.2.3 Arbitrage qualité/quantités

Nous adoptons ici une démarche identique à celle utilisée dans le modèle à effort exogène. Afin de mettre en évidence la substitution entre les marges intensive et extensive, nous nous intéressons cependant au programme d'optimisation sous contrainte non-linéaire (2.4), dans lequel l'effort – donc les prix – est une variable endogène.

Pour chaque variable de pratique β, $\beta = \{X, l, h^{nc}, h^c, e\}$, les conditions du premier ordre du programme définissent les demandes optimales comme une fonction implicite des paramètres du mode de rémunération : $\beta = \beta(f, y)$. Le nombre d'actes optimal s'en déduit par la contrainte technologique (ii) : $A(f, y) = h^c(f, y)\ e(f, y)$. Comme précédemment, l'effet du passage à la rémunération mixte sur les choix de pratique peut être décomposé entre effets substitution et revenu :

$$\Delta\beta|_{RM} \approx \frac{\widetilde{\beta}}{\partial f}\Delta f + \frac{\partial\beta}{\partial y}E_U\ \Delta\widetilde{U}\Big|_{RM}$$

Preuve Pour les médecins qui choisissent de l'adopter, l'impact de la rémunération mixte sur les choix optimaux correspond à : $\Delta\beta|_{RM} = \beta(f_{RM}, y_{RM}) - \beta(f_{RA}, y_{RA})$. Pour tout système de rémunération C ($C \in \{RA, RM\}$), la fonction de dépense correspond par définition à : $y_C = E(f_C, \widetilde{U_C})$. L'effet du passage à la rémunération mixte peut donc être approximé par l'expression :

$$\Delta\beta|_{RM} \approx \frac{\partial\beta}{\partial f}\Delta f + \frac{\partial\beta}{\partial y}\left[E_f \Delta f + E_U \left.\Delta\widetilde{U}\right|_{RM}\right] = \left[\frac{\partial\beta}{\partial f} + \frac{\partial\beta}{\partial y}E_f\right]\Delta f + \frac{\partial\beta}{\partial y}E_U \Delta \left.\widetilde{U}\right|_{RM} \tag{2.8}$$

Blomquist (1989) propose une analyse systématique de la théorie du consommateur sous contrainte non-linéaire. En particulier, le lemme de Shephard et la décomposition de Slutsky peuvent être facilement adaptés à ce cas.

En adoptant les notations proposées par l'auteur, soit $g(f, \beta) = X - f.e\,[T - h^o - l]$ la contrainte budgétaire. Le lemme de Shephard dans le cas non-linéaire s'écrit : $\dfrac{\partial E}{\partial f} = \dfrac{\partial g}{\partial f} = g'_f$. Définissant les demandes Hicksiennes, $\widetilde{\beta}(f, U)$,comme les solutions du programme de minimisation de la dépense pour un niveau d'utilité U donné, l'équation de Slutsky devient par ailleurs : $\dfrac{\partial\beta}{\partial f} = \dfrac{\partial\widetilde{\beta}}{\partial f} - \dfrac{\partial\beta}{\partial y}.g'_f$.

Ensemble, ces résultats impliquent donc que : $\dfrac{\partial\beta}{\partial f} + \dfrac{\partial\beta}{\partial y}.E_f = \dfrac{\partial\widetilde{\beta}}{\partial f}$, d'où provient le résultat par substitution dans (2.8). ■

a) Effets revenu

La modélisation que nous avons adopté suggère un certain nombre d'hypothèses quant à la forme de la fonction d'utilité qui permettent de circonscrire les signes des effets revenus.

D'une part, notre analyse considère l'effort comme un critère de qualité des soins dispensés, puisqu'il constitue un indicateur du temps consacré aux soins et à leur explication, de la vigilance du médecin, de la pertinence du diagnostic, etc. Pour toutes ces raisons, le niveau de santé atteint par les patients est supposé décroissant de l'effort (voir Section 2.2.1). Dans la mesure où le niveau de santé affecte négativement le bien-être des médecins, l'effort devrait donc apparaître comme un mal dans leur fonction d'utilité.

D'autre part, un nombre significatif de médecins consacrent une partie importante de leur temps aux activités non-cliniques (Tableau 2.3) et ce y compris lorsque, comme sous la rémunération à l'acte, celles-ci ne donnent lieu à aucune rémunération. Cette caractéristique nous a conduit à considérer les heures de travail non-cliniques comme une forme particulière de loisir, au sens où elles participent à accroître le bien-être des médecins. Prenant acte des résultats obtenus par la plupart des travaux empiriques consacrés à l'offre de travail (Pencavel, 1986), nous supposons que tous les loisirs, le loisir pur comme les heures de travail non-clinique, sont des biens normaux. Sous ces hypothèses, l'effet de l'introduction de la rémunération mixte sur les variables de pratique dépend alors exclusivement de la sensibilité des demandes compensées aux variations de prix.

***Lemme* 2.1.** *Si :*

- *Les deux types de loisir sont des biens normaux ;*
- *L'effort est un mal ;*

alors la sensibilité des choix de pratique aux variations induites par la rémunération mixte ne dépend que de la sensibilité au prix des demandes Hicksiennes.

Preuve Nous utilisons ici les résultats présentés dans la preuve de la Proposition 2.1. On a ainsi : $E_U > 0$ par définition de la fonction de dépense, $\Delta\widetilde{U}\Big|_{RM} > 0$ en raison du passage volontaire à la RM et $\Delta f|_{RM} = -\alpha P \leq 0$. Les hypothèses présentées dans le Lemme 2.1 fournissent en outre la forme des effets revenu.

Si l'effort est un mal, sa demande non-compensée est décroissante du niveau de revenu. On a alors :

$$\Delta e|_{RM} \approx \frac{\partial \widetilde{e}}{\partial f} \underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial e}{\partial y}}_{<0} \underbrace{E_U}_{>0} \underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}$$

Une condition suffisante pour que l'effort diminue avec le passage à la rémunération mixte est donc que : $\dfrac{\partial \widetilde{e}}{\partial f} > 0$.

Si les deux types de loisir sont des biens normaux, on a :

$$\Delta l|_{RM} \approx \frac{\partial \widetilde{l}}{\partial f}\underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial l}{\partial y}}_{>0}\underbrace{E_U}_{>0}\underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}, \text{ et } \Delta h^{nc}|_{RM} = \frac{\partial \widetilde{h}^{nc}}{\partial f}\underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial h^{nc}}{\partial y}}_{>0}\underbrace{E_U}_{>0}\underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}$$

et $\dfrac{\partial \widetilde{l}}{\partial f} < 0, \dfrac{\partial \widetilde{h}^{nc}}{\partial f} < 0$ sont donc des conditions suffisantes pour que les deux types de loisir augmentent simultanément dans le passage à la rémunération mixte.

Par la contrainte d'allocation du temps (ii), la sensibilité des heures de travail totales au revenu hors-tavail est : $\dfrac{\partial h^c}{\partial y} = \dfrac{\partial}{\partial y}\left[T - l - h^{nc}\right] = -\dfrac{\partial l}{\partial y} - \dfrac{\partial h^{nc}}{\partial y} < 0$. L'effet du passage à la rémunération mixte peut donc être résumé par :

$$\Delta h^c|_{RM} \approx \frac{\partial \widetilde{h}^c}{\partial f}\underbrace{\Delta f}_{<0} + \underbrace{\frac{\partial h^c}{\partial y}}_{<0}\underbrace{E_U}_{>0}\underbrace{\Delta\, \widetilde{U}\Big|_{RM}}_{>0}$$

et $\dfrac{\partial \widetilde{h}^c}{\partial f} > 0$ est une condition suffisante pour que les heures cliniques augmentent avec le passage à la rémunération mixte. ■

b) Effets prix

L'analyse des effets prix est considérablement complexifiée par la non-linéarité de la contrainte budgétaire. Sa linéarisation, fondée sur la définition de prix virtuels, permet cependant de retrouver les résultats standards de la théorie du consommateur.[32]

On note $\pi_\alpha, \alpha = \{l, h^{nc}, e\}$ les prix virtuels associés aux variables de pratique, c'est à dire les prix tels que les fonctions de demande $\beta(f, y)$ résultent d'une contrainte linéaire dans les prix virtuels. Ils correspondent par définition à : $\pi_l = fe$, $\pi_{nc} = fe$, $\pi_e = -fh^c$ et le programme d'optimisation (2.4) est alors formellement équivalent au programme

[32] A notre connaissance, l'utilisation de cette technique d'analyse remonte au travail de Becker & Lewis (1973) étudiant l'arbitrage entre quantité et qualité des enfants dans les choix de fertilité. Cette technique de linéarisation utilisant les prix virtuels a ensuite fait l'objet d'analyses systématiques par Edlefsen (1981) puis Blomquist (1989).

linéaire :

$$\begin{aligned} Max \quad & \widetilde{U}\left(X, l, h^{nc}, e\right) \\ s.c. \quad & X + \pi_l l + \pi_{nc} h^{nc} = y + \pi_e e \end{aligned} \tag{2.9}$$

Preuve Soient γ_1 le prix implicite de la consommation et L le lagrangien associé au programme (2.4). Les conditions du premier ordre s'écrivent :

$$\begin{array}{lcll} \dfrac{\partial \widetilde{L}}{\partial X} & = & \dfrac{\partial \widetilde{U}}{\partial X} - \gamma_1 = 0 & \Rightarrow \quad \gamma_1 = \widetilde{U} m_X \\ \dfrac{\partial \widetilde{L}}{\partial l} & = & \dfrac{\partial \widetilde{U}}{\partial X} - \gamma_1 fe = 0 & \Rightarrow \quad \dfrac{\widetilde{U} m_l}{\widetilde{U} m_X} = fe \\ \dfrac{\partial \widetilde{L}}{\partial h^{nc}} & = & \dfrac{\partial \widetilde{U}}{\partial h^{nc}} - \gamma_1 fe = 0 & \Rightarrow \quad \dfrac{\widetilde{U} m_{nc}}{\widetilde{U} m_X} = fe \\ \dfrac{\partial \widetilde{L}}{\partial e} & = & \dfrac{\partial \widetilde{U}}{\partial e} + \gamma_1 f[T - h^{nc} - l] = 0 & \Rightarrow \quad \dfrac{\widetilde{U} m_e}{\widetilde{U} m_X} = -fh^c \\ \dfrac{\partial \widetilde{L}}{\partial \gamma_1} & = & y + fe[T - h^{nc} - l] - X = 0 & \end{array} \tag{2.10}$$

Par définition, les prix virtuels sont les prix de la contrainte linéaire telle que les fonctions de demande qui en résultent sont identiques aux fonctions de demande issues de ces CPO. Dans le cas classique d'une optimisation sous contrainte linéaire, la demande optimale du consommateur pour le bien x satisfait $TMS_{x,x_0} = p_x$, où x_0 désigne le bien numéraire. En appliquant ce résultat aux CPO (2.10), les prix virtuels, notés $\pi_\alpha, \alpha = \{l, h^{nc}, e\}$ sont donc définis par les conditions :

$$\pi_l = TMS_{l,X} = fe\,;\ \pi_o = TMS_{nc,X} = fe\,;\ \pi_e = TMS_{e,X} = -fh^c \tag{2.11}$$

Il convient également de noter que, au voisinage de l'équilibre, les demandes Hicksiennes et Marchaliennes concordent. Dans cet intervalle, les prix virtuels peuvent donc s'écrire en fonction des demandes Hicksiennes, $\pi_l = f\widetilde{e}\,;\ \pi_o = f\widetilde{e}\,;\ \pi_e = -f\widetilde{h}^c$.

Les demandes qui résultent du programme linéaire (2.9) sont alors mécaniquement identiques à celles qui résolvent le programme non linéaire (2.4). ■

Par définition, les fonctions de demande qui apparaissent dans la décomposition de l'effet du passage à la rémunération mixte (2.8) sont solutions du programme linéaire. En particulier, les fonctions de demande Hicksiennes peuvent donc être exprimées comme

des fonctions implicites des prix virtuels :

$$\widetilde{l} = \widetilde{l}(\pi_l, \pi_{nc}, \pi_e, \widetilde{U}) \, ; \, \widetilde{h}^{nc} = \widetilde{h}^{nc}(\pi_l, \pi_{nc}, \pi_e, \widetilde{U}) \, ; \, \widetilde{e} = \widetilde{e}(\pi_l, \pi_{nc}, \pi_e, \widetilde{U}) \tag{2.12}$$

Sous l'hypothèse que les heures totales de travail de tout médecin qui choisit la rémunération mixte ouvrent droit à un *per diem* ($h^c + h^{nc} \geq \bar{h}$), les prix des deux types de loisir sont identiques ($\pi_l = \pi_{nc} = f.e$). Le loisir total, $L = T - h^c$, est donc un agrégat Hicksien. En s'appuyant sur cette propriété, le modèle est analysé en termes d'abord d'arbitrage entre les heures totales de loisir et l'effort, puis d'allocation du loisir optimal entre les différentes occupations.

On note $\pi_L = \pi_l = \pi_{nc}$ le prix virtuel du loisir total. Les fonctions de demande Hisksiennes s'écrivent en fonction de ce prix : $\widetilde{e}(\pi_L, \pi_e, \widetilde{U})$ et $\widetilde{L}(\pi_L, \pi_e, \widetilde{U})$. Comme nous l'avons vu plus haut (Section 2.1.1) la littérature d'économie de la santé admet communément qu'une augmentation du taux de rémunération des actes tendra à accroître simultanément les heures de travail clinique et l'effort (en diminuant le temps consacré à chaque patient). Ces résultats proviennent de diverses études théoriques et/ou empiriques dont l'analyse isole l'une des variables de pratique (heures ou effort). En intégrant dans l'analyse l'arbitrage entre les marges intensive et extensive, il apparaît cependant que ce résultat n'est valide que pour certaines valeurs des élasticités-prix croisées entre ces variables.

Proposition 2.2. *Les demandes compensées de l'effort et des heures cliniques sont croissantes du taux de rémunération des actes si :*

- *Condition nécessaire :* $\eta_{h^c,\pi_e} = \eta_{e,\pi_L} < 1$ *;*
- *Condition suffisante :* $(1 - \eta_{h^c,\pi_e})^2 = (1 - \eta_{e,\pi_L})^2 > \eta_{e,\pi_e}\eta_{h^c,\pi_L}$

Preuve Les prix virtuels permettent de contourner la non-linéarité de la contrainte budgétaire due à l'endogénéité des prix. Ils dépendent donc des variables de pratique et leur réaction aux variations du taux de rémunération des actes, au voisinage de l'équilibre, correspond donc à :

$$\frac{\partial \pi_L}{\partial f} = \frac{\partial \pi_l}{\partial f} = \frac{\partial \pi_{nc}}{\partial f} = \widetilde{e} + f\frac{\partial \widetilde{e}}{\partial f} \; ; \; \frac{\partial \pi_e}{\partial f} = -\left[\widetilde{h}^c + f\frac{\partial \widetilde{h}^c}{\partial f}\right]$$

En utilisant ce résultat, l'effet compensé d'une variation du taux de rémunération des actes, f, sur l'effort optimal s'écrit :

$$\frac{\partial \widetilde{e}(\pi_L, \pi_e, \widetilde{U})}{\partial f} = \frac{\partial \widetilde{e}}{\partial \pi_L}\left[\widetilde{e} + f\ \frac{\partial \widetilde{e}}{\partial f}\right] + \frac{\partial \widetilde{e}}{\partial \pi_e}\left[\widetilde{h}^c + f\ \frac{\partial \widetilde{h}^c}{\partial f}\right]$$

Quelques manipulations algébriques permettent de faire apparaître les expressions définissant les prix virtuels, présentés en (2.11). En notant η les élasticités-prix compensées, l'expression précédente s'écrit alors :

$$\eta_{e,f} = \frac{\eta_{e,\pi_L} + \eta_{e,\pi_e} + \eta_{e,\pi_e}\eta_{h^c,f}}{1 - \eta_{e,\pi_L}} \tag{2.13}$$

En procédant de la même façon, on obtient l'expression de l'élasticité prix compensée du loisir total : $\eta_{L,f} = \eta_{L,\pi_L}(1 + \eta_{e,f}) + \eta_{L,\pi_e}(1 + \eta_{h^c,f})$. Par la contrainte d'allocation du temps, cette expression s'écrit de façon équivalente en termes d'heures cliniques. Sachant que $\widetilde{L} = T - \widetilde{h}^c$, on a en effet $\frac{\partial \widetilde{L}}{\partial f} = -\frac{\partial \widetilde{h}^c}{\partial f} \Leftrightarrow \frac{\partial \widetilde{L}}{\partial f}\ \frac{f}{\widetilde{L}} = -\frac{\partial \widetilde{h}^c}{\partial f}\ \frac{f}{\widetilde{L}}$ et donc : $\widetilde{L}\ \eta_{L,f} = -\widetilde{h}^c\ \eta_{h^c,f}$. Après substitutions, l'élasticité-prix compensée des heures cliniques devient :

$$\eta_{h^c,f} = \frac{\eta_{h^c,\pi_L} + \eta_{h^c,\pi_e} + \eta_{h^c,\pi_L}\eta_{e,f}}{1 - \eta_{h^c,\pi_e}} \tag{2.14}$$

Par définition des demandes compensées, les effets prix propres sont négatifs : $\frac{\partial \widetilde{e}}{\partial \pi_e} \leq 0$ et $\frac{\partial \widetilde{h}^c}{\partial \pi_L} = -\frac{\partial \widetilde{L}}{\partial \pi_L} \geq 0$. Lorsque l'analyse inclut deux variables de pratique, et sous hypothèse de substitution nette entre la consommation et les deux types de loisir, les équations d'Euler imposent donc que les effets croisés soient positifs : $\frac{\partial \widetilde{e}}{\partial \pi_L} \geq 0$ et $\frac{\partial \widetilde{h}^c}{\partial \pi_e} = -\frac{\partial \widetilde{L}}{\partial \pi_e} \leq 0$. Comme le prix virtuel de l'effort est négatif ($\pi_e = -f\ \widetilde{h}^c$), toutes les élasticités-prix ($\eta_{\beta,\pi} = \frac{\partial \widetilde{\beta}}{\partial \pi_\beta}\ \frac{\pi_\beta}{\widetilde{\beta}}$, $\beta = \{h^c, e\}$) sont positives.

Les demandes compensées d'effort et d'heures cliniques de travail réagissent positivement aux variations de prix si : $\eta_{h^c,f} > 0$ et $\eta_{e,f} > 0$. Dans ce cas, les numérateurs des expressions (2.13) et (2.14) sont tous deux positifs. Il est donc nécessaire que les dénominateurs le soient également, c'est à dire que : $1 - \eta_{e,\pi_L} > 0$ et $1 - \eta_{h^c,\pi_e} > 0$. La matrice de Slutsky est symétrique, les effets prix croisés sont donc reliés selon l'expression : $\frac{\partial \widetilde{L}}{\partial \pi_e} = \frac{\partial \widetilde{e}}{\partial \pi_L}$. En utilisant les résultats issus de la contrainte d'allocation du temps, $\frac{\partial \widetilde{L}}{\partial \pi_e} = -\frac{\partial \widetilde{h}^c}{\partial \pi_e}$, cette propriété se traduit en termes d'élasticités prix sous la

forme : $-\dfrac{\partial \widetilde{h^c}}{\partial \pi_e}\ f\ \dfrac{h^c}{h^c} = \dfrac{\partial \widetilde{e}}{\partial \pi_L}\ f\ \dfrac{e}{e} \Leftrightarrow \eta_{h^c,\pi_e} = \eta_{e,\pi_L}$. Au total, la condition nécessaire s'écrit donc : $\eta_{h^c,\pi_e} = \eta_{e,\pi_L} < 1$.

Après substitution de (2.14) dans (2.13) et simplifications, les signes des élasticités-prix peuvent être étudiés à partir des relations suivantes :

$$\eta_{e,f} = \frac{\eta_{e,\pi_e}\left(1+\eta_{h^c,\pi_L}\right) + \eta_{e,\pi_L}\left(1-\eta_{e,\pi_L}\right)}{\left(1-\eta_{e,\pi_L}\right)^2 - \eta_{e,\pi_e}.\eta_{h^c,\pi_L}}$$
$$\eta_{h^c,f} = \frac{\eta_{h^c,\pi_L}\left(1+\eta_{e,\pi_e}\right) + \eta_{h^c,\pi_e}\left(1-\eta_{h^c,\pi_e}\right)}{\left(1-\eta_{h^c,\pi_e}\right)^2 - \eta_{e,\pi_e}.\eta_{h^c,\pi_L}} \tag{2.15}$$

La condition nécessaire garantit que les dénominateurs sont positifs. En s'appuyant sur la symétrie de la matrice de Slutsky, une condition suffisante est alors que les numérateurs soient positifs, *i.e.* : $\left(1-\eta_{h^c,\pi_e}\right)^2 = \left(1-\eta_{e,\pi_L}\right)^2 > \eta_{e,\pi_e}\eta_{h^c,\pi_L}$. ■

Sous les conditions décrites dans la Proposition 2.2, le déplacement le long d'une courbe d'utilité induit par un accroissement du taux de rémunération des actes (rotation de la contrainte budgétaire) conduit à une augmentation des demandes optimales d'effort et d'heures de travail clinique ; donc à une réduction du temps consacré au loisir pendant les semaines de travail, L. La modélisation que nous avons adopté introduit cependant une distinction importante entre les allocations possibles de ce temps de loisir. Les heures qui ne sont pas consacrées à la réalisation d'actes médicaux peuvent en effet être utilisées à des activités productives si elles prennent la forme d'heures de travail non-clinique. Accroître le temps consacré à ces activités est, au demeurant, l'un des objectifs qui a présidé à l'introduction de la rémunération mixte (Section 2.1.3). Au-delà de la réaction du temps total de loisir au variables de rémunération, l'allocation du temps entre les loisirs participe donc à la détermination du profil de pratique des médecins.

Proposition 2.3. *Les configurations de signes possibles sont résumées dans le Tableau 2.5. Les effets prix du loisir ne sont simultanément négatifs que sous les conditions (1a), (2b), (3b), (4a), (5) et (6a,b).*

Tableau 2.5 – Déterminants théoriques de l'allocation du loisir

Cas	η_{l,π_e}	η_{h^{nc},π_e}	$(\eta_{h^c,\pi_L} - \eta_{e,\pi_e})$	$\eta_{l,f}$	$\eta_{h^{nc},f}$
(1)	+	−	+	$+/-^a$	−
(2)	+	−	−	−	$+/-^b$
(3)	−	+	+	−	$+/-^b$
(4)	−	+	−	$+/-^a$	−
(5)	−	−	+	−	−
(6)	−	−	−	$+/-^a$	$+/-^b$

[a] Négatif si : $\eta_{l,\pi_e}(\eta_{h^c,f} - \eta_{e,f}) < (1 - \eta_{e,f})\eta_{l,p_x}$.
[b] Négatif si : $\eta_{h^{nc},\pi_e}(\eta_{h^c,f} - \eta_{e,f}) < (1 - \eta_{e,f})\eta_{l,p_x}$.

Preuve En utilisant l'expression des demandes Hicksiennes du programme linéarisé (2.12), l'effet compensé d'une variation de prix sur la demande d'heures de travail non-cliniques s'écrit :

$$\frac{\partial \widetilde{h}^{nc}}{\partial f} = \frac{\partial \widetilde{h}^{nc}}{\partial \pi_l}\left[\widetilde{e} + f\ \frac{\partial \widetilde{e}}{\partial f}\right] + \frac{\partial \widetilde{h}^{nc}}{\partial \pi_{nc}}\left[\widetilde{e} + f\ \frac{\partial \widetilde{e}}{\partial f}\right] - \frac{\partial \widetilde{h}^{nc}}{\partial \pi_e}\left[\widetilde{h}^c + f\ \frac{\partial \widetilde{h}^c}{\partial f}\right] \tag{2.16}$$

Allocation du loisir. Un raisonnement identique s'applique à le demande compensée de loisir pur. D'après la définition des prix virtuels au voisinage de l'équilibre, ces expressions peuvent être converties en termes d'élasticités compensées. L'arbitrage qui guide l'allocation du loisir est alors décrit par le système :

$$\begin{aligned}
\eta_{l,f} &= [\eta_{l,\pi_l} + \eta_{l,\pi_{nc}} + \eta_{l,\pi_e}] + \eta_{e,f}\,[\eta_{l,\pi_l} + \eta_{l,\pi_{nc}}] + \eta_{l,\pi_e}\eta_{h^c,f} \\
\eta_{h^{nc},f} &= [\eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_e}] + \eta_{e,f}\,[\eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_{nc}}] + \eta_{h^{nc},\pi_e}\eta_{h^c,f}
\end{aligned}$$

soit encore :

$$\begin{aligned}
\eta_{l,f} &= (1 + \eta_{e,f})\,(\eta_{l,\pi_l} + \eta_{l,\pi_{nc}} + \eta_{l,\pi_e}) + \eta_{l,\pi_e}\,(\eta_{h^c,f} - \eta_{e,f}) \\
\eta_{h^{nc},f} &= (1 + \eta_{h^{nc},f})\,(\eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_{nc}} + \eta_{h^{nc},\pi_e}) + \eta_{h^{nc},\pi_e}\,(\eta_{h^c,f} - \eta_{e,f})
\end{aligned}$$

En notant p_x le prix de la consommation et sachant que les demandes compensées sont homogènes de degré 0 dans les prix, les équations d'Euler s'écrivent : $\dfrac{\partial \widetilde{\beta}}{\partial \pi_l}\pi_l + \dfrac{\partial \widetilde{\beta}}{\partial \pi_{nc}}\pi_{nc} + \dfrac{\partial \widetilde{\beta}}{\partial \pi_e}\pi_e = \dfrac{\partial \widetilde{\beta}}{\partial p_x}p_x$, $\beta \in \{l, h^{nc}\}$. Après substitution de ce résultat dans le système d'équations précédent, les expressions qui gouvernent l'allocation du loisir sont :

$$\eta_{l,f} = -\eta_{l,p_x}(1 + \eta_{e,f}) + \eta_{l,\pi_e}(\eta_{h^c,f} - \eta_{e,f}) \tag{2.17}$$

$$\eta_{h^{nc},f} = -\eta_{h^{nc},p_x}(1 + \eta_{e,f}) + \eta_{h^{nc},\pi_e}(\eta_{h^c,f} - \eta_{e,f}) \tag{2.18}$$

Les signes des variations compensées du loisir et des heures de travail non-clinique sont déduits de ces expressions.

Etude des signes. Si la consommation et les deux types de loisir sont des substituts Hicksiens, on a : $\eta_{l,p_x} > 0$ et $\eta_{h^{nc},p_x} > 0$. Par ailleurs, la manipulation des expressions (2.15) permet d'obtenir :

$$\eta_{h^c,f} - \eta_{e,f} = \frac{\eta_{e,\pi_e} - \eta_{h^c,\pi_L}}{(1 - \eta_{h^c,\pi_e})^2 - \eta_{e,\pi_e}.\eta_{h^c,\pi_L}}$$

Sous les conditions de la Proposition 2.2, le dénominateur de cette expression est positif. La preuve de cette proposition a en outre établit que $\eta_{L,\pi_e} = \eta_{l,\pi_e} + \eta_{h^{nc},\pi_e} < 0$. Les élasticités des heures de travail non-clinique et du loisir pur ne peuvent donc pas être simultanément positifs.

L'ensemble de ces résultats est utilisé pour construire le Tableau 2.5, dont les quatre premières colonnes décrivent les signes possibles des termes du membre de droite de (2.17) et (2.18), les deux dernières colonnes les signes induits des élasticités compensées. ■

En vertu du Lemme 2.1, l'étude des signes des effets compensés permet de mettre en évidence les conditions suffisantes à ce que la rémunération mixte ait l'effet traditionnellement attendu sur les variables de pratique : une diminution de l'effort et des heures de travail clinique ainsi qu'un accroissement des heures de travail non-clinique. Comme l'indiquent nos résultats, résumés ci-dessous, seules quelques configurations spécifiques des préférences confirment ces attentes.

Résumé

Sous les hypothèses de la Proposition 2.2 et du Lemme 2.1 :

- *Le passage à la rémunération mixte devrait diminuer l'effort ainsi que les heures de travail clinique ;*
- *Dans les cas (1a), (2b), (3b), (4a), (5) et (6a,b) du Tableau 2.5, cette baisse des heures cliniques est partagée entre une augmentation des heures de travail non-clinique et une augmentation du loisir pur ;*
- *En conséquence, le temps total de travail diminue ;*
- *Dans tous les autres cas, les effets substitution et revenu agissent en sens opposé, et l'effet de la rémunération mixte sur l'allocation du loisir est ambigu.*

L'analyse économétrique, consacrée à l'estimation des préférences des médecins, laisse aux comportements observés le soin de trancher ces ambiguïtés.

2.3 Modèle économétrique

Conformément au cadre de notre analyse théorique, le modèle économétrique est spécifié en termes de maximisation d'utilité, selon la forme réduite (2.4). L'estimation s'appuie sur les comportements de pratique annuels des médecins spécialistes du Québec. Les variables de pratique décrites ci-dessus sont donc désormais définies sur une base annuelle plutôt que journalière. Afin de permettre une analyse plus fine de l'allocation du loisir, nous distinguons en particulier le loisir pris pendant les semaines de travail des semaines de loisir dans l'année, notées S ($S = 52 - W$).[33] Les heures de travail (clinique et non-clinique) sont alors mesurées par la moyenne annuelle des heures de travail hebdomadaires réalisées pendant les semaines de travail. Les actes sont, quant à eux, mesurés en termes de quantités annuelles. La distinction introduite par la rémunération mixte entre actes facturables et non-facturables repose en grande partie sur les caractéristiques techniques des actes pratiqués.[34] Nous tenons donc compte de la possibilité d'utilités (ou de désutilités) marginales différentes en incluant séparément ces deux types d'actes dans les préférences des médecins.

En résumé, la démarche économétrique adoptée consiste donc à estimer des préférences de la forme : $U = U(S, l, h^{nc}, AF, ANF, X)$, d'où résultent les choix de pratique

[33] Les résultats empiriques obtenus par Hanoch (1980) et Blank (1988) confirment l'existence d'une substitution imparfaite entre ces deux types de loisir.

[34] Une grande partie des actes non-facturables est constituée, par exemple, des visites de contrôle qui suivent la délivrance de soins.

optimaux des médecins en termes de :

(i) temps hebdomadaire de travail clinique (actes médicaux), h^c ;

(ii) temps hebdomadaire de travail non-clinique (administration, enseignement), h^{nc} ;

(iii) semaines de travail annuelles, W ;

(iv) quantités d'actes facturables réalisés pendant l'année, AF ;

(v) quantités d'actes non-facturables réalisés pendant l'année, ANF.

Ce modèle structurel est identifié grâce aux variations de prix induites par l'introduction de la rémunération mixte. Comme nous l'avons vu, la contrainte budgétaire qui en résulte présente différentes non-linéarités, issues à la fois de la combinaison des modes de rémunération mixte et à l'acte le long de la contrainte budgétaire efficiente et de l'endogénéité des prix. Suivant une tradition récente en économétrie de l'offre de travail (van Soest, 1995), cette difficulté est surmontée en discrétisant l'ensemble de choix. L'estimation des paramètres de la fonction d'utilité repose alors sur un modèle de choix discret, dont la spécification est présentée dans la Section 2.3.1. Nous apportons ensuite un certain nombre de modifications à ce cadre de base, commandées par les spécificités de la base de données utilisée (Section 2.3.2).

2.3.1 Modèle de choix discrétisé : éléments de base

Pour chaque variable de pratique, nous considérons un nombre fini de niveaux possibles entre lesquels les médecins choisissent. Le nombre de niveaux retenu pour chaque variable de pratique est donc un élément important de la mise en oeuvre du modèle. Notre objectif en la matière est de recouvrir au mieux le large éventail des choix de pratique observés dans l'échantillon. Nous avons donc fait le choix de conserver un nombre important de niveaux pour chaque variable de pratique. Plus précisément, nous avons retenu $N_c = 7$ niveaux de discrétisation pour les heures de travail clinique, $N_{nc} = 8$ niveaux pour les heures de travail non-clinique, $N_w = 6$ niveaux pour les semaines de

Tableau 2.6 – Distribution de l'échantillon entre les niveaux de discrétisation

Heures				**Semaines**		**Actes**[a]			
h^c		h^{nc}		W		AF		ANF	
0	3.12%	0	66.89%	0	0.12%	0	36.15%	0	60.16%
20	12.96%	15	27.22%	10	0.53%	140000	55.38%	100000	28.03%
40	55.16%	30	4.34%	20	0.82%	280000	7.96%	200000	10.22%
60	24.92%	45	1.19%	30	1.36%	420000	0.39%	300000	1.50%
80	3.26%	60	0.30%	40	26.43%	560000	0.09%	400000	0.08%
100	0.57%	75	0.04%	50	70.75%	700000	0.02%	500000	0.01%
120	0.01%	90	0.02%	-	-	840000	0.02%	-	-
-	-	105	0.01%	-	-	-	-	-	-

[a] En Dollars constants (base 1996).

Note. Pour chaque variable de pratique (heures de travail clinique, h^c, heures de travail non-clinique, h^{nc}, semaines de travail, W, actes facturables, AF, et actes non-facturables, ANF), pourcentage d'observations pour lesquelles le choix observé discrétisé concorde avec le niveau de discrétisation correspondant.

travail, $N_{AF} = 7$ niveaux pour les actes facturables et $N_{ANF} = 6$ niveaux pour les actes non-facturables. Comme l'illustre le Tableau 2.6 cette stratégie assure une assez large diffusion de l'échantillon entre les niveaux retenus.

Une alternative consiste alors en une combinaison particulière de variables de pratique, c'est à dire un ensemble de valeurs : $j = \{c_j, nc_j, w_j, ANF_j, AF_j\}$ désignant respectivement le ${c_j}^{eme}$ niveau d'heures de travail clinique, $c_j \in \{1, ..., N_c\}$, le ${nc_j}^{eme}$ niveau d'heures de travail non-clinique, etc .... L'ensemble des niveaux de discrétisation définit donc un ensemble d'alternatives J, de dimension : $dim(J) = N_c \times N_o \times N_w \times N_{NBA} \times N_{BA} = 14112$.

L'estimation du modèle consiste à retenir les valeurs des paramètres de la fonction d'utilité qui maximisent la vraisemblance de l'alternative effectivement choisie. La mise en oeuvre de l'estimation nécessite donc de spécifier la forme de la fonction d'utilité, qui guide le choix au sein de l'ensemble J. On note V_j l'utilité annuelle que retire un médecin

représentatif des choix de pratique dans l'alternative j. Une hypothèse devenue classique dans la littérature (McFadden, 1974) consiste à prendre en compte les erreurs de mesure propres à chaque alternative en décomposant l'utilité, V, entre une part déterministe, u_j, et une erreur de mesure indépendante entre les alternatives, ϵ_j : $V_j = u_j + \epsilon_j$.

La partie déterministe de l'utilité est spécifiée selon une fonction d'utilité translog, qui constitue une approximation du second ordre de toute fonction d'utilité correctement spécifiée et permet de prendre en compte une grande variété de profils de substitution (Christensen, Jorgenson & Lau, 1975). Formellement, la composante déterministe de la fonction d'utilité peut être définie, sous forme condensée, comme[35] :

$$\begin{aligned} u_j &= \mathbf{G}'\ Z_j + Z_j'\ \mathbf{B}\ Z_j + \gamma_{ANF}\ \ln ANF_j \\ &+ \mathbf{B_{ANF}}'\ Z_j\ \ln ANF_j + \beta_{ANF}\ (\ln ANF_j)^2 \end{aligned} \tag{2.19}$$

Notations Z_j désigne le vecteur colonne des variables de pratique associées à l'alternative j, à l'exception des actes non-facturables :

$$Z_j = \left[\ln(h_j^{nc}), \ln(R - W_j), \ln(T - h_j^{nc} - h_j^c), \ln(AF_j), \ln(X_j)\right]'$$

où T est le nombre d'heures totales disponibles dans une semaine (égal à $7 \times 24 = 168$ dans l'application) et R $(= 52)$ le nombre total de semaines disponibles dans l'année.

Les paramètres à estimer sont contenus dans les matrices :

$$\mathbf{B} = \begin{pmatrix} \beta_{nc} & \beta_{nc}^S & \beta_{nc}^l & \beta_{nc}^{AF} & \beta_{nc}^x \\ \beta_S^{nc} & \beta_S & \beta_S^l & \beta_S^{AF} & \beta_S^x \\ \beta_l^{nc} & \beta_l^S & \beta_l & \beta_l^{AF} & \beta_l^x \\ \beta_{AF}^{nc} & \beta_{AF}^S & \beta_{AF}^l & \beta_{AF} & \beta_{AF}^x \\ \beta_x^{nc} & \beta_x^S & \beta_x^l & \beta_x^{AF} & \beta_x \end{pmatrix}; \mathbf{G} = \begin{pmatrix} \gamma_{nc} \\ \gamma_S \\ \gamma_l \\ \gamma_{AF} \\ \gamma_x \end{pmatrix}; \mathbf{B_{ANF}} = \begin{pmatrix} \beta_{ANF}^{nc} \\ \beta_{ANF}^S \\ \beta_{ANF}^l \\ \beta_{ANF}^{AF} \\ \beta_{ANF}^x \end{pmatrix}$$

auxquels s'ajoutent γ_{ANF} et β_{ANF}. La matrice $\mathbf{B}$ est symétrique par définition, de sorte que : $\beta_k^j = \beta_j^k$ $\forall k \neq j$ tels que $j, k \in \{nc, ANF, S, AF, x\}$. ■

[35] Dans ce qui suit, l'indice propre aux individus, i, est négligé par souci de clarté aussi souvent que possible.

Compte tenu de cette spécification, un médecin choisit l'alternative j si : $V_j \geq V_k, \quad \forall k \neq j$. Sa contribution individuelle à la vraisemblance est donc la probabilité de cet évènement. Si les ϵ_j, $j \in J$ sont supposés i.i.d. selon une Gumbel (distribution à valeurs extrêmes de type I), cette probabilité s'écrit[36] :

$$\begin{aligned} P(j) &= P[V_j \geq V_k, \quad \forall k \neq j] \\ &= P[\epsilon_j \geq u_k - u_j + \epsilon_k, \quad \forall k \neq j] \\ &= \frac{e^{u_j}}{\sum_{k=1}^{J} e^{u_k}} \end{aligned} \tag{2.20}$$

L'estimation des paramètres de la fonction d'utilité nécessite donc de connaître le niveau d'utilité atteint par l'individu dans chaque alternative. Pour une valeur donnée des paramètres, l'utilité correspond au niveau de bien-être associé au comportement de pratique dans l'alternative j, tel que décrit par la fonction d'utilité (2.19). Comme cette utilité dépend du niveau de consommation offert par le revenu de pratique, l'estimation du modèle requiert en particulier de générer le revenu issu de la pratique dans chaque alternative. Pour ce faire, nous utilisons la modélisation de la contrainte budgétaire présentée dans la Section 2.1, en nous appuyant sur les équations (2.1) à (2.3) pour calculer le niveau de consommation associé à toute combinaison particulière des variables de pratique.

L'ensemble de ces éléments définit un Logit polytomique. Contrairement à d'autres modèles, cette spécification impose en particulier que les termes d'erreur soient indépendants entre les alternatives. Ces derniers ne peuvent donc pas prendre en compte l'hétérogénéité inobservable propre aux individus. Comme l'a souligné l'analyse de la Section 2.2, les choix des médecins en matière de mode de rémunération reposent pourtant de façon cruciale sur la forme de leurs préférences. Il est donc particulièrement important de prendre en compte l'hétérogénéité individuelle dans un modèle destiné à analyser

[36] Voir, par exemple, Train (2003, p.78) pour une dérivation complète des probabilités d'un Logit multinomial.

ce choix. On peut par exemple s'attendre à ce que les individus en début de carrière tendent à réaliser un nombre important d'actes et d'heures de travail. Ce profil de pratique étant mieux rémunéré sous la rémunération à l'acte que sous la rémunération mixte, les individus les plus jeunes montreraient alors une plus forte propension à rester à la rémunération à l'acte. Ce type d'hétérogénéité observable est introduit dans la Section 2.4.1.

Outre ces caractéristiques individuelles observables, nous tenons compte de l'hétérogénéité inobservable en estimant la distribution des préférences des médecins de l'échantillon, plutôt que les préférences elles-mêmes. A cette fin, un certain nombre des coefficients de la fonction d'utilité (2.19) sont supposés être aléatoires. Les statistiques descriptives présentées dans la Section 2.1.3 suggèrent que les médecins appelés à choisir la rémunération mixte diffèrent de ceux qui resteront à la rémunération à l'acte principalement en termes d'heures consacrées au travail non-clinique et de quantité d'actes non-facturables réalisés. C'est donc sur l'utilité marginale de chacune de ces variables que nous permettons aux préférences de se distinguer. Dans la fonction d'utilité (2.19), les termes linéaires associés aux heures de travail non-clinique et aux actes non-facturables sont ainsi supposés suivre des lois normales : $\gamma_k \equiv N(\overline{\gamma}_k, \sigma_k)$, $k \in \{nc, ANF\}$, indépendantes entre elles et indépendantes des ϵ_j, $\forall j$. La moyenne et l'écart-type de ces variables aléatoires sont estimés conjointement avec les paramètres déterministes de la fonction d'utilité. Pour ce faire, les contributions individuelles à la vraisemblance (2.20) doivent être adaptées afin de tenir compte de l'incertitude sur les préférences. Conditionnellement aux valeurs prises par γ_{nc} et γ_{ANF}, la contribution à la vraisemblance de l'individu considéré correspond au logit polytomique décrit ci-dessus. La contribution inconditionnelle de l'individu i, qui a choisit l'alternative j_i, est alors :

$$l_i = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(j_i)\ \phi\left(\frac{\gamma_{nc} - \overline{\gamma}_{nc}}{\sigma_{nc}}\right) \phi\left(\frac{\gamma_{ANF} - \overline{\gamma}_{ANF}}{\sigma_{ANF}}\right) d\gamma_{nc}\ d\gamma_{ANF}$$

où ϕ est la fonction de densité de la loi normale centrée réduite.

L'estimation du modèle requiert donc, désormais, le calcul d'une intégrale bidimensionnelle. Pour faire l'économie de l'importante charge de calcul de l'intégration numérique, nous utilisons une méthode d'intégration par simulation. Les intégrales sont alors approximées par la valeur moyenne de $l_i|\{\gamma_{nc}, \gamma_{ANF}\}$ calculée sur r tirages aléatoires dans les distributions de γ_{nc} et γ_{ANF}. Les tirages sont réalisés en utilisant des séquences d'Halton, qui permettent de minimiser la variance de simulation pour un nombre donné de tirages, r (Train, 1999). Cette méthode d'estimation correspond alors au Maximum de Vraisemblance Simulé, qui est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance exact tant que $\sqrt{r}$ s'accroît plus rapidement que la taille de l'échantillon (Gourieroux & Monfort, 1993).

2.3.2 Aspects spécifiques

Les données utilisées pour réaliser l'estimation du modèle imposent un certain nombre d'adaptations de l'architecture de base présentée ci-dessus. Une première difficulté est imputable au niveau de finesse choisit pour la discrétisation des variables de pratique. La cohérence de la modélisation des choix discrétisés oblige en effet à former l'espace des choix en considérant l'ensemble des combinaisons possibles entre les niveaux discrétisés des variables de pratique. L'ensemble de choix inclut en conséquence un nombre important d'alternatives qui violent les contraintes de faisabilité auxquelles font face les médecins. A titre d'exemple, le modèle devrait théoriquement autoriser le choix d'une alternative dans laquelle le nombre d'actes réalisés est maximum, mais où le nombre d'heures de travail clinique est nul. Ce type de choix n'est à l'évidence jamais observé ; et s'avère, en réalité, indisponible. Afin d'alléger l'estimation, nous nous limitons donc au sous-ensemble des alternatives techniquement réalisables – au sens où elles sont choisies au moins une fois par un médecin au cours de la période d'observation – noté $J^C \subset J$. Cette stratégie d'estimation revient à conserver le même sous-ensemble

d'alternatives quelle que soit l'alternative choisie. Comme le montre McFadden (1978) cette propriété de "conditionnement uniforme" assure que l'estimation du modèle Logit reste convergente malgré la réduction de l'espace de choix. La méthode de sélection que nous utilisons est encore plus restrictive, puisqu'elle repose sur une distribution dégénérée (voir (2.21) ci-dessous) qui se limite à réaliser l'estimation sur un sous-ensemble d'alternatives constant. L'estimation du modèle selon cette stratégie reste donc convergente.

Preuve Nous adaptons ici au cas spécifique qui est le nôtre la preuve fournie par Train (2003, ch. 3) du résultat de McFadden (1978).

Nous nous intéressons à l'estimation des paramètres qui déterminent le choix à l'intérieur de l'ensemble J mais en s'appuyant sur un sous-ensemble $K \in J$. Soit $q(K|\,j)$ la probabilité que le sous-ensemble K soit utilisé pour évaluer la contribution individuelle à la vraisemblance lorsque l'alternative choisie est j. La propriété de *conditionnement uniforme* qualifie les situations dans lesquelles cette probabilité est constante pour tous les choix $j \in K$. Dans notre cas, cette distribution est dégénérée et s'écrit :

$$q(J^C\big|\,j) = \begin{cases} 1 & \text{si } K \equiv J^C \quad \forall j \in J^C \\ 0 & \text{si } K \equiv J^C \quad \forall j \notin J^C \\ 0 & \text{si } K \neq J^C \quad \forall j \in J \end{cases} \tag{2.21}$$

La probabilité inconditionnelle de choisir l'alternative j est notée P_j, définie en (2.20). Soit $P(j|\,J^C)$ la probabilité de choisir l'alternative j dans le sous-ensemble de choix J^C. En utilisant la règle de Bayes, ces probabilités sont liées par la probabilité jointe de sélectionner le sous-ensemble J^C et que l'alternative j soit choisie :

$$P(j, J^C) = q(J^C\big|\,j)\ P_j = P(j|\,J^C)\ Q(J^C) \tag{2.22}$$

où $Q(J^C) = \sum_{j \in J^C} P_j\ q(J^C\big|\,j)$ est la probabilité marginale de sélectionner le sous-ensemble J^C parmi l'ensemble des alternatives contenues dans J. La relation (2.22) se simplifie selon :

$$P(j|\,J^C) = \frac{q(J^C\big|\,j)\ P_j}{\sum_{j \in J^C} P_j\ q(J^C|_j)} = \frac{P_j}{\sum_{j \in J^C} P_j} \tag{2.23}$$

où le second membre de droite s'obtient en utilisant le conditionnement uniforme retenu dans notre application (2.21).

Dans une spécification en termes de Logit mixte, l'utilité est définie conditionnellement à l'hétérogénéité inobservable : $u_j = u_{j|\epsilon}$ puis intégrée dans sa distribution. Les probabilités conditionnelles

de l'hétérogénéité inobservable s'écrivent :

$$P_{j|\epsilon} = \frac{e^{u_{j|\epsilon}}}{\sum_{i \in J} e^{u_{i|\epsilon}}}$$

Après substitution dans (2.23), on obtient :

$$P(j|\, J^C, \epsilon) = \frac{p_{j|\epsilon}}{\sum_{j \in J^C} P_{j|\epsilon}} = \frac{\frac{e^{u_{j|\epsilon}}}{\sum_{i \in J} e^{u_{i|\epsilon}}}}{\sum_{k \in J^C} \frac{e^{u_{k|\epsilon}}}{\sum_{i \in J} e^{u_{i|\epsilon}}}} = \frac{e^{u_{j|\epsilon}}}{\sum_{i \in J^C} e^{u_{i|\epsilon}}}$$

La probabilité conditionnelle correspond donc à la probabilité d'un Logit multinomial associée au choix de l'alternative j dans l'ensemble de choix J^C. La contribution individuelle à la vraisemblance correspond à l'intégrale de cette probabilité conditionnelle dans la distribution de l'hétérogénéité. Sous hypothèse de normalité, on a :

$$P(j|\, J^C, \epsilon) = \int_{-\infty}^{\infty} \frac{e^{u_{j|\epsilon}}}{\sum_{i \in J^C} e^{u_{i|\epsilon}}}\ \phi(\epsilon) d\epsilon$$

La maximisation de la fonction de vraisemblance formée à partir de ces contributions est convergente. La réduction de l'ensemble de choix restreint cependant la quantité d'information utilisée et affecte donc l'efficacité des estimateurs. ■

La seconde difficulté est imputable à la distinction introduite par la rémunération mixte entre actes facturables et non-facturables. Si ces deux types d'actes sont rémunérés et, par conséquent, observés sous la rémunération à l'acte, les actes non-facturables sont par définition inobservables sous la rémunération mixte. Il ne donnent en effet lieu à aucune rémunération spécifique (voir Section 2.1.3 ci-dessus) et ne sont donc pas déclarés par les médecins. Un médecin qui a opté pour la rémunération mixte peut cependant être en partie rémunéré selon le mode de rémunération à l'acte, dès lors que certaines de ses heures de travail ne sont pas couvertes par un *per diem*. Dans ce cas, les actes non-facturables réalisés sont rémunérés à taux plein, donc observés. Soit m le niveau d'actes non-facturables discrétisés réalisé dans une période de travail non couverte par un *per diem*, par un médecin qui a choisit la rémunération mixte (noté $d_i = 1$). Le choix associé, ANF_m, constitue alors un plancher pour les actes non-facturables effectivement réalisés par cette observation. Ainsi, les actes non-facturables observés sont ANF_m alors même que les actes non-facturables effectifs sont $NBA \in \{NBA_m, NBA_{m+1}, \ldots, NBA_{N_{NBA}}\}$.

Pour ces observations, la fonction de vraisemblance doit donc incorporer l'incertitude quant au niveau effectivement choisi. Par définition des variables discrétisées, les niveaux de discrétisation d'une variable de pratique sont mutuellement exclusifs. La contribution à la vraisemblance d'un individu qui exerce sous la rémunération mixte et dont les choix observés sont $\{Z_j, ANF_m\}$ est donc la somme des probabilités de choix parmi les ANF_j :

$$\begin{aligned}
P\left(j|_{d_i=1}\right) &= P\left[\{Z_j, ANF_m\}|_{d_i=1}\right] = P\left(Z_j, ANF_m\right) \cup P\left(Z_j, ANF_{m+1}\right) \cup \ldots \cup P\left(Z_j, ANF_{N_{ANF}}\right) \\
&= \sum_{l=m}^{N_{ANF}} \frac{\exp\left(u(Z_j, ANF_l)\right)}{\sum_{k=1}^{JC} e^{u_k}} \\
&= \sum_{l=m}^{N_{ANF}} \frac{\exp\left(\mathbf{G}'\ Z_j + Z_j'\mathbf{B}Z_j + \gamma_{ANF} \ln ANF_l + \mathbf{B_{ANF}}'\ Z_j\ \ln ANF_l + \beta_{ANF}(\ln ANF_l)^2\right)}{\sum_{k=1}^{JC} e^{u_k}} \\
P\left(j|_{d_i=1}\right) &= \frac{\exp(\mathbf{G}'\ Z_j + Z_j'\mathbf{B}Z_j)}{\sum_{k=1}^{JC} e^{u_k}} \sum_{l=m}^{N_{ANF}} \exp\left(\gamma_{ANF} \ln ANF_l + \mathbf{B_{ANF}}'\ Z_j \ln ANF_l + \beta_{ANF}(\ln ANF_l)^2\right)
\end{aligned}$$

Pour les médecins qui ont choisit la rémunération mixte, les probabilités sont donc corrigées de façon à prendre en compte l'incertitude quant à l'alternative selectionnée au sein du sous-ensemble d'alternatives choisi. Pour les médecins qui sont restés à la rémunération à l'acte, en revanche, les actes non-facturables sont observables en toutes circonstances. Leur contribution reste donc conforme à l'expression (2.20). Au total, la probabilité que l'individu i choisisse l'alternative j devient donc :

$$P(j_i) = \left(\frac{e^{u_{j_i}}}{\sum_{k=1}^{JC} e^{u_k}}\right)^{1-d_i} P\left[\{Z_j, ANF_m\}|_{d_i=1}\right]^{d_i} \tag{2.24}$$

En résumé, nous estimons un logit mixte dont la vraisemblance s'écrit : $\prod_{i=1}^{N} l_i$ où l_i est décrit par (2.21) et $P(j_i)$, la probabilité de choisir l'alternative j_i, par (2.24). La spécification retenue comprend deux paramètres aléatoires et 25 paramètres constants. L'estimation comporte donc 29 paramètres à estimer – auxquels s'ajoutent les paramètres d'hétérogénéité individuelle, voir Section 2.4.1 – en utilisant 12 842 observations du comportement de pratique des médecins spécialistes du Québec. Le modèle est estimé par la méthode du maximum de vraisemblance simulé. Pour chaque paramètre aléatoire, nous réalisons 20 tirages d'Halton spécifiques à chaque individu. La fonction de vraisemblance est alors évaluée, pour chaque tirage, en calculant le niveau d'utilité atteint par l'individu dans chacune des J^C alternatives. Compte tenu de ces caractéristiques, l'estimation du modèle requiert d'importantes capacités tant de calcul que de mémoire.[37] L'estimation est rendue possible par la parallélisation des calculs (Swann, 2002), qui consiste à répartir l'évaluation de la vraisemblance entre plusieurs processeurs (20 ici). Le programme, présenté dans l'Annexe (Section 2.A), a été développé en langage Ox (Doornik & Ooms, 2001 ; Cribari-Neto & Zarkos, 2003).

2.4 Résultats : les vertus de la flexibilité

Le modèle est identifié empiriquement grâce aux variations de prix induites par la rémunération mixte. A cette fin, des observations sur les comportements de pratique avant et après la réforme ont été recueillies. Ces données doivent subir un certain nombre de transformations pour fournir le pendant empirique de la modélisation que nous avons adopté. Les résultats d'estimation et, en particulier, les simulations réalisées montrent que la liberté d'adoption de la rémunération mixte est un élément clé de son succès.

[37] En utilisant les niveaux de discrétisation présentés dans le Tableau 2.6, chaque itération nécessite ainsi le calcul de plus de 640 niveaux d'utilité par individu.

2.4.1 Présentation des données

Les données que nous utilisons recouvrent les comportements de pratique et un certain nombre de caractéristiques individuelles de l'ensemble des médecins spécialistes exerçant au Québec entre 1996 et 2002. Elles résultent de la combinaison des informations fournies par deux institutions Québécoises : le *Collège des médecins du Québec* (CMQ) et la *Régie d'Assurance Maladie du Québec* (RAMQ).

Le CMQ est l'organisation professionnelle représentative des médecins du Québec. Il réalise chaque année une enquête auprès de ses membres, destinée à recueillir de l'information sur leurs caractéristiques individuelles (telles que la spécialisation, l'âge ou le sexe), les caractéristiques institutionnelles et géographiques de leur établissement de rattachement ainsi que l'allocation de leur temps de travail. Les médecins sont ainsi appelés à évaluer le temps qu'ils consacrent à leur travail en termes d'heures (nombre moyen d'heures hebdomadaires) et de semaines (nombre annuel), puis la répartition de ces heures, en pourcentage, entre le temps consacré aux patients – activités cliniques – et le temps consacré respectivement à l'enseignement, aux activités administratives et à la recherche. Malgré la remarquable stabilité du questionnaire au cours des années, la question portant sur les semaines de travail n'apparaît qu'en 1996, 1997, 1998 et 2002. Ce changement temporaire dans la collecte des données nous oblige donc à abandonner les observations couvrant la période 1999-2001, pour lesquelles les semaines de travail sont manquantes. La rémunération mixte ayant été introduite au quatrième trimestre de l'année 1999, nous sommes donc conduits à abandonner les 3 années qui suivent immédiatement la réforme. Les choix de pratique que nous conservons après la réforme résultent donc d'une longue période d'ajustement au nouveau mode de rémunération.

La RAMQ est l'organisation publique en charge de la rémunération des médecins au Québec. A ce titre, elle reçoit de chaque praticien une déclaration décrivant son activité professionnelle, à partir de laquelle la rémunération est calculée. Les données de prix et de productivité (nombre d'actes réalisés) que nous en obtenons, sur une

base trimestrielle, sont donc très peu sujettes à des problèmes d'erreur de mesure. Ces données administratives offrent en particulier une description parfaite du mode de rémunération sous lequel se déroule la pratique. Conformément au modèle théorique, nous nous concentrons sur le choix entre la rémunération à l'acte et la rémunération mixte. Seuls les médecins (62.68% de l'échantillon, correspondant à 12,819 observations de décisions annuelles) dont l'intégralité du revenu provient de l'un ou l'autre de ces modes de rémunération sont donc conservés dans l'estimation.

Les deux ensembles de données sont combinés grâce à un identifiant codé, propre à chaque médecin. Cette variable nous permet également de retracer les choix multiples d'un même individu entre les périodes. Les choix de pratique des 4544 médecins spécialistes retenus sont donc observés en panel de 1996 à 1998 et en 2002.

2.4.2 Construction des variables

Ces données doivent subir un certain nombre de transformations avant de fournir des mesures en adéquation avec notre modélisation de la marge intensive (heures et semaines de travail), de la marge extensive (quantités d'actes) et des paramètres de la contrainte budgétaire.

a) Heures de travail : mesures de la marge extensive

Pour les années retenues dans l'estimation (1996-1998 et 2002), le nombre de semaines de travail est directement disponible. Les semaines de loisir annuelles – qui sont l'argument de la fonction d'utilité estimée – sont calculées par différence, sur la base de 52 semaines par an.

Les variables d'heures hebdomadaires sont calculées en multipliant les heures to-

tales par le pourcentage de temps consacré à chaque activité. Le pourcentage consacré aux activités cliniques permet ainsi, sans autre transformation, d'obtenir une mesure du nombre d'heures cliniques réalisées. D'après notre définition – qui recouvre celle des activités admissibles au *per diem* sous la rémunération mixte, voir Section 2.1.3 – les heures non-cliniques regroupent l'enseignement et les tâches administratives, mais excluent le temps consacré à la recherche. Notre mesure des heures totales de travail diffère donc de celles qu'ont déclaré les médecins, puisqu'elles correspondent à la somme des heures de travail clinique et non-clinique et ignorent par conséquent le temps consacré à la recherche.

Prendre en compte les changements de rémunération introduits par la rémunération mixte nécessite en outre de ventiler les variables de pratique selon le mode de rémunération sous lequel elles ont exercées (comme l'indique, par exemple, l'expression (2.3)). Un médecin qui a choisi la rémunération mixte est en effet rémunéré selon le mode de rémunération à l'acte pour toutes les heures de pratique qui n'ont pas donné lieu au versement d'un *per diems*. Dans le cadre de notre modélisation en termes de choix discrets, cette disposition impose de prédire le nombre de *per diem* reçus dans chaque alternative. Pour ce faire, nous utilisons une approximation fondée sur la part des heures de travail qui peuvent, compte tenu des restrictions imposées par la rémunération mixte, être admises au *per diem*[38]. Cette proportion, notée θ_i pour chaque individu i, est définie formellement par l'expression : $\theta_i = \dfrac{\overline{d}.N_i}{h^c + h^o}$, où N_i, le nombre de *per diems* hebdomadaire moyen, est définit par (2.2).

[38]la solution exacte, mais extrêmement coûteuse en termes de complexité du modèle, consiste à distinguer les variables de pratique selon le mode de rémunération sous lequel elles ont été exercées. Nous avons fait le choix de la simplicité.

b) Actes : mesure de la marge intensive

Outre les heures de travail clinique, notre analyse intègre l'ajustement des marges intensives par le biais de la quantité d'actes accomplis. Chaque médecin réalise en général une grande variété d'actes, qui différent considérablement tant en termes de temps qu'en termes d'effort (attention, expertise, ...). Les taux de rémunération des actes sont ajustés en conséquence et reflètent, au moins en partie, cette diversité. Afin d'obtenir une mesure du nombre d'actes qui soit à la fois unique et fidèle à l'intensité de la pratique, nous avons donc construit un indice de quantités où, pour chaque type d'acte (correspondant, formellement, à un code d'acte dans la taxinomie de la RAMQ), la quantité délivrée est pondérée par son taux de rémunération.[39]

Pour constituer une mesure fiable du nombre d'actes fourni, les indices de quantités doivent être préservés des variations dues à l'évolution des prix. Les pondérations sont donc maintenues constantes, en utilisant le prix des actes à une année de base (1996). Au cours des six années d'observation, pourtant, de nombreux actes apparaissent ou deviennent, au contraire, obsolètes en raison du progrès des connaissances médicales. Une seconde année de base est donc utilisée (2000) et les indices de quantité prennent alors la forme d'indices de Laspeyres chaînés.[40] Les dispositions de la rémunération mixte nous conduisent, par ailleurs, à considérer deux mesures d'actes, selon qu'ils sont ou non facturables sous ce mode de rémunération. Bien que cette distinction engendre également des différences dans les taux de rémunération des actes (voir Section 2.1.3), les pondérations utilisées sont également maintenues constantes pour les deux indices (égales au prix qui rémunère les actes, à l'année de base, sous le mode rémunération à

[39] A titre d'illustration, un dermatologue qui réalise 4 visites primaires et 6 visites de contrôle totaliserait, en l'absence de pondération, 10 actes. Une visite primaire nécessite pourtant un entretien approfondi avec le patient ainsi qu'un diagnostic complet, et dure en moyenne 45 minutes, tandis qu'une visite de contrôle se limite en général à une vingtaine de minutes. Les rémunérations de ces actes reflètent ces différences, puisqu'elles sont respectivement, en 1996, de 47$ et 16.50$.

[40] Le chaînage permet de convertir les indices calculés selon les prix de la seconde année en indices basés sur la première. Diewert (1993) propose une présentation détaillée de cette technique.

l'acte).

Formellement, le nombre d'actes réalisés par le médecin i à la période t, $A_i^t = \{ANF_i^t, AF_i^t\}$ est alors mesuré par la variable :

$$A_i^t = \begin{cases} \sum\limits_{a \in \mathscr{A}} A_{a,i}^t \; p_{a_s}^{1996} & \text{si } 1996 \leq t < 2000 \\ \sum\limits_{a \in \mathscr{A}} \left(A_{a,i}^t \; p_{a_s}^{2000} \right) \dfrac{\sum\limits_{a \in \mathscr{A}} A_{a,i}^{2000} \; p_{a_s}^{1996}}{\sum\limits_{a \in \mathscr{A}} A_{a,i}^{2000} \; p_{a_s}^{2000}} & \text{si } 2000 \leq t \leq 2002 \end{cases} \tag{2.25}$$

Notations $p_{a_s}^t$ désigne le prix, à la période t et sous la rémunération à l'acte, de l'acte a lorsqu'il est réalisé par un médecin de la spécialité s ; $A_{a,i}^t$ le nombre d'actes de type a réalisés par le médecin i à la période t. Les pondérations restant inchangées que les actes soient facturables ou non, la variable A_i^t désigne indifféremment les premiers, $A_i^t = ANF_i^t$, ou les seconds, $A_i^t = AF_i^t$. Seul le groupe d'actes considéré, $\mathscr{A}$, s'en trouve affecté. Il regroupe l'ensemble des actes pour lesquels le taux de réduction sous la rémunération mixte, α, est strictement positif, dans le calcul de l'indice d'actes facturables : $\mathscr{A}_F = \{a : \alpha_a > 0\}$; et l'ensemble des actes pour lesquels le taux de réduction sous la rémunération mixte est strictement nul, dans le calcul de l'indice d'actes non-facturables : $\mathscr{A}_{NF} = \{a : \alpha_a = 0\}$ ■

c) Prix des actes : simulation du revenu potentiel

Le niveau de consommation (*i.e.* le revenu réel) associé à chaque alternative est calculé en utilisant la contrainte budgétaire développée dans la Section 2.1. Pour un mode de rémunération $d_i \in \{0; 1\}$ donné, le revenu potentiel correspond donc au bénéfice tiré des variables de pratique, décrit par la contrainte (2.3).

La partie du revenu qui provient de la rémunération des actes résulte, en particulier, du produit entre le nombre d'actes réalisés et le prix des actes sous le mode de rémunération choisi. Ce calcul nécessite ainsi de disposer d'une variable reflétant les taux de

rémunération des actes, à chaque période et sous chaque mode de rémunération, qui soit cohérente avec la mesure utilisée pour les quantités (nombre d'actes réalisés, voir ci-dessus). A cette fin, les prix des actes sont agrégés sous forme d'indices de prix.

Dans le calcul de ces indices, le prix est pondéré par le nombre moyen d'actes à l'une des années de base (1996 ou 2000), reflétant ainsi la valorisation monétaire d'un profil de pratique représentatif. Le choix de la pondération répond au souci d'isoler les mesures de prix des variations de pratique dues au passage à la rémunération mixte. Ainsi, seuls les médecins qui sont restés à la rémunération à l'acte sont pris en compte pour calculer les quantités moyennes à l'année de base. De plus, ces même pondérations sont utilisées pour calculer l'indice de prix sous la rémunération à l'acte comme sous la rémunération mixte.

En notant p_s^t l'indice de prix auquel font face les médecins de la spécialité s à la période t (où $p = P$ pour la rémunération à l'acte et $p = (1-\alpha)\ P = PF$ pour la rémunération mixte), le revenu tiré des actes réalisés est alors mesuré par : $A_i^t\ \Delta P_s^t$.

Preuve Les indices de prix de la spécialité s à la période t, p_s^t, sont mesurés par :

$$p_s^t = \begin{cases} \sum\limits_{a_s \in \mathscr{A}} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^t & \text{si } 1996 \leq t < 2000 \\ \sum\limits_{a_s \in \mathscr{A}} (\overline{A}_{a_s}^{\,2000}\ p_{a_s}^t)\ \dfrac{\sum\limits_{a_s \in \mathscr{A}} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^{2000}}{\sum\limits_{a_s \in \mathscr{A}} \overline{A}_{a_s}^{\,2000}\ p_{a_s}^{2000}} & \text{si } 2000 \leq t \leq 2002 \end{cases} \tag{2.26}$$

où $\overline{A}_{a_s}^{\,t}$ désigne le nombre moyen d'actes de type a réalisés, à la période t, par les médecins de la spécialité s qui sont restés à la rémunération à l'acte pendant l'ensemble de la période d'observation (1996-2002). Ces pondérations sont utilisées pour calculer l'indice de prix sous la rémunération à l'acte, $p_s^t = P_s^t$, comme sous la rémunération mixte $p_s^t = (1-\alpha_s^t)\ P_s^t = PF_s^t$. Seul le groupe d'actes considéré, $\mathscr{A}$, est adapté en fonction de l'indice calculé ($\mathscr{A} = \mathscr{A}_F$ dans le calcul de l'indice de prix des actes facturables, $\mathscr{A} = \mathscr{A}_{NF}$ dans le calcul de l'indice de prix des actes non-facturables). La seconde équation dans (2.26) reflète le chaînage entre les deux années de base.

La rapport $\Delta P_s^t = \dfrac{\sum\limits_{a_s} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^t}{\sum\limits_{a_s} \overline{A}_{a_s}^{\,1996}\ p_{a_s}^{1996}}$ mesure donc la revalorisation nominale subie par le panier

d'actes entre la première année de base (1996) et l'année en cours (t). Le revenu tiré des actes correspond au nombre d'actes mesurés aux prix de 1996 et réévalués selon cette mesure : $A_i^t \ \Delta P_s^t = \sum_a A_{a,i}^t \ p_{a_s}^{1996} \ \Delta P_s^t$. ■

Dans le cas de la rémunération à l'acte ($d_i = 0$) cette quantité suffit à décrire le revenu potentiel du médecin i. La quantité totale d'actes réalisés correspond en effet, dans ce cas, à la somme des actes facturables et non-facturables ; et le revenu potentiel dans l'alternative j sous la rémunération à l'acte est la valeur monétaire de l'ensemble des actes réalisés, soit : $X_{j,t}^{RA} = \left(AF_j^t + ANF_j^t\right) \ \Delta P^t$.

Le revenu sous la rémunération mixte, quant à lui, tient compte tant des actes réalisés que des heures de travail. Il concorde cependant avec le revenu de la rémunération à l'acte pour la partie de la pratique qui n'est pas incluse dans un *per diem*. Le calcul du revenu nécessite donc, dans ce cas, de ventiler les variables de pratique en fonction du mode de rémunération sous lequel elles ont exercées. Nous adoptons pour ce faire l'approximation présentée plus haut, fondée sur la proportion des heures de travail qui sont admissibles à un *per diem*, θ. Le revenu potentiel associé à l'alternative j pour un médecin qui a choisi la rémunération mixte est donc : $X_{j,t}^{RM} = S_j^t N_j D + \theta_{j,t} \ AF_j^t \ PF^t + (1 - \theta_{j,t})(AF_j^t + ANF_j^t)P^t$ où le nombre de *per diems* dans l'alternative j, N_j, est définit par (2.2).

Les variables ainsi définies nous permettent donc de calculer le revenu potentiel dans l'alternative j sous chacun des modes de rémunération disponibles. Dans la mesure où le passage à la rémunération mixte est un choix volontaire de la part du médecin (ou, en tout cas, supposé tel, voir Note (20)), nous retenons dans chaque alternative le revenu maximum parmi ceux qui résultent des modes de rémunération disponibles (qui se réduisent à la rémunération à l'acte jusqu'en 1999). Le choix du mode de rémunération est donc une variable implicite de notre modèle, représenté par la valorisation optimale des choix de pratique. Cette stratégie consiste en effet à retenir le choix de rémunération individuellement rationnel, puisque les variables de pratique sont fixes dans

une alternative donnée. Si, comme on doit s'y attendre, l'utilité marginale du revenu est positive, les médecins doivent donc, pour des choix de pratique donnés, adopter le mode de rémunération qui maximise le revenu potentiel. La rémunération potentielle dans l'alternative j est donc : $X_{j,t} = \max\left\{X_{j,t}^{RA}; X_{j,t}^{RM}\right\}$, où $X_{j,t}^{RM} = 0$ si $t \leq 1999$.

d) Plafonds et taux de rémunération : simulation du revenu effectif

Dans l'analyse traditionnelle de la théorie du consommateur, c'est par l'intermédiaire de la consommation que le revenu influence le bien-être des agents. A ce titre, la fonction d'utilité que nous estimons ne dépend pas du revenu potentiel mais du revenu effectif, qui correspond au revenu effectivement versé aux praticiens. Comme nous l'avons vu plus haut (Section 2.1.2), ces quantités diffèrent sensiblement en raison, notamment, des mesures de différenciation et de plafonnement des rémunérations. Le revenu ajusté qui en résulte doit en outre être corrigé de l'inflation afin de fournir une mesure réelle plutôt que nominale de la consommation.

Les dispositions qui gouvernent le calcul des taux de rémunération différenciée font intervenir un large éventail de caractéristiques individuelles (telles que la région et la ville de pratique) dont certaines nous sont indisponibles. Nos données contiennent cependant le niveau de revenu trimestriel de chaque médecin avant et après application du taux de rémunération. Pour chacun d'entre eux, nous utilisons donc une approximation du taux individuel, τ_i, calculée à partir du rapport moyen, sur l'ensemble de la période, entre ces deux niveaux de revenu.

Les seuils au-delà desquels la rémunération est soumise aux mesures de plafonnement sont propres aux spécialités de pratique, et constituent une information publique, fournie par la RAMQ (le détail de ces montants est fourni dans la Section 2.1.2). Les modalités de leur mise en œuvre nous obligent cependant à définir un plafond propre à chaque individu, fondé sur son profil de pratique moyen au cours de la période. Ces dis-

positions d'application dépendent en effet de façon très importante de l'établissement médical où s'est déroulée la pratique. Ainsi, les revenus issus des activités en urgence sont exclus de l'assiette de calcul du plafond de 1996 à 2001. A partir de 2001, cette exclusion s'étend à l'ensemble des revenus liés à la pratique en hôpital. Pour tenir compte des frais professionnels, les revenus provenant de la pratique en cabinet privé sont en outre diminués d'une proportion fixe.

Un traitement exact de ces dispositions aurait, une fois encore, nécessité de distinguer les variables de pratique en fonction de l'établissement ou elles s'exercent, multipliant d'autant le nombre d'alternatives. Par souci de simplicité, nous avons plutôt choisi d'ajuster le niveau des plafonds en utilisant la ventilation moyenne du revenu entre les établissements. Cette méthode permet alors de définir un plafond virtuel, propre à l'individu, correspondant au plafond auquel le revenu de l'individu est effectivement soumis compte tenu de la répartition de ses activités entre les établissements. La mesure de plafond utilisée, $C_{i,t}$, est alors formellement définie par :

$$C_{i,t} = \frac{\widetilde{C}_{s,t}}{s_i^e + s_i^p \ (1 - a_s)}$$

Notations Les dispositions légales qui gouvernent l'application des plafonds sont décrites par le seuil applicable à la spécialité du professionnel, s, à la période t, noté $\widetilde{C}_{s,t}$ et le taux de réduction appliqué aux revenus en cabinet privé, a_s ($a_s = 35\%$ pour toutes les spécialités à l'exception de la radiologie diagnostique qui bénéficie d'une réduction de $a_s = 75\%$).

La répartition des activités du médecin i sur l'ensemble de la période est décrite par la ventilation de son revenu, R_i, entre le revenu issu de la pratique en cabinet privé, R_i^p, et les revenus hors cabinet privé admissibles au plafond (*i.e.* excluant les revenus à l'urgence, ainsi que les revenus hospitaliers à partir de 2001), R_i^e. Pour chaque individu, cette décomposition permet de définir les parts de la pratique réalisées dans chaque type d'établissement comme : $s_i^e = R_i/R_i^e$ et $s_i^p = R_i/R_i^p$. Le revenu du médecin i à la période t est alors soumis au plafond si : $s_i^e.R_{i,t} + s_i^p.R_{i,t}(1-a) \geq \widetilde{C}_{s,t}$. Le plafond virtuel qui s'applique au revenu global est donc : $R_{i,t} \geq \dfrac{\widetilde{C}_{s,t}}{s_i^e + s_i^p.(1-a)} \equiv C_{i,t}$. ■

Tableau 2.7 – Prédiction de la consommation effective

Variable	Coefficient	(Ecart-type)
Revenu prédit	0.97***	(0.005)
Constante	43081.77***	(4336.695)
R^2	0.83	

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Régression linéaire. La variable endogène est la consommation effective observée du médecin, la variable *Revenu prédit* correspond à la consommation effective simulée pour les choix observés.

L'ensemble de ces variables fournit, pour un revenu potentiel donné, le revenu effectif de chaque observation dans chaque alternative selon l'expression (2.1). Ce revenu potentiel nominal est enfin converti en revenu réel, en utilisant les données d'inflation fournies par *Statistique Canada*.[41] Le taux d'inflation annuel moyen pour l'ensemble de la période est de 1.92%.

L'ensemble de ces variables permet de construire la contre-partie empirique de la contrainte budgétaire des médecins. Suivant en cela une longue tradition en économie appliquée de l'offre de travail (Blundell & MaCurdy, 1999), ces variables sont en effet utilisées pour simuler le niveau de consommation dans toute alternative. Pour l'alternative choisie, la comparaison avec le revenu effectivement obtenu par le praticien fournit une évaluation de la qualité de cette prévision. A cette fin, le Tableau 2.7 présente les résultats de la régression de la consommation effective observée sur sur son niveau prédit par le modèle dans l'alternative choisie. La qualité de la modélisation peut être évaluée selon deux dimensions. D'un part, la consommation prédite recouvre une large proportion des variations de la consommation effective (83%). D'autre part, le pouvoir explicatif de la consommation prédite, mesuré par son coefficient dans la regression (0.97), est très proche d'une prévision parfaite (pour laquelle le coefficient serait égal à 1, indiquant que toute variation de la consommation prédite recouvre une variation

[41] Les données sont librement disponibles sur le site web de l'institution. Nous utilisons l'indice des prix à la consommation annuel pour le Québec.

identique de la consommation effective observée).

2.4.3 Résultats d'estimation

Cette contrainte budgétaire "empirique" permet d'évaluer la fonction d'utilité des médecins dans chaque alternative. L'estimation du modèle présenté dans la Section 2.3 consiste à retenir la combinaison de paramètres qui rendent optimale l'alternative choisie. A titre préliminaire, le modèle est estimé sur le sous-échantillon des chirurgiens. Ces préférences participent à lever les indéterminations du modèle théorique de la Section 2.2, en fournissant une appréciation de l'élasticité empirique des choix de pratique aux variations des incitations. Elles permettent également d'anticiper l'effet sur l'offre de soins produit par tout changement des paramètres de la contrainte budgétaire. Les modalités d'instauration de la rémunération mixte choisies par les autorités du Québec peuvent ainsi être comparées à des dispositions alternatives, telles que sa surpression ou sa généralisation.

a) Préférences estimées

Les chirurgiens représentent 9.65% (1237 observations) des observations de l'échantillon, regroupant 495 individus. Avec un taux d'adhésion à la rémunération mixte de plus de 36% en 2002 (voir Tableau 2.1), ce sous-échantillon présente notamment l'avantage de regrouper des individus aux choix de rémunération très variables. Le Tableau 2.8 présente les profils de pratique moyens au sein de cette spécialité selon le mode de rémunération choisi (ou imposé avant 1999). Cette variabilité des choix de rémunération semble s'appuyer sur une importante diversité des comportements de pratique, qui recouvre les différences commentées plus haut (Section 2.3). Une exception notable est l'apparition d'un effet de sélection en termes d'heures de travail non-clinique. Cet aspect est intégré dans le modèle économétrique où l'hétérogénéité inobservable vis-

Tableau 2.8 – Profil de pratique des chirurgiens

		h	h^c	h^{nc}	W	AF^a	ANF^a
Médecins	Avant 1999	59.19	49.12	10.07	45.52	172.46	16.96
RM	2002	52.44	46.65	5.78	44.03	117.68	9.57
Médecins	Avant 1999	54.02	46.18	7.84	44.86	142.63	33.95
RA	2002	53.03	48.87	4.15	45.26	159.90	35.21
Total		54.43	47.05	7.38	44.94	147.19	29.61

[a] En milliers de Dollars constants (base 1996).

Note. *Moitié supérieure* : Profil de pratique moyens des chirurgiens qui ont obtenu une partie de leur rémunération sous la rémunération mixte au cours de la période, avant (première ligne) et après (deuxième ligne) l'avoir adoptée. *Moitié inférieure* : profil de pratique des chirurgiens dont 100% du revenu provient de la rémunération à l'acte, avant (troisième ligne) et après (dernière ligne) l'introduction de la réforme.

à-vis de ces heures de travail est prise en compte par un coefficient aléatoire.

Les résultats d'estimation des préférences des chirurgiens, décrites par la fonction d'utilité (2.19), sont présentées dans le Tableau 2.9. L'hétérogénéité inobservable est introduite progressivement dans les spécifications 2 (heures de travail non-cliniques) et 3 (actes non-facturables). Comme on pouvait s'y attendre, la prise en compte de l'hétérogénéité inobservable permet au modèle de décrire de mieux en mieux les préférences des individus de l'échantillon. Nous retenons en conséquence la spécification 3, qui incorpore l'hétérogénéité par rapport aux heures de travail non-clinique et aux actes non-facturables.

La qualité de l'estimation peut être appréciée par la capacité des préférences estimées à recouvrir la distribution des pratiques réelles. A cette fin, le Tableau 2.10 compare les prédictions du modèle estimé pour le comportement en 2002 (colonne centrale) aux comportements observés à la fois sur l'ensemble de la période (première colonne) et en 2002, qui est l'unique année de réforme incluse dans l'échantillon. Dans l'ensemble, le modèle recouvre avec une précision très satisfaisante les variations du comportement de pratique. La diminution de l'effort et l'accroissement important de la consommation en 2002, par rapport à leur niveau sur l'ensemble de la période, sont en particulier

Tableau 2.9 – Paramètres estimés de la fonction d'utilité Translog

	Specification 1		**Specification 2**		**Specification 3**	
	Coef. Estimé	t de Student	Coef. Estimé	t de Student	Coef. Estimé	t de Student
$\gamma^{nc}, \overline{\gamma}^{nc}$	9.100	9.48***	9.543	9.82***	9.547	9.78***
σ_{nc}	.	.	1.040	8.14***	0.879	9.18***
γ^{L}	3.526	2.58***	3.540	2.59***	3.568	2.61***
γ^{l}	207.237	17.19***	206.923	16.43***	206.409	15.78***
$\gamma^{ANF}, \overline{\gamma}^{ANF}$	5.716	6.50***	5.720	6.47***	7.703	6.23***
σ_{ANF}	.	.	.	.	0.979	2.65***
γ^{AF}	4.011	6.79***	3.989	6.83***	3.965	6.38***
γ^{X}	-1.195	1.23	-1.251	1.29*	-1.324	1.29*
β_{L}^{nc}	-0.060	2.03**	-0.061	2.05**	-0.061	2.06**
β_{l}^{nc}	-1.545	8.03***	-1.552	7.95***	-1.558	7.91***
β_{ANF}^{nc}	-0.036	5.19***	-0.038	5.30***	-0.037	5.19***
β_{AF}^{nc}	-0.001	0.13	-0.005	0.51	-0.002	0.16
β_{X}^{nc}	-0.017	1.39*	-0.011	0.88	-0.014	0.95
β_{l}^{L}	-0.147	0.54	-0.151	0.55	-0.158	0.58
β_{ANF}^{L}	-0.014	1.31*	-0.014	1.32*	-0.016	1.36*
β_{AF}^{L}	-0.012	0.88	-0.012	0.90	-0.013	0.92
β_{X}^{L}	0.013	0.77	0.014	0.81	0.014	0.85
β_{ANF}^{l}	-0.071	1.30*	-0.072	1.32*	-0.076	1.39*
β_{AF}^{l}	-0.538	5.64***	-0.542	5.56***	-0.546	5.63***
β_{X}^{l}	0.084	0.69	0.093	0.77	0.095	0.78
β_{AF}^{ANF}	-0.008	2.29**	-0.009	2.59***	-0.007	1.63*
β_{x}^{ANF}	0.050	1.16	0.060	1.43*	0.075	1.74**
β_{X}^{AF}	-0.029	0.50	-0.019	0.33	-0.029	0.45
β^{nc}	-0.644	11.98***	-0.876	16.42***	-0.834	15.60***
β^{L}	-1.009	10.38***	-1.007	10.36***	-1.006	10.34***
β^{l}	-21.356	17.28***	-21.323	16.51***	-21.262	15.85***
β^{ANF}	-0.373	11.49***	-0.381	11.71***	-0.395	11.31***
β^{AF}	-0.227	13.26***	-0.235	13.65***	-0.333	6.67***
β^{X}	0.069	1.37*	0.070	1.42*	0.081	1.46*

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Logit mixte, estimé par le maximum de vraisemblance simulé. La forme fonctionnelle estimée est décrite par (2.19). L'hétérogénéité inobservable est prise en compte par le coefficient de la partie linéaire de la fonction d'utilité, en supposant une distribution normale : $\gamma_k \equiv N(\overline{\gamma}_k, \sigma_k)$. *Spécification 1* : Logit multinomial ; *Spécification 2* : heures de travail non-clinique ; *Spécification 3* : heures de travail non-clinique et actes non-facturables.

Tableau 2.10 – Qualité du modèle estimé

	Observé Ensemble	Prédit 2002	Observé 2002
Heures hebdomadaires totales	54.62	55.92	53.04
——— cliniques (h^c)	47.21	48.77	48.70
——— non cliniques (h^{nc})	7.42	7.16	4.33
Semaines (W)	45.96	46.28	45.71
Actes[a] totaux	165.59	167.55	163.05
——— facturables (AF)	144.52	145.03	144.55
—— non facturables (ANF)	21.07	22.52	18.51
Effort $\left(e = \frac{ANF + AF}{h^c * W}\right)$	76.33	74.23	73.24
Revenu annuel[a] (X)	169.29	228.93	222.92

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. Comportements de pratique moyens observés sur l'ensemble de la période (première colonne) ou en 2002 (dernière colonne) et comportements de pratique moyens prédits par le modèle estimé pour l'année 2002 (colonne centrale).

correctement prédites à partir des préférences estimées. Les prédictions du modèle restent cependant fortement influencées par les comportements réels qui ont permis son estimation. Ainsi, le modèle prévoit difficilement la chute importante des heures non cliniques observée en 2002 (4.3 heures en moyenne pour cette année, contre plus de 7 pour l'ensemble de la période) et tend à prédire un comportement proche de celui qui a valu pendant l'ensemble de la période. Cette imprécision se reporte sur les heures de travail totales, légèrement surestimées elles-aussi.

Les propriétés dérivées de ces préférences estimées tendent également à en confirmer la validité. En raison de la forme analytique de la fonction d'utilité Translog, il faut cependant noter que les effets marginaux dépendent non seulement des paramètres de la fonction d'utilité, mais également du niveau des variables de pratique. Nous fournissons donc une évaluation des propriétés locales moyennes des préférences estimées, évaluées en utilisant pour chaque individu le niveau observé des variables de pratique. Nous

utilisons pour ce faire l'année 1998, exempte de la rémunération mixte. Le Tableau 2.11 présente en particulier la distribution de l'échantillon en termes d'utilité marginale en fonction du sexe des individus. L'estimation par discrétisation, adoptée ici, exclu *a priori* les points intérieurs de l'ensemble budgétaire. La cohérence de la méthode nécessite donc que l'utilité marginale du revenu soit positive (van Sœst, 1995, p.68). Quel que soit le sexe des individus, cette hypothèse est respectée par une écrasante majorité des observations (première colonne, 99.7% au total). L'utilité marginale du loisir (deuxième colonne) reflète une forte préférence des chirurgiens en faveur du travail. Ce résultat est assez intuitif au regard de la part qu'occupe le travail dans une semaine-type d'activité (60 heures de travail hebdomadaires en moyenne, Tableau 2.8). Les préférences à l'égard des heures de travail non-clinique (dernière colonne), enfin, confirment assez largement le cadre adopté dans l'analyse théorique, considérant ces heures de travail comme une forme particulière de loisir. Un écart important apparaît en fonction du sexe, les femmes manifestant une préférence beaucoup plus forte pour ces activités d'administration et d'enseignement.

L'effet propre de la rémunération mixte sur l'offre de soins dépend de la réponse des variables de pratique à la variation simultanée du taux de rémunération des actes et du *per diem*. La prochaine section en propose une évaluation, à partir de simulations de dispositifs alternatifs de rémunération. Elles permettent en particulier de lever les ambiguïtés mises en évidence par l'analyse théorique.

Tableau 2.11 – Utilités marginales

	X	l	h^{nc}
Femme	99.86	27.67	45.91
Homme	99.74	29.20	39.91
Total	99.77	28.85	41.27

Note. Proportion des individus de l'échantillon pour lesquels l'utilité marginale du revenu (1° colonne), du loisir (2°) et des heures non-cliniques (3°) est positive, en fonction du sexe. En %.

Tableau 2.12 – Variations induites par l'introduction de la rémunération mixte

	RA	RM volontaire	Variation
Heures hebdomadaires totales	54.29	55.92	3 %
——— cliniques (h^c)	47.21	48.77	3.3 %
——— non cliniques (h^{nc})	7.08	7.16	1.1 %
Semaines (W)	45.99	46.28	.6 %
Actes[a] totaux	176.70	167.55	-5.2 %
——— facturables (AF)	152.33	145.03	-4.8 %
—— non facturables (ANF)	24.36	22.52	-7.6 %
Effort $\left(e = \frac{ANF + AF}{h^c * W}\right)$	81.38	74.23	-8.8 %
Revenu annuel[a] (X)	161.92	228.93	41.4 %

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. Comportements de pratique moyens prédits par le modèle pour l'année 2002 selon que le schéma de rémunération correspond à la rémunération à l'acte (colonne de gauche) ou au dispositif de rémunération mixte existant (colonne centrale). La variation (dernière colonne) correspond au taux de variation relatif entre la première et la deuxième colonne.

b) Simulations

L'estimation d'un modèle structurel permet en effet de générer les réponses optimales des médecins à tout changement hypothétique de la contrainte budgétaire. les paramètres estimés permettent d'identifier l'alternative qui rend maximale l'utilité sous la contrainte budgétaire supposée ; les comportements de pratique simulés correspondent alors aux choix de pratique inclus dans cette alternative. Les simulations sont réalisées pour l'année 2002, qui est la seule année d'existence de la rémunération mixte incluse dans l'échantillon.

Le Tableau 2.12 compare les comportements de pratique simulés de l'ensemble des médecins selon que la rémunération mixte existe dans ses dispositions actuelles (colonne centrale) ou que le dispositif pré-réforme a été maintenu, contraignant l'ensemble des

médecins de l'échantillon à conserver la rémunération à l'acte (première colonne). Par définition, seuls le médecins qui choisissent librement d'adopter la rémunération mixte lorsqu'elle est disponible sont affectés par ce changement. La variation induite par l'introduction de la rémunération mixte (dernière colonne) correspond donc à l'effet propre de la réforme sur les comportements de pratique des chirurgiens du Québec. La rémunération mixte engendre d'abord un important accroissement de revenu (plus de 41%) pour les médecins qui la choisissent. Comme nous l'avons vu (Tableau 2.3), ces médecins se caractérisent par un niveau de revenu plus faible à heures de travail comparables. En ce sens, il semble donc que la rémunération mixte parvienne à rétablir l'équité des rémunérations offertes aux médecins indépendemment de la diversité de leur pratique.

En termes d'offre de soins, la réforme affecte principalement le niveau d'effort consentit par les médecins. Le léger accroissement des heures de travail clinique (+1.1%) s'accompagne en effet d'une baisse beaucoup plus importante du nombre d'actes fournis (−5.2%), facturables comme non facturables. Les médecins consacrent en conséquence un temps plus important à la réalisation de chacun des actes (+8.8%). Les heures de travail non-cliniques sont, quant à elles, peu affectées par le passage à la rémunération mixte. L'augmentation simultanée des heures de travail clinique et non-clinique engendre cependant un accroissement non négligeable (+3%) des heures passées au travail. Enfin, les semaines de travail apparaissent assez insensibles au mode de rémunération conformément au résultat obtenu par les études classiques consacrées au sujet (Sloan, 1975). Dans l'ensemble, ces résultats ne satisfont que partiellement les objectifs poursuivis lors de l'introduction de la rémunération mixte. Si le temps consacré aux actes en constitue une mesure adéquate, l'objectif de promotion de la qualité des soins semble être atteint. La rémunération mixte ne provoque, en revanche, qu'une diversification très modérée des activités de pratique. Au regard de l'important effet de sélection remarqué plus haut sur cette variable (Tableau 2.8), la rémunération mixte semble donc avoir pour seul effet de rémunérer les médecins qui manifestent une préférence marquée pour les activités non-cliniques.

Tableau 2.13 – Variations induites par une rémunération mixte obligatoire

	RA	RM obligatoire	Variation
Heures hebdomadaires totales	54.29	48.50	-11.9 %
——— cliniques (h^c)	47.21	41.71	-13.2 %
——— non cliniques (h^{nc})	7.08	6.78	-4.3 %
Semaines (W)	45.99	46.51	1.1 %
Actes[a] totaux	176.70	160.03	-10.4 %
——— facturables (AF)	152.33	143.66	-6.0 %
—— non facturables (ANF)	24.36	16.38	-48.8 %
Effort $\left(e = \frac{ANF + AF}{h^c * W}\right)$	81.38	82.48	1.3 %
Revenu annuel[a] (X)	161.92	190.72	15.1 %

[a]En milliers de Dollars. Les actes sont mesurés en Dollars constants (base 1996).

Note. Comportements de pratique moyens prédits par le modèle pour l'année 2002 selon que le schéma de rémunération correspond à la rémunération à l'acte (colonne de gauche) ou à un dispositif qui contraint l'ensemble des médecins à adopter la rémunération mixte (colonne centrale). La variation (dernière colonne) correspond au taux de variation relatif entre la première et la deuxième colonne.

D'une façon générale, ces variations – quantitativement faibles – se traduisent par une importante augmentation du coût du système de santé en termes de rémunération des médecins. L'introduction de la rémunération mixte s'avère donc un instrument puissant de rééquilibrage des rémunérations entre médecins, accompagnée d'une légère amélioration de la qualité. Elle apparaît ainsi comme un instrument plus efficace sur le plan des objectifs politiques que de l'efficience économique. Cette conclusion quant à l'efficacité de la combinaison d'instruments choisie doit cependant être nuancée par l'effet des modalités de mise en œuvre adoptées.

Afin de l'évaluer, le Tableau 2.13 présente les changements qui auraient prévalu si la rémunération mixte avait été rendue obligatoire pour l'ensemble des chirurgiens. La plupart des résultats précédents sont alors renversés, selon une ampleur démultipliée. A l'exception du revenu, qui continue à augmenter quoique beaucoup plus légèrement,

l'offre de soins est en effet réduite dans toutes ses dimensions par cette version de la réforme. Une très importante diminution des heures de travail clinique, associée à une baisse substantielle mais moins marquée du nombre d'actes délivrés, conduit ainsi à une diminution du temps consacré aux patients (accroissement de l'effort). La baisse des heures de travail non clinique vient par ailleurs renforcer celle du temps de travail clinique, qui résultent en une baisse importante du temps de travail des médecins.

Bien que les instruments utilisés semblent quantitativement inappropriés à promouvoir l'efficience de l'offre de soins, la rémunération mixte tire donc un parti important de la liberté d'adoption laissée aux médecins.

2.5 Conclusion

L'étude des effets attendus et observés de l'introduction de la rémunération mixte sur les choix de pratique des médecins a permis, dans ce chapitre, d'approfondir l'analyse quant à l'influence des rémunérations sur l'arbitrage qualité/quantité de l'offre de soins. Combinant une rémunération fixe et un taux de rémunération des actes réduit, la rémunération mixte est explicitement destinée à encourager la diversification des activités médicales des médecins et l'amélioration de la qualité des soins fournis aux patients. Ce changement dans les incitations s'accompagne potentiellement d'un effet de sélection, puisque les médecins du Québec peuvent, après l'introduction de la réforme en 1999, conserver l'ancien système de rémunération.

L'analyse théorique de l'effet attendu de son adoption a mis en évidence les propriétés des préférences des médecins suffisantes à ce que la réforme parvienne à atteindre ses objectifs. Sous les conditions traditionnellement retenues quant au goût pour le loisir (bien normal) comme pour la consommation (utilité marginale positive), la réaction des médecins à la réforme dépend uniquement de la sensibilité des choix de pratique aux

variations de prix. En raison de l'arbitrage entre marge intensive (temps consacré aux patients) et marges extensives (temps de travail et nombre d'actes) cet effet est plus ambigu que ne le laissent présager les travaux existants. Certaines configurations des préférences (valeurs relatives des élasticités croisées) assurent cependant que le passage à la rémunération mixte aboutisse à un double accroissement des heures de travail consacrées à l'enseignement ou à l'administration des établissements, et du temps consacré à chaque patient. Ces effets positifs en termes de diversification et d'amélioration de la qualité peuvent s'accompagner d'une réduction du temps total de travail des médecins.

L'analyse économétrique du comportement des médecins, observé de 1996 à 1998 et en 2002, permet de lever ces indéterminations pour les médecins appartenant à l'échantillon retenu. Les préférences des médecins sont estimées grâce à une base de données originale, combinant des données d'enquête sur le temps consacré au travail par les médecins et des données administratives sur les quantités d'actes délivrés, et leur valorisation monétaire. Les résultats permettent de simuler le comportement optimal des médecins sous divers modes de rémunération. En comparaison des choix qu'aurait engendré le maintien de la seule rémunération à l'acte, la rémunération mixte a eu pour effet d'augmenter légèrement l'ensemble des heures de travail et, surtout, d'accroître de façon substantielle le temps consacré à chaque acte médical. L'effet le plus important reste cependant un accroissement considérable du revenu versé aux médecins, suite à la prise en compte d'activités qui étaient jusqu'alors exclues des rémunérations. En ce sens, la rémunération mixte s'avère être une réforme coûteuse dont les effets sur la santé sont modestes, mais permettant de promouvoir l'équité des rémunérations entre les médecins.

Les instruments choisis par les autorités en charge de la politique de santé au Québec (montant de la rémunération fixe et taux de réduction de la rémunération des actes) conduisent donc à des résultats très mitigés au regard de l'efficience économique. La rémunération mixte tire cependant un bénéfice considérable du caractère volontaire de son adoption. Si l'instauration obligatoire du nouveau mode de rémunération avait été

préférée au dispositif en vigueur, en effet, la réforme aurait eu pour effet d'abaisser l'offre de soins dans toutes ses dimensions, du nombre d'heures de travail au temps consacré à chaque acte, et aurait simultanément accru le revenu versé aux praticiens. S'ils conduisent à douter de l'efficacité des niveaux de rémunération choisis, nos résultats abondent par conséquent dans le sens des revendications de plus en plus fréquentes en faveur d'une plus grande liberté dans les choix de rémunération.[42]

Ces conclusions proviennent d'une première estimation, utilisant le sous-ensemble des médecins chirurgiens. Bien que cette population constitue à de nombreux égards (diversité de la pratique, variété des choix de rémunération) un échantillon représentatif de la population des médecins, elles ne sauraient donc être définitives sans que soit évalué l'effet de la rémunération mixte sur l'ensemble des spécialités de pratique. Il faut noter, en particulier, que l'offre de pratique des chirurgiens est en partie rationnée par la disponibilité des équipements dans les établissements hospitaliers. Les médecins de cette spécialité peuvent n'être en conséquence que partiellement maîtres du nombre d'actes réalisés. Pour dépasser cette limite potentielle, il s'agirait alors d'estimer les préférences de tous les médecins en activité afin de simuler la réponse optimale des choix de pratique à la réforme dans l'ensemble du Québec.

Si elles sont confirmées, ces conclusions disqualifient moins la combinaison d'instruments incluse dans le dispositif de rémunération mixte que les niveaux de rémunération choisis pour son application. Comme l'a montré l'analyse théorique, en effet, l'efficacité de la réforme dépend assez largement des propriétés locales des préférences des médecins, et une plus grande efficacité du dispositif n'est pas à exclure *a priori*. L'estimation

[42] «*Il est urgent de compléter le paiement à l'acte par d'autres éléments de rémunération, selon le type d'exercice ou les efforts effectués en termes de qualité des soins. De plus en plus de médecins, notamment parmi les plus jeunes, sont prêts à une telle évolution. Pourquoi ne pas leur offrir ce choix, tout en permettant à ceux qui le souhaitent de continuer à être payés uniquement à l'acte ?*»"Sécu : la solitude de l'assuré" P-Y. Geoffard, *Libération* (3 octobre 2005). Les articles, cités plus haut, de Encinosa, Gaynor & Rebitzer (1997) et Barro & Beaulieu, (2003) mettent en évidence l'efficacité de ce type d'auto-sélection.

d'un modèle structurel permet de simuler les choix de pratique qui résultent de tout système de rémunération. Un prolongement naturel de notre analyse consisterait par conséquent à chercher la combinaison optimale entre rémunération fixe et taux de rémunération des actes, c'est à dire les niveaux de rémunération (fixe et variable) tels que l'amélioration des soins soit maximale pour un coût minimum. Une seconde lacune, difficile à combler, tient à ce que notre analyse néglige la décision de groupe qui préside à l'adoption de la rémunération mixte. Si cet aspect pourrait être intégré à l'analyse théorique, les clauses de confidentialité nous interdisent en revanche d'accéder aux informations quant au groupe d'appartenance des médecins. Une investigation empirique s'avère de ce fait difficilement envisageable.

Enfin, notre analyse laisse de côté deux aspects importants de l'effet des incitations sur les comportements de pratique, que les données dont nous disposons pourraient permettre d'appréhender. D'une part, il semble établi que les variations dans les taux de rémunération des actes sont susceptibles de donner lieu à un phénomène de demande induite. Cet effet a déja, au Québec, été observé dans le passé (Rochaix, 1993) et pourrait modifier de façon importante l'analyse coûts–bénéfices de l'effet de la réforme. Outre les variables de revenu, les données administratives fournies par la *Régie de l'Assurance Maladie du Québec* contiennent également des informations détaillées sur le nombre de patients traités et le nombre de visites réalisées par chaque médecin. L'effet de l'introduction de la rémunération mixte sur ces mesures de la demande de soins pourrait donc fournir une évaluation de l'ampleur de la demande induite engendrée par la réforme. Outre l'introduction de la rémunération mixte, la période couverte par nos données contient également, d'autre part, un certain nombre d'ajustements (montants et modalités d'application) dans les mesures de plafonnement qui s'appliquent au revenu des médecins. Compte tenu de l'originalité de cette mesure, connaître son effet sur les choix de pratique participerait à approfondir la compréhension du rôle des incitations dans la politique de santé. La mise en œuvre du dispositif s'appuie sur une distinction des activités médicales en fonction, notamment, de l'établissement de pratique. Un traitement adéquat de cet aspect nécessiterait une discrétisation spécifique des variables de pratique, et doit donc faire l'objet de nouvelles investigations.

Annexes

2.A Programme Ox

Le programme, écrit en langage Ox, a été adapté à partir du programme GAUSS développé par Train.[43] La structure du programme et les noms de variables ont été autant que possible conservées. Nous décrivons dans un premier temps les éléments spécifiques à notre version. Les matrices requises ont été crées à l'aide du logiciel STATA. Les commandes qui contrôlent la parallélisation requièrent le package OxMPI (Doornik, Shephard & Hendry, 2004).

2.A.1 Lexique des variables et matrices utilisées

```
 //* Code Files *//


par.ox
--------
Contains all usefull controls for the current estimation.
debug.ox
--------
Contains parameters controling extensive printing in debug mode
logit.ox
--------
Main code for one processor running
logitMPI.ox
--------
Main code for parallelized running


 //* Input files *//


xb.mat
--------
File containing NOBS rows padded with individual caracteristics values (0 or 1 for sex)
```

[43] Le programme est librement disponible sur la page personnelle de l'auteur, sous le titre "Mixed Logit Estimation for Panel Data using Maximum Simulated Likelihood".

```
for each variable
you want to include in the estimation.
cons.mat
--------
File containing the constant variables values in each alternative.
Must contain NALT*sumc(CONS) elements.
yvec.mat
--------
Contains the id of chosen alternative for each observation.
RM.mat
--------
Contains choice uncertainty for each observation. 0 if choice is certain. Otherwise, contains
the number of
alternatives subsequent to the one designated in yvec.mat between which the observation
may have chosen.
rescale.mat
--------
Contains 2 columns : the first indicates the id of the variable (1 for the first, 2 for
the second, ...) to
be rescaled the second the rescale factor. rows(rescale)n can be lower than NVAR.
xmat.mat
--------
Contains the values of non-constant variables in each alternative for each observation.
times.mat
--------
Identifies the number of times for which each of the NP agents chose an alternatives-eg.
the number of choices
each consumer made. For example, if the first agent faced 3 choice situations (e.g., made
a choice in each of three time periods)
and the second agent faced 7 choice situations, then TIMES is 3, 7, .... The sum of TIMES
over its NP elements
must equal the number of observations NOBS.


 //* par.ox *//


PREDICT
--------
Controls prevision behavior in the last likelihood calculation

 0 = Never compute predicted probability for the sample
 1 = Compute predicted probability for each individual in the sample at estimated parameters
 2 = Predict only mode : only compute predicted probabilities at starting values
NPRED
--------
The number of variables involved in prediction for PREDICT!=0. The matrix of predicted
values for each observation
witll contain NPRED columns, each (c) resulting from alternative j estimated probability
product with variable c.
Variables must be sorted according to NPRED order in Xmat.

RESCALE_START
```

```
--------
Controls starting values rescaling : starting values will be rescaled according to variables
rescale factor if 1.
Usefull while restarting an interrupted estimation.
CENSOR
--------
XXX
CONS
--------
1*NVAR vector identifying parameters associated with constant variable. A variable is
constant if its value
is the same for each observation. Constant variable matrix thus contains only one row,
used for every observation.
idXB
--------
Vector containing one row per individual caracteristics variable. If the variable is of
dummy type, put the
number of single values for which you a want a parameter estimate (e.g. 1 for sexe).
sumc(iDXB) must be the number of columns in xb.mat
iDB
--------
NVAR*1 vector associating parameters with individual caracteristics variables.
No interaction between parameter and individual variables when 0. If >0, put the id of
variable in iDXB.
As an example, if you want the second and fourth parameters among seven to be estimated
in interaction with sexe :
iDXB=<1>, iDB<0 ; 1 ; 0 ; 1 ; 0 ; 0 ; 0> ;
START
--------
Controls starting values : if string, starting values are loaded from "start.mat". The
file must contains
NFC+NN*2 values. Otherwise, put every numerical value, which will be used as default starting
for all parameters.
TITLE
--------
String containing the title for current estimation
PRINT
--------
Controls printing during iteration. Intermediate results will appear every PRINT iteration.
HALT
--------
Controls halton sequence. Loaded from file HMNAME if 1, created if 0.
DRAWS
--------
Controls simulation draws
 1 = SIMPLE
 2 = HALTON

QUICK
--------
Controls the gradient computation method used for optimisation.
 0 = Numerical Derivatives (Robust but time consumming)
 1 = Analytical Gradient (Quicker)
GRAD
```

```
--------
XXX
GRADOBS
--------
XXX


 //* MPI management *//


tag
--------
tag is the variable used by the master for sending tasks to nodes
 10 = Compute likelihood
 20 = Likelihood computation for probabilities prediction
 30 = Numerical derivatives for hessian calculation
seek
--------
One row per sequentially loaded file. Each contains the adress in the corresponding file
where loading has
previously stopped.


 //* debug.ox *//


DEBUG
--------
Controls extensive printing. Main calculation steps printed if 1.

debug_ll
--------
Commands a session without optimisation. Likelihood computation will be performed only
once.
debug_err
--------
If DEBUG, the matrix of random draws will be printed once.
debug_v
--------
If DEBUG, informations about the matrix containing utility based on fixed coefficients
will be printed once.
debug_ev
--------
If DEBUG, informations about the matrix containing utility will be printed once.
debug_expev
--------
If DEBUG, informations about the matrix containing exp(utility) will be printed once.
debug_denom
--------
If DEBUG, the denominator in individual contribution to likelihood will be printed once.
debug_p00
--------
If DEBUG, the probability of chosen alternative will be printed once.
```

```
debug_y
--------
If DEBUG, the vector of NALT dummies for choice will be printed once.
debug_p1
--------
If DEBUG, the sum over alterantive probabilities will be printed once (should be 1).
debug_mpi
--------
Nodes and master prints communiactions steps.
debug_halt
--------
Number of drawn Halton sequences to print if DEBUG.


 //* logit.ox : Program specific variables *//
IDXB
--------
Describe xb.mat : number of variable for each dummy in first column, position of
the first in the second.
IDB
--------
Describe parameters : individual caracteristic to be intercted with in first column,
number of estimated coeficients in the second.
```

2.A.2 Programme

```
#include <oxstd.h>
#import <maximize>
#include <par.ox>
#include <debug.ox>
static decl   XMAT, XCONS, YVEC, TIMES, YPERM, XB, RM, RSCLMAT, IDXB, IDB, MATCENSOR;
const decl   IDA = IDNC, HMNAME = "hm15.asc";
decl      id, numproc, procname, SIM=0, HM, tag;
decl      IDCONS, IDSPEC, SPEC, NEVAR, N, MyN, MyObs;
       start();
       data();
       rescale(rsclmat);
       rescB(b, const rsclmat);
       logitll( b, ll, score, hess);
       halton();
       haltonserial(n, s);
       cdfinvn(p);
       compll(b, ll, score, hess);
       score(b);
       ll(b, ll, score, hess);
       ScoreContributions(const func, vP, const avScore);
#include <single.ox>
//#include <mpi.ox>
main()
  {if (SINGLE)
     single();
```

```
26    else
27      mpi();
28    decl BETA;
29
30    OxMPIBarrier();        // Nodes synchronization for sequential printing
31    if (id==0)             // Master Prints session title
32      {println("");
33      println(date(), " ", time());
34      println(TITLE);
35      println(NP, " Individuals, ", NOBS, " observations, ", NALT, " Alternatives.");
36      println("");}
37
38    OxMPIBarrier();        // Nodes synchronization before names printing
39
40                    // Will contain people allocated to nodes
41    decl nodesNP = zeros(numproc,1);
42
43    if (id==0)             // Allocates observations between nodes
44                    // (fixing for non cylindered panel data)
45      {decl   allocate = cumulate(reshape(loadmat("times.mat",1),NP,1)),
46           i = 1, node = 1, aimed = floor(NOBS/(numproc-1)), done = 0;
47      while (i <= NP-1)
48           {if (aimed < allocate[i][0])
49                    // Number of raws read allocated to node
50           {nodesNP[node][0] = i - sumc(nodesNP[0 : node-1][0]);
51           done = allocate[i-1][0];
52           allocate = allocate - done;
53           node = node + 1;}
54         i = i + 1;}
55
56                    // Master in charge with the residual
57    nodesNP[0][0] = NP - sumc(nodesNP);}
58
59    OxMPIBcast(&nodesNP,0);    // Sends people vector to nodes
60    MyN = nodesNP[id][0];
61    println("Process ", id+1, " of ", numproc," on ", procname, ", MyN = ", MyN);
62    OxMPIBarrier();
63    data();                // Data loading
64
65    tag = 10 + QUICK;
66    if(id == 0)
67      {BETA = start();      // Master loads and rescales starting values
68      if (RESCALESTART*RESCALE)
69        {rescB(&BETA, RSCLMAT);
70        print("Rescaled ");}
71      print("Starting values : ", BETA);
72      decl lik;
73      if (PREDICT!=2)        // No computation in predict only mode
74
75          //* One likelihood calculation for running time indication *//
76
77        {println("");
78        println("Starting calculations...");
79                    // Calculations starting time
```

```
80          decl timeconv0 = OxMPIWtime(), time0 = time();
81          compll(BETA, &lik, 0, 0);
82          decl timeconv1 = OxMPIWtime();
83          println("");      // end time
84          println("Calculation started at ", time0, " finished at ", time());
85          print("First calculated likelihood : ", lik, " in ", timeconv1-timeconv0, "s");
86
87            //* Optimization *//
88
89          if (debugll==0)
90            {decl std;
91            MaxControl(NITER,1);
92            println("");
93            print("Starting optimization");
94            if (QUICK)
95              print(" using analytical gradient");
96            time0 = time();
97            timeconv0 = OxMPIWtime();
98            decl optim = MaxBFGS(compll, &BETA, &lik, 0, 1-QUICK);
99            if (optim == 4 )
100             println("Estimation failed, you should try a DEBUG session");
101           timeconv1 = OxMPIWtime();
102           println("Estimation started at ", time0, " finished at ", time());
103           println(MaxConvergenceMsg(optim), " obtained in ", timeconv1-timeconv0 , "s"
104 );
105           //* Estimation results *//
106
107           println("");
108           println("Preparing results...");
109
110           tag = 30;      // First derivatives calculation
111           decl H = score(BETA);
112           if (ROBUST)      // Numerical 2nd derivatives for robust variances
113             Num2Derivative(compll, BETA, &std);
114           else
115             std = H;
116
117           decl sigma = sqrt((diagonal(invertgen(std, 3)*H*invertgen(std,3)))');
118           decl t = BETA./sigma;
119           decl PrintB = BETA;
120
121           if (RESCALE)   // Descale before printing and saving
122             {decl descl = ones(rows(RSCLMAT), 2);
123                    // Create descaling matrix
124             descl[][0] = RSCLMAT[][0];
125             descl[][1] = 1 ./ RSCLMAT[][1];
126                    // Descale parameter and standard errors values
127             rescB(&PrintB, descl);
128             rescB(&sigma, descl);}
129
130           savemat("sigma.mat", sigma, 1);
131           savemat("beta.mat", PrintB, 1);
132
133                    // Results printing
```

```
134           decl printer = PrintB   sigma   t;
135
136           println(" Parameter Estimated Standard");
137           println(" value Error t");
138           println(" -------------------------");
139           println("Fixed Coefficients :");
140
141           decl k = 0, estimated = < >, pr = 0, pr1 = -1;
142           while (k <= NFC-1)
143             {pr1 = pr + IDB[k][columns(IDB)-1]-1;
144             print(constant(IDFC[k],IDB[k][columns(IDB)-1],1)   printer[pr : pr1][]);
145             estimated = estimated | constant(IDFC[k],IDB[k][columns(IDB)-1],1);
146             pr = pr + IDB[k][columns(IDB)-1];
147             k = k+ 1;}
148           if (NNC > 0)
149             {println("Normally distributed coefficients : ");
150             println(shape(IDNC | IDNC, 2*NNC,1)   printer[pr1 + 1 : ][]);
151             estimated = estimated | shape(IDNC | IDNC, 2*NNC,1);}
152           if (ROBUST)
153             println("Uses robust standard errors.");
154           else
155             println("Uses non robust standard errors. WARNING not reliable for random
156 coefficients.");
157           print("Final loglikelihood : ", lik);
158           savemat("results.mat", estimated   PrintB   sigma   t,1);}
159         }
160       if (PREDICT!=0)        // Predicted probabilities at BETA while PREDICT=1 or 2
161         {if (GRAD)
162           {tag = 50 + GROBS;
163           decl G = score(BETA);
164           println("");
165           savemat("gradient.mat", G, 1);}
166         SIM = 1;
167         println("");
168         println("Preparing forecasts...");
169         tag = 20;         // Send job to nodes
170         OxMPIBcast(&tag,0);
171                     // Mean over observations
172         if (RESCALE)      // Descale before forecasts
173           {decl DSCLMAT = RSCLMAT[][];
174           DSCLMAT[][1] = 1 ./ DSCLMAT[][1];
175           rescale(DSCLMAT);
176           rescB(&BETA, DSCLMAT);}
177         decl FC;
178         compll(BETA, &FC, 0, 1-QUICK);
179         FC = FC ./NOBS;
180                     // Print prediction results
181         decl printer = cumulate(ones(NPRED,1))   FC   (FC[][0]-FC[][1])./FC[][1];
182         println("");
183         println(" Variable Predicted Actual Rel. Error");
184         println("----------------------------");
185         println(printer);
186         savemat("forecast.mat",printer,1);}
187
```

```
188       tag = 99;           // Announce the end to nodes
189       OxMPIBcast(&tag,0);
190       println("End");}
191     else
192       {while(tag!=99)        // Nodes loops on calculations
193         {if (debugmpi)
194           println("Process ", id+1, " waiting for tag...");
195         OxMPIBcast(&tag,0);// Stop looping when tag==99
196         if (debugmpi)
197           println("Process ", id+1, " received tag : ", tag);
198         if (tag!=99)
199           {decl LL;      // Parameters in current iteration
200           OxMPIBcast(&BETA,0);
201           if (debugmpi)
202             println("Process ", id+1, " last received beta : ", BETA[NVAR-1]);
203           if (tag==20)   // Likelihood calculated for prediction
204             SIM=1;
205
206           if (tag!=30)    // Likelihood for My Obs at received BETA
207             ll(BETA, &LL, 0, QUICK);
208           if (tag==30)   // Numerical 1st derivative
209             score(BETA);
210           }
211         else           // Computations stops when tag=99
212           {if (debugmpi)
213               println("Process ", id+1, " about to exit...");}
214         }
215       }
216     if (debugmpi)
217       println("Process ", id+1, " has finished jobs");
218
219     OxMPIFinalize();        // Break communications
220   }
221           //* Data loading *//
222   data()
223     {NEVAR = NFC + 2*NNC;      // Number of estimated parameters
224   //    NVAR = columns(IDFC)+columns(IDNC);
225
226     SPEC = (CONS .== 0);      // Dummies identifying specific and constant variables
227     decl VARS = SPEC*ones(NVAR,1), VARC = CONS*ones(NVAR,1);
228     if (id == 0 && DRAWS == 2)    // Master generates halton sequences
229       HM = halton();
230   //* Nodes specific sequential loading *//
231     decl    seek = zeros(7,1),    // Opens files to be read
232         ftimes = fopen("times.mat", "r"), fxmat = fopen("xmat.mat", "r"),
233         fy = fopen("yvec.mat", "r"), frm = fopen("RM.mat", "r");
234
235     if (DRAWS == 2)
236       decl fhalt = fopen("temphalt.mat", "r");
237     if (sumc(iDB)!= 0)
238       decl fxb = fopen("xb.mat", "r");
239     if (CENSOR == 1)
240       decl fcensor = fopen("censor.mat", "r");
241     decl master = 1;        // For coherency with people allocation, loading
```

```
242    while (master >= 0)        // must start with nodes (master=1), finish with
243                       // master (master=0)
244      {for(decl i = 0; i <= (numproc-2)*master; i=i+1)
245        {if (id == i + master )
246          {if (MyN>0)      // Debug the case where master's MyN=0
247            {println("");
248            println("Loading process ", id+1, " datas...");
249
250                  // Load each individual's number of choices in process i
251            fseek(ftimes,"c", seek[0]);
252            fscan(ftimes, "
253            seek[0] = fseek(ftimes);
254            MyObs = sumc(TIMES[]);
255            print(" MyObs : ", MyObs);
256 //            println("TIMES ok");
257
258                  // Load specific variables values in process i
259            fseek(fxmat,"c", seek[1]);
260            fscan(fxmat, "
261            seek[1] = fseek(fxmat);
262 //            println("XMAT ok");
263
264                  // Load chosen alternative id in process i
265            fseek(fy,"c", seek[2]);
266            fscan(fy, "
267            seek[2] = fseek(fy);
268 //            println("YVEC ok");
269
270                  // Load RM in process i
271            fseek(frm,"c", seek[3]);
272            fscan(frm, "
273            seek[3] = fseek(frm);
274 //            println("RM ok");
275
276                  // Distribute generated halton sequences to nodes
277            if (DRAWS == 2)
278              {fseek(fhalt,"c", seek[4]);
279              fscan(fhalt,"
280              seek[4] = fseek(fhalt);
281 //               println("HM ok");
282              }
283                  // Load individual specific variables
284            if (sumc(iDB)!= 0)
285              {fseek(fxb,"c", seek[5]);
286              fscan(fxb,"
287              seek[5] = fseek(fxb);
288 //               println("XB ok");
289              }
290                  // Load censoring variable
291            if (CENSOR == 1)
292              {fseek(fcensor,"c", seek[6]);
293              fscan(fcensor,"
294              seek[6] = fseek(fcensor);
295 //               println("MATCENSOR is ", rows(MATCENSOR), " * ", columns(MATCENSOR));
```

```
296                 }
297               }
298             }
299                       // Seek of the loaded files sent to all procs
300         OxMPIBcast(&seek,i+master);}
301       master = master - 1;}    // Switch to master's loading
302     fclose(frm);
303     fclose(fy);
304     fclose(fxmat);
305     fclose(ftimes);
306     if (DRAWS==2)
307       fclose(fhalt);
308     if (sumc(iDB)!= 0)
309       fclose(fxb);
310   //* Common simultaneous loading *//
311
312     if (RESCALE)           // Load rescaling matrix
313       {RSCLMAT = loadmat("rescale.mat",1);
314       RSCLMAT = shape(RSCLMAT,2,NVAR)';}
315                       // Create dependent variable permutation matrix : 1 if alternative
316   is chosen; 0 otherwise
317     YPERM = zeros(MyObs,NALT);
318     decl i = 0;
319     while (i <= MyObs-1)
320       {YPERM[i][(YVEC[i]-1) : (YVEC[i]-1+RM[i])] = 1;
321       i = i + 1;}
322   //    println("iDB ", iDB);
323     IDB = iDB   ones(NVAR,1);             // Individual Caracteristics pointer
324     if (sumc(IDB[][0])!= 0)
325       {IDXB = idXB   zeros(rows(idXB),1);
326                       // Augments BETA dimensions with individual intercepts
327       decl i = 0;
328       while (i <= NVAR-1)
329         {if (IDB[i][0]!=0)
330           {decl col = 0;
331           while (col <= columns(IDB)-2)
332             {IDB[i][columns(IDB)-1] = IDB[i][columns(IDB)-1] + IDXB[IDB[i][col]-1][0];
333             // Number of b[i] estimated
334             NEVAR = NEVAR + IDXB[IDB[i][col]-1][0];
335             col = col + 1;}}
336         i = i + 1;}
337
338       i = 1;             // Variables adress in XB
339       while (i <= rows(IDXB)-1)
340         {IDXB[i][1] = IDXB[i-1][0] + IDXB[i-1][1];
341         i = i + 1;}}
342   //    println("IDXB :",IDXB);
343   //    println("IDB ",IDB);
344
345     if (MyN>0)              // Check RM treatment
346       {decl temp = sumc((sumc(YPERM'))')-sumc(RM)-MyObs;
347       if (VERBOSE)
348         {println("");
349         if (temp==0)
```

```
350           println("RM correctly treated in process ", id+1 ) ;
351         else
352           println("YPERM contains", temp , "more choices than what RM indicates in process
353 ", id+1 ) ;
354         }
355
356    IDCONS = IDSPEC = zeros(NVAR,1) ;
357    decl k = 0, idc = 1, ids = 1 ;
358    while (k <= NVAR-1)        // Identify the position of constant and specific variables
359      {if (CONS[k])
360        {IDCONS[k] = idc ;
361        idc = idc + 1 ;}
362      else if(SPEC[k])
363        {IDSPEC[k] = ids ;
364        ids = ids + 1 ;}
365      else
366        {if (VERBOSE)
367          println("Variable ", k+1, " is identified neither as constant nor as specific") ;}
368      k = k + 1 ;}
369    if (DEBUG)
370      {if (id==0)
371        println("IDCONS ", IDCONS, "IDSPEC ", IDSPEC) ;}
372                    // Load constant variables values
373    decl X1 = loadmat("cons1.mat", 1), X2 = loadmat("cons2.mat", 1) ;
374    XCONS = X1 | X2 ;
375    if (DEBUG && id==0)
376      {println("X1 contains : ", rows(X1)) ;
377      println("X2 contains : ", rows(X2)) ;
378      println("VARC : ", VARC, " NVAR : ", NVAR) ;}
379
380    if (VERBOSE && id==0)      // Master checks files
381                    // TIMES file
382        {decl temp=loadmat("times.mat",1) ;
383        temp = reshape(temp,NP,1) ;
384        println("") ;
385        if (sumc(temp)==NOBS)
386          println("TIMES file seems to be correct.") ;
387        else
388          println("Check TIMES file : does not fit NOBS.") ;
389                    // XCONS files
390        println("") ;
391        if (rows(XCONS) == VARC*NALT)
392          println("Constant variables correctly loaded.") ;
393        else
394          println("Constant variables file contains ", rows(XCONS), " elements, it should
395 have ", VARC*NALT) ;
396        println("") ;
397        if (sumc(iDB)!= 0)
398          println(columns(XB), " Individual caracteristics variables loaded.") ;
399        }
400      }
401                    // Reshape constant variables vector to its true dimensions   VARC
402 rows, NALT columns
403    XCONS = shape(XCONS, NALT, VARC)' ;
```

```
404   rescale(RSCLMAT);        // Rescale variables values using RSCLMAT
405 }
406         //* Starting values (master only) *//
407 start()
408   {decl b;
409   if (isstring(START))      // Starting values loaded if START="Y"
410     {b = loadmat(START,1);
411     b = reshape(b,NEVAR,1);
412     if (VERBOSE)
413       {println("");
414       println("Last estimation values loaded as starting");}
415     }
416   else               // START taken as default for all parameters otherwise
417     {b = ones(NEVAR,1)*START;
418     if (VERBOSE)
419       {println("");
420       println("Default value used as starting");}
421     }
422   return b;
423 }
424         //* Rescale datas using rsclmat *//
425 rescale(rsclmat)
426   {if (RESCALE)
427     {decl km;
428     if (id==0)          // Master prints the current operation
429       {if (VERBOSE)
430         {println("");
431         println("Rescaling data...");}
432       if (DEBUG)
433         {if (rsclmat[][0] > NVAR)
434           println("RSCLMAT identifies a variable that is not in the data set.");
435         println(" ");
436         println(" Variable Mult. Factor");
437         print(rsclmat[][0]   rsclmat[][1]);}
438       }
439
440     decl j = rows(rsclmat)-1;
441     decl i = 0;
442     while (i <= j)        // Rescale the variables
443       {if (CONS[rsclmat[i][0]-1])
444         {XCONS[IDCONS[rsclmat[i][0]-1]-1][] =
445           XCONS[IDCONS[rsclmat[i][0]-1]-1][] * rsclmat[i][1];}
446       else if (SPEC[rsclmat[i][0]-1])
447           {km = (IDSPEC[rsclmat[i][0]-1]-1)*NALT;
448           if (DEBUG)
449             println("km = ", km);
450           XMAT[][(km) :(km+NALT)-1] =
451             XMAT[][(km) :(km+NALT)-1] * rsclmat[i][1];}
452       i = i + 1;}
453     if (id==0)
454       println(" ...done");
455     }
456 }
457         //* Rescale parameters (master only) *//
```

```
458 rescB(b, const rescl)
459   {decl j = rows(rescl)-1, idb = 0, i = 0, colB = 0, maxB = columns(IDB)-1;
460   while (i <= NFC-1)        // Rescale fixed starting values
461     {decl ki = 0, l = 0;
462     while (l <= j)
463       {if (rescl[l][0] .== IDFC[i])
464         {decl nbB = 0;
465         while (nbB <= maxB-1)
466           {colB = IDB[IDFC[i]-1][nbB]-1;
467           if (0 <= colB)
468             ki = ki + IDXB[colB][0];    //
469           nbB = nbB + 1;}
470         (b[0])[idb : idb + ki ] =(b[0])[idb : idb + ki ] / rescl[l][1];
471         idb = idb + ki + 1;}
472       l = l + 1;}
473     i = i + 1;}
474
475   i = 0;                // Rescale normal starting values
476   while (i <= NNC-1)
477     {decl l = 0;
478     while (l <= j)
479       {if (rescl[l][0] .== IDNC[i])
480         {(b[0])[idb] =(b[0])[idb] / rescl[l][1];
481         (b[0])[idb + 1] =(b[0])[idb + 1] / (rescl[l][1]);
482         idb = idb + 2;}
483       l = l + 1;}
484     i = i + 1;}
485 }
486         //* Nodes coordination for likelihood comptutation (master only) *//
487
488 compll(b, LIK, score, hess)
489   {if (debugmpi)
490     decl timell = time();
491
492   if (debugmpi)
493     println("Process ", id+1, " sending tag : ", tag );
494
495   OxMPIBcast(&tag,0);      // Send tag to nodes
496   OxMPIBcast(&b,0);        // Send iteration parameters to nodes
497
498   if (debugmpi)
499     println("Process ", id+1, " last received b in compll : ", b[NVAR-1] );
500   decl LL;             // Iteration likelihood is saved in LL
501   LIK[0] = ll(b, &LL, score, hess);
502
503   if (debugmpi)
504     println("Calculated likelihood : ", LIK[0], " in : ", timespan(timell));
505
506   return 1;
507 }
508         //* LogLikelihood computation *//
509
510 ll(b, ll, score, hess)
511   {decl Mylik, Myscore = zeros(NEVAR,1);
```

```
512   if (debugmpi)
513     println("Process ", id+1, " last received b in ll : ", b[NVAR-1]);
514
515   if (MyN>0)             // Each process computes its observations loglikelihood if positive
516     {logitll(b, &Mylik, score, hess);
517     if (tag == 10 + QUICK || tag == 30)
518       Mylik = sumc(Mylik);
519     if (score)
520       Myscore = Myscore + sumc(score[0])';}
521
522   if (debugmpi)
523     println("Process ", id+1, " Mylik : ", Mylik );
524
525   if (score)
526     (score[0]) = OxMPIReduce(Myscore, MPISUM, 0);
527                    // LogLikelihoods summed over nodes and sent to master
528   return (ll[0]) = OxMPIReduce(Mylik, MPISUM, 0);
529 }
530           //* Score or Hessian computation *//
531 score(b)
532   {decl score;
533   OxMPIBcast(&tag,0);
534   OxMPIBcast(b,0);
535   if (MyN>0)             // Each node computes MyObs's hessian,
536     {if (ROBUST)         // stored in score
537       {ScoreContributions(logitll, b, &score);
538       if (tag!= 50+GROBS)
539         score = score'*score;}
540     else
541       {decl lik;
542       if (tag == 50+GROBS)
543         logitll(b, &lik, &score, 0);
544       else
545         logitll(b, &lik, 0, &score);}}
546   if (debugmpi)
547     println("Process ", id+1, " H first row : ", score[0][]);
548
549                    // Return sum over nodes hessians
550   return OxMPIReduce(score, MPISUM, 0);
551 }
552           //* Individual contributions to likelihood computation *//
553
554 logitll(b, LL, score, hess)
555   {decl X, err;
556   if (SIM)
557     decl PP=zeros(NPRED,2);      // For each predicted variable, contains the sum of
558 predicted levels
559   if (debugmpi)
560     println("Process ", id+1, " last received b in logitll : ", b[NVAR-1]);
561
562   decl time0 = time();
563   decl v = zeros(MyObs,NALT);
564   decl p0 = zeros(MyN,1);      // Simulated probability
565   decl der = zeros(MyN,NEVAR);// Jacobian matrix
```

```
566    decl derobs = zeros(MyObs,NEVAR);
567    decl p0obs = zeros(MyObs,1);
568    decl maxB = columns(IDB)-1, k = 0, idb = 0;
569  //* Adds variables with fixed coefficients *//
570
571    while (k <= NFC-1)
572    //* Constructs the relevant alternatives variables *//
573      {if (CONS[IDFC[k]-1])
574        X = XCONS[IDCONS[IDFC[k]-1]-1][].* ones(MyObs,NALT);
575      else if (SPEC[IDFC[k]-1])
576           {decl km = (IDSPEC[IDFC[k]-1]-1)*NALT;
577           X = XMAT[][(km) :(km+NALT-1)];}
578    //* Alternative constant *//
579      v = v + b[idb].*X[][ : ];    // Matrix (NOBS * NALT)
580    //* Interaction terms *//
581      idb = idb + 1;
582      decl nbB = 0;
583      while (nbB <= maxB-1)
584        {decl colB = IDB[IDFC[k]-1][nbB]-1;
585        if (0 <= colB)
586          {decl ki = 0;
587          while (ki <= IDXB[colB][0]-1)
588            {v = v + b[idb + ki].*XB[][IDXB[colB][1] + ki].*X[][ : ];
589            ki = ki + 1;}
590          idb = idb + ki;}
591        nbB = nbB + 1;}
592      k = k + 1;}
593
594    if (DEBUG*debugv)
595      {println("v has ", rows(v), " rows and ", columns(v), " columns" );
596      println("v first individual : ", v[0 :TIMES[0]-1][]);
597      println("ev first element : ", exp(v[0][0]));
598      debugv = 0;
599      }
600
601  //* Loop on individuals   : random coefficients *//
602    decl rd = 0, n = 0;
603    while (n <= MyN-1)
604
605    //* Loads random draws for simulation *//
606
607      {if (DRAWS == 1)        // Random draw
608        {decl NECOL = max(NVAR-NFC,1);
609        err = rann(NREP,NECOL);}
610
611      else if (DRAWS == 2)    // Halton sequence for individual n
612        {err = HM[(NREP*n) :(NREP*n+NREP-1)][];
613        if (DEBUG*debugerr)
614          println("HM = ", HM);}
615
616      if (DEBUG*debugerr*DRAWS)
617        {println("n = ", n, " err :", err);
618        debugerr = 0;}      // err has NREP rows and NECOL columns for each person
619
```

```
620      decl p00 = ones(NREP,1);// Simulated probabilities for individual n
621                      // Score vector for individual n, one row per simulation
622      decl g = zeros(NREP,NEVAR);
623  //      idb = idb + 1;
624  //      print("idb : ", idb);
625    //* loop over individual i observations *//
626      decl t=1, ev;
627      while (t<=TIMES[n])
628        {decl kmm = rd + t -1;
629      ev = v[kmm][];      // Contains utillity derived from fixed coefficients for each
630  alt (columns)
631      decl k = 0;
632      while (k <= NNC-1) // Adds variables with normal coefficients
633           {if (CONS[IDNC[k]-1])
634             X = XCONS[IDCONS[IDNC[k]-1]-1][];
635           else if (SPEC[IDNC[k]-1])
636             {decl km = (IDSPEC[IDNC[k]-1]-1)*NALT;
637             X = XMAT[kmm][(km) :(km+NALT-1)];}
638                      // Matrix (NREP * NALT)
639        ev = ev + (b[idb+(2*k)] + (b[idb+(2*k)+1] .* err[][k])) .*X[][ : ];
640          if (DEBUG*debugev)
641            {println("ev is ", rows(ev), " * ", columns(ev));
642            println("ev first raw : " , ev[0][]);
643            debugev = 0;
644            }
645          k = k+1;}
646                      // Individual i dummies for choice at period t
647        decl y = YPERM[kmm][];
648        if (DEBUG*debugy)
649          {println("y", y);
650          debugy = 0;}
651
652        ev = exp(ev);      // Probabilities based on exp(U)
653        if (DEBUG*debugexpev)
654          {println("exp(ev) has " , rows(ev), " r ", columns(ev), " c" );
655          debugexpev = 0;}
656                    // Probability denominator
657        decl denom;
658        if (CENSOR==1)
659          denom = (ev * MATCENSOR[kmm][]')';    // Matrix (1 * NREP)
660        else
661          denom = (sumr(ev))'; // Matrix (1 * NREP)
662        if (DEBUG*debugdenom)
663          {println("denom = " , denom);
664          debugdenom = 0;}
665
666                    // Chosen alternative probability
667        if (CENSOR == 1)
668          {p0obs[kmm] = meanc(((sumr(ev.*y.* MATCENSOR[kmm][]))'./denom)');
669          p00 = p00.*((sumr(ev.*y.* MATCENSOR[kmm][]))'./denom)';}
670        else
671          {p0obs[kmm] = meanc(((sumr(ev.*y))'./denom)');
672          p00 = p00.*((sumr(ev.*y))'./denom)';}
673        if (DEBUG*debugp00)
```

```
674           {println("p00", p00);
675           debugp00 = 0;}
676                     // Row vector of NALT choice probabilities
677         decl p1;
678         if (CENSOR == 1)
679           p1 = ((ev .* MATCENSOR[kmm][])./denom'); // Matrix (NREP * NALT), row sum should
680 be unitary
681         else
682           p1 = (ev./denom'); // Matrix (NREP * NALT), row sum should be unitary
683         if (DEBUG*debugp1)
684           {println("p1 :", p1);
685           println("sum : ", sumr(p1));
686           debugp1 = 0;}
687
688
689         if (SIM)        // In prediction mode : increments predicted levels with individual
690 n contribution
691           {decl k = 0, F = zeros(NPRED,NALT), p = meanc(p1);
692           while (k < NPRED)
693             {if (CONS[k])
694               F[k][] = XCONS[(IDCONS[k]-1)][];
695             else if (SPEC[k])
696               F[k][] = XMAT[kmm][((IDSPEC[k]-1)*NALT) : ((IDSPEC[k]-1)*NALT+NALT-1)];
697             k = k + 1;}
698           F = exp(F);
699           decl CORRECT = (F .!= 1);
700           F = F .* CORRECT;
701                     // Column vector containing the sum of predicted choices
702 //           if (CENSOR == 1)
703 //              PP[][0] = PP[][0] + F*p' .* MATCENSOR[kmm][];
704 //           else
705             PP[][0] = PP[][0] + F*p';}
706
707 //* Gradient specific calculations *//
708
709         if (hess || score || DEBUG)
710   //* Increment gradient with calculated probabilities *//
711           {decl idb = 0, k = 0;
712         //* Constructs the relevant alternatives variables *//
713           while (k <= NFC-1)
714             {if (CONS[IDFC[k]-1])
715               X = XCONS[IDCONS[IDFC[k]-1]-1][];
716             else if (SPEC[IDFC[k]-1])
717               {decl km = (IDSPEC[IDFC[k]-1]-1)*NALT;
718               X = XMAT[kmm][(km) :(km+NALT-1)];}
719
720                     // Alternative constant
721             g[][idb] = g[][idb] - sumc((p1.*X[][])')
722             + sumc(((ev.*X[][]).*y)')./(sumc((ev.*y)'));
723                     // Interaction terms
724             idb = idb + 1;
725             decl nbB = 0;
726             while (nbB <= maxB-1)
727               {decl colB = IDB[IDFC[k]-1][nbB]-1;
```

```
728                if (0 <= colB)
729                  {decl ki = 0;
730                  while (ki <= IDXB[colB][0]-1)
731                    {g[][idb] = g[][idb] - sumc((p1.*X[][]*XB[kmm][IDXB[colB][1] + ki])');
732                    g[][idb] = g[][idb]+ sumc((ev.*X[][]*XB[kmm][IDXB[colB][1] + ki].*y)')
733 ./(sumc((ev.*y)'));
734                    ki = ki + 1;}
735                  idb = idb + ki;}
736                nbB = nbB + 1;}
737              k = k+1;}
738
739           k = 0;        // Normally distributed coefficients
740           while (k <= NNC-1)
741             {if (CONS[IDNC[k]-1])
742               X = XCONS[IDCONS[IDNC[k]-1]-1][];
743             else if (SPEC[IDNC[k]-1])
744               {decl km = (IDSPEC[IDNC[k]-1]-1)*NALT;
745               X = XMAT[kmm][(km) :(km+NALT-1)];}
746             g[][NFC+(2*k)] = g[][NFC+(2*k)]    + sumc((X[][ : ].*y)');
747             g[][NFC+(2*k)+1] = g[][NFC+(2*k)+1] + sumc((err[][k].*X[][ : ].*y)');
748             k = k+1;}
749           derobs[kmm][] = meanc(p0obs[kmm].*g);
750           }
751         t = t+1;}
752
753       p0[n] = meanc(p00);      // Individual i contribution to likelihood
754                     // Increment hessian factor
755       der[n][] = der[n][] + meanc(p00.*g);
756       rd = rd + TIMES[n];      // Next individual address
757       n = n + 1;}
758
759
760     LL[0] = log(p0);        // Returned likelihood vector
761
762     if (SIM)
763       {decl i = 0;
764       YPERM[][]=0;
765       while (i<=MyObs-1)
766         {YPERM[i][(YVEC[i]-1) : (YVEC[i]-1)] = 1;
767         i = i + 1;}
768       for (decl k = 0; k < NPRED; k = k + 1)
769         {decl F;
770         if (CONS[k])      // Construct the vector of actual values
771           {F = XCONS[(IDCONS[k]-1)][];
772           F = exp(F);
773           decl CORRECT = (F .!= 1);
774           F = F .* CORRECT;
775           PP[k][1] = sumc((F*YPERM')');}
776         else if (SPEC[k])
777           {F = XMAT[][((IDSPEC[k]-1)*NALT) : ((IDSPEC[k]-1)*NALT+NALT-1)];
778           F = exp(F);
779           decl CORRECT = (F .!= 1);
780           F = F .* CORRECT;
781           PP[k][1] = sumc((diagonal(F*YPERM')'));}}
```

```
782                     // Sum of predicted and actual levels returned to ll
783       LL[0] = PP;
784       return 1;}
785
786  //* Hessian & Gradient : Final results *//
787    if (hess || score || DEBUG)
788      {decl det, H;
789      der = der ./p0;         // First derivative matrix
790
791      if (hess || DEBUG)      // Return hessian matrix (NVAR*NVAR) if required
792        {H = der'*der;
793        if (hess)
794          (hess[0]) = H;
795        det = determinant(H);
796        if (det == 0)      // Check Hessian determinant
797          println(id+1, " : Singular Hessian!");}
798      if (score)           // Return score matrix (NVAR*1) if required
799        {if (tag == 51)
800          (score[0]) = derobs./p0obs; //
801        else
802          (score[0]) = der;}
803
804      if (DEBUG)           // Print results for debugging
805        {println("Process : ", id+1);
806        decl sg = (sumc(der))';
807        print("Iteration LogLikelihood ", sumc(log(LL[0])));
808        println("First Derivatives matrix has ", rows(der), " rows and ", columns(der),
809  " columns");
810        println("First Derivatives matrix, diagonal : ", (diagonal(sg*sg'))');
811        decl invhess = invertgen(H, 3);
812        println("Hessian determinant : ", det);
813        println("Inverse Hessian first row : ", invhess[0][]);
814        println(" Iteration Standard ");
815        println(" Parameters Errors Gradient");
816        println("--------------------");
817        println(b   sqrt(diagonal(invhess))'   sg);
818        println(" ");
819        println("Likelihood calculated in ", timespan(time0), "ms");}
820      }
821    return 1;}
822        //* Halton sequences generation (master only) *//
823
824  halton()
825    {decl hm = < >;
826    if (HALT==1)           // Halton sequence loaded from HMNAME file
827    {hm = loadmat(HMNAME, 1);
828    println("Halton sequences loaded in process ", id+1, ".");}
829
830    else
831      {if (HALT==0)        // Halton sequence created
832      println("Creating Halton sequences in process ", id+1, " ....");
833      else
834        println("HALT must be 0 or 1. Default : Halton sequence created.");
835
```

```
836                     // Number of random estimated parameters (set to 1 if 0)
837       decl NECOL = max(NVAR-NFC,1);
838                     // Prime numbers vector
839       decl prim = < 2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, 61,
840  67, 71, 73, 79, 83, 89, 97, 101, 103, 107, 109, 113 >;
841       if (VERBOSE)
842         println("Halton sequences are based in primes : ", prim[0 :NECOL-1]);
843       print;
844
845       decl h = 1, hm1;      // Sequences generation
846       while (h <= NECOL)
847         {hm1 = haltonserial(10+NREP*NP, prim[h]);
848       if ((h <= NNC) || (h > (NNC+NUC+NTC)))
849           {hm1 = cdfinvn(hm1);
850                     // Truncation for inverse-normal extreme values
851       hm1 = hm1.*(hm1 .<= 10) + 10 .* (hm1 .> 10);
852     hm1 = hm1.*(hm1 .>= -10) -10 .* (hm1 .< -10);}
853
854         hm = hm   hm1[10 :rows(hm1)-1][];
855       h = h + 1;}
856       println("Finished Halton sequences.");
857
858       if (DEBUG*debughalt)
859         println("Halton sequences : ", hm[0 :debughalt][]);
860
861       if (SAVH)           // Sequences saved in specified file if required
862         {savemat(HMNAME, hm, 1);
863         println("Halton sequences saved.");}
864
865                     // Sequences saved in temporary file for nodes distribution
866       savemat("temphalt.mat", hm, 1);
867       }
868     return hm;
869  }
870         //* Halton serial numbers (master only) *//
871  haltonserial(n,s)          // Use the pattern described in Train, "Halton Sequences for
872  Mixed Logit." http ://elsa.berkeley.edu/wp/train0899.pdf
873     {decl j, y, x;
874                     // Create n+1 Halton numbers including the initial zero
875     decl k = floor(log(n + 1) ./ log(s));
876     decl phi = < 0 >, i = 1;
877     while (i <= k)
878       {x = phi;
879       j = 1;
880       while (j < s)
881         {y = phi + (j / s^i);
882       x = x | y;
883       j = j + 1;}
884       phi = x;
885       i = i + 1;}
886     x = phi;
887     j = 1;
888     while ((j < s) && (rows(x) < (n+1)))
889       {y = phi + (j / s^i);
```

```
890       x = x | y;
891       j = j + 1;}
892
893     phi = x[1 :n];          // Starting at the second element gets rid of the initial zero
894     return phi;
895   }
896         //* Inverse Normal function (master only) *//
897   cdfinvn(p)
898     {decl   p0 = -0.322232431088,      q0 = 0.0993484626060,
899         p1 = -1.0,             q1 = 0.588581570495,
900         p2 = -0.342242088547,      q2 = 0.531103462366,
901         p3 = -0.0204231210245,      q3 = 0.103537752850,
902         p4 = -0.453642210148*1e-4,   q4 = 0.38560700634*1e-2;
903     if ((p > 1.0) || (p < 0.0))
904       {println("Error : Probability is out of range.");
905     break;}
906                       // Create masks for p = 0 or p = 1
907     decl mask0 = (p .== 0), mask1 = (p .== 1),inf0 = mask0 .* (-1e+300),inf1 = mask1 .*
908   (1e+300);
909
910                       // Create masks for handling p > 0.5 and p >= 0.5
911     decl maskgt = (p .> 0.5), maskeq = (p!= 0.5);
912     decl sgn = (maskgt .== 0) * (-1) + maskgt;
913                       // Convert p > 0.5 to 1-p
914     decl pn = ( maskgt - p) .* sgn + mask1 + mask0;
915                       // Computation of function for p < 0.5
916     decl y=sqrt(sqrt((-2*log(pn)).^2));
917     decl norms = y + ((((y*p4+ p3).*y + p2).*y + p1).*y + p0)./
918   ((((y*q4 + q3).*y + q2) .*y + q1).*y + q0);
919
920                       // Convert results for p > 0.5 and p = 0.5
921     norms=((norms.*sgn).*maskeq).*(1-mask0)
922         .*(1-mask1)+mask0.*inf0+mask1.*inf1;
923
924     return norms;
925   }
926         //* Michael Creel's routine for numerical derivatives of a vector
927   (http ://ideas.repec.org/c/boc/bocode/x981001.html)*//
928   const decl SQRTEPS =1E-8;    // appr. square root of machine precision
929   const decl DIFFEPS1=5E-6;      // Rice's formula : log(DIFFEPS)=log(MACHEPS)/3
930   static dFiniteDiff1(const x)
931     {return max( (fabs(x) + SQRTEPS) * SQRTEPS, DIFFEPS1);
932   }
933   ScoreContributions(const func, vP, const avScore)
934     {decl i, cp = rows(vP), left, right, fknowf = FALSE, p, h, f, fm, fp, v;
935                       // get 1st derivative by central difference
936     for (i = 0; i < cp; i++)
937       {p = double(vP[i][0]);
938       h = dFiniteDiff1(p);
939       vP[i][0] = p + h;
940       right = func(vP, &fp, 0, 0);
941       if(i==0)
942         v = new matrix[rows(fp)][cp];
943       vP[i][0] = p - h;
```

```
944       left = func(vP, &fm, 0, 0);
945       vP[i][0] = p;    // restore original parameter
946       if (left && right)
947                       // take central difference
948         v[][i] = (fp - fm) / (2 * h);
949       else
950         return FALSE;}
951
952     avScore[0] = v;
953     return TRUE;
954   }
```

Chapitre 3

Demande de travail au noir en environnement concurrentiel : la malédiction de Bertrand[1]

«Mais s'il est si parfait que vous le déclarez,
Ce zèle qui vous pousse et dont vous vous parez,
D'où vient que pour paraître il s'avise d'attendre
Qu'à poursuivre sa femme il ait su vous surprendre,
Et que vous ne songez à l'aller dénoncer
Que lorsque son honneur l'oblige à vous chasser ?»

Tartuffe ou l'imposteur (Acte V, Scène VII), Molière (1664).

Le pouvoir de taxation fournit aux gouvernants les ressources financières nécessaires à l'exercice du monopole de la contrainte légitime qui fonde son existence. [2] Il est donc

[1] Ce chapitre est inspiré d'un travail en cours, réalisé en collaboration avec Jean-Louis Rullière, dans le cadre d'un contrat de recherche avec l'*Agence Centrale des Organismes de Sécurité Sociale*.

[2] «*[...] l'État moderne est un groupement de domination de caractère institutionnel qui a cherché*

indissociable des missions qui incombent à l'Etat. Ainsi, bien que la nature de ces missions partage les économistes, et détermine assez largement les contours des "écoles" du choix publique (Mueller, 2003), la nécessité de garantir la collecte des taxes demeure une exigence unanimement partagée. Pour les économistes, cette position de principe est renforcée par les coûts économiques associés au développement de l'économie souterraine, en termes de croissance (Loayza, 1997) comme de développement (Carillo & Pugno, 2004).

Du point de vue de l'analyse économique, la transition des agents économiques vers le secteur de l'économie informelle[3] apparaît comme une réaction naturelle à une lourdeur excessive des charges liées à l'activité sur le marché légal. Une abondante littérature est consacrée à en évaluer la nature et identifie quatre causes principales susceptibles de participer au développement de ce secteur (voir la synthèse proposée par Schneider & Enste (2000) pour plus de détails). Se détourner du marché légal est d'abord un moyen de fuir les coût directs de la légalité, inhérents au niveau des taxes elle-mêmes (Jung, Snow & Trandel, 1994 ; Trandel & Snow, 1999) et aux lourdeurs administratives (Friedman, Johnson, Kaufmann & *al.*, 2000). A ces coûts directs s'ajoutent les coûts d'opportunité imposés par le cadre légal. Le passage à l'économie informelle permet ainsi de se soustraire aux contraintes qui pèsent sur le marché du travail (limitation de la durée du temps de travail, des horaires, etc.) et de conserver le bénéfice des transferts sociaux dont le versement est conditionnel à des critères de revenu (Lemieux, Fortin & Frechette, 1994).

(avec succès) à monopoliser, dans les limites d'un territoire, la violence physique légitime comme moyen de domination et qui, dans ce but, a réuni dans les mains des dirigeants les moyens matériels de gestion.», M. Weber (1919, p.32), Le savant et le politique.

[3]Les termes d'économie *souterraine* (*underground economy*) ou *informelle* (*informal economy*) sont utilisés indifféremment dans la littérature pour désigner la production de biens légaux réalisée en dehors de la loi. Cette production se fait alors à partir de travail non déclaré, appelé ici travail *au noir* (*black*, ou *underground*, *work*). Nous nous conformons ici à cette tradition, dont Gërxhani (2004) propose une synthèse historique.

Dans ces conditions, le choix du secteur d'activité résulte alors d'un simple arbitrage coût/bénéfice. L'offre de travail au noir peut ainsi se comprendre comme le résultat d'un choix en deux étapes, où l'allocation du temps de travail entre secteurs s'ajoute à l'arbitrage consommation/loisir traditionnel (Cowell, 1981 ; 1985 ; 1990). De la même façon, la demande de travail qui émane des producteurs s'adresse à l'un ou l'autre des secteurs selon leur rentabilité respective (Rauch, 1991). Lorsqu'offre et demande sont prises en compte simultanément selon ces mécanismes, ces analyses permettent alors de mettre en évidence l'effet des variables de politique économique (taux de taxe, niveaux et conditions des transferts sociaux, services publics, etc) sur le développement de l'économie informelle et sa relation avec le secteur légal.[4] Ces variables constituent autant d'instruments sur lesquels il est possible d'agir pour contenir le développement de l'économie souterraine. Leur portée est cependant limitée par les objectifs propres à la politique économique poursuivie et l'instauration d'une politique de répression spécifiquement dédiée à lutter contre le travail au noir s'y ajoute en général. En réaction à un certain nombre de débats récents[5], l'objectif de ce chapitre est d'évaluer l'efficacité potentielle de l'addition d'un nouvel instrument à ce dispositif : la dénonciation du travail au noir par les firmes concurrentes.

Dans la tradition de Becker (1968), les instruments de la politique de répression (contrôle et amende en cas de détection) s'intègrent à l'arbitrage coût–bénéfice qui décrit le comportement de demande de travail au noir d'une firme isolée (voir Chapitre 1 pour une description plus détaillée). Une entreprise devrait ainsi recourir au travail au noir si le bénéfice attendu de la fraude est supérieur à l'amende espérée. Dans le cas de la demande de travail au noir, cependant, l'activité illégale est commise par des entreprises qui évoluent dans un environnement concurrentiel. Le bénéfice de l'illé-

[4] « *The theoretical literature on the underground economy in a general equilibrium setting concentrates either on the problems raised by a segmented labour market – the wage gap and unemployment especially – or on the effects of taxation and regulations and the provision of public services in determining the proportion of the underground economy.*»Carillo & Pugno (2004, p.258).

[5] C. Rollot "Se taire ou cafter ? La dénonciation des actes frauduleux fait débat en France", *Le Monde Economie* (11 Octobre 2005).

galité dépend alors non seulement du comportement du fraudeur mais également de la stratégie choisie par ses concurrents. Afin d'évaluer l'effet de la dénonciation sur le comportement de demande de travail au noir, l'analyse doit donc prendre en compte cette interaction. Par cet intermédiaire, ce chapitre propose une évaluation théorique et expérimentale des incitations à embaucher au noir qui ressortent de l'environnement économique dans lequel les firmes interagissent – caractéristiques du marché et de la politique de répression notamment.

A cet égard, le modèle de Bertrand (Bertrand, 1883) est connu pour cristalliser les mécanismes fondamentaux de la concurrence (Tirole, 1994, ch.5 p.21), jusqu'à conduire au *paradoxe de Bertrand* selon lequel la tarification au coût marginal résulte de toute concurrence en prix. Si l'interaction est répétée, ce résultat doit être nuancé et le modèle de Bertrand fournit une représentation plus réaliste de la concurrence. La perspective de profits futurs permet en effet, dans ce cas, de mettre en œuvre un accord de collusion tacite en maintenant le prix à un niveau supérieur au coût marginal. L'intensité de la concurrence reste cependant une variable centrale, puisque le fonctionnement du marché tend vers une situation de concurrence pure et parfaite à mesure que la taille du marché s'accroît (Friedman, 1971). Ce résultat théorique, devenu classique, est confirmé empiriquement. Considérant des marchés expérimentaux formés de deux à quatre firmes, Abbink & Brandts (2004) montrent, par exemple, que le prix d'équilibre est fortement décroissant par rapport au nombre de firmes présentes, même s'il n'atteint jamais le coût marginal. Dufwenberg & Gneezy (2000) rendent plus forte encore cette conclusion, en l'élargissant aux marchés où l'interaction n'est pas répétée.

Qu'elle soit instantanée ou répétée, la concurrence rend donc de plus en plus difficile l'obtention de profits positifs. A ce titre, le bénéfice de la fraude fiscale inhérent à la demande de travail au noir devrait donc être décroissant du nombre de concurrents présents sur le marché. Ainsi, même lorsque l'évasion est rentable au prix concurrentiel on peut s'attendre à ce que la concurrence élimine naturellement les possibilités de profit liées à la fraude, et décourage le recours au travail au noir. Nous montrons dans

un premier temps que cette conclusion ne résiste que partiellement à l'analyse. Si la guerre des prix tend effectivement à éliminer les profits issus de l'évasion, l'intensité de la concurrence oblige les firmes à la choisir. Lorsque la demande de travail au noir est prise en compte, le nouvel équilibre du marché correspond donc à un état d'évasion généralisée, mais qui n'améliore pas les profits réalisés par les firmes. L'interaction fondée sur le niveau du prix conduit alors à la *malédiction de Bertrand*, au sens où les firmes se trouvent dans l'impossibilité de sélectionner l'état du marché où l'embauche est légale alors même que l'évasion n'améliore pas leur situation.

Si l'appartenance au marché fournit aux firmes une information sur les décisions de leurs concurrentes, un second instrument s'ajoute cependant à l'interaction. Dans ce cas, la dénonciation de l'évasion fiscale pourrait en effet doter les firmes d'un pouvoir coercitif contre la généralisation de l'évasion, et ainsi mettre un terme à la malédiction de Bertrand. De fait, ce mécanisme est implicitement utilisé par les autorités en charge de la lutte contre le travail au noir en France, puisque les URSSAF[6] réalisent une part importante de leurs contrôles en réaction à une dénonciation par un concurrent.

L'analyse économique de la dénonciation s'est originellement consacrée à l'efficacité de la repentance. L'attrait de cet instrument tient en particulier à ce qu'il fait porter les coûts de détection sur les contrevenants (Kaplow & Shavell, 1994) et permet en conséquence de réduire les coûts de contrôle des activités illégales à dissuasion constante (Innes, 1999a ; 1999b). Encourager l'aveu des comportements illégaux est alors un choix optimal pour les autorités chargées de la répression (Feess & Heesen, 2002). La notion de dénonciation généralise cette approche au cas où le comportement illégal d'autrui est révélé par un fraudeur qui en a connaissance. En ce sens, elle fait appel à l'analyse des activités illégales en groupe, dans lesquelles les comportements illégaux impliquent

[6] *Union de Recouvrement des Cotisations de Sécurité Sociale et d'Allocations Familiales*. Organisées en antennes locales (103 au total), elles sont plus particulièrement chargées du recouvrement des cotisations sociales, et gèrent en conséquence l'immatriculation des professionnels, l'encaissement des cotisations et contributions, le contentieux et le contrôle. Cette dernière fonction est confiée à 1400 inspecteurs répartis sur l'ensemble du territoire national.

la participation de plusieurs agents économiques. La dénonciation est alors destinée à interrompre la coopération entre les fraudeurs (Møllgaard, 2002). Elle met les criminels dans une situation de dilemme du prisonnier et conserve dans ce cadre ses propriétés d'efficacité en termes de dissuasion à un coût minimum (Berentsen, Brügger & Lörtscher, 2004 ; Feess & Walzl, 2004).

En matière de régulation de la concurrence, ces principes se sont traduits par l'addition de clauses de clémence (*leniency clauses*) au dispositif réglementaire anti-concurrentiel de toutes les grandes zones commerciales (Europe et Etats-Unis en particulier, voir Feess & Walzl (2003), Motchenkova & Kort (2004) et Motchenkova (2004a) pour des analyses théoriques comparées de ces dispositifs). Le recours à cet instrument témoigne de la confiance des autorités dans l'efficacité de la dénonciation, puisqu'il consiste à offrir au fraudeur – partiellement ou totalement – une exonération des sanctions encourues s'il dénonce l'activité illégale à laquelle il participe.

L'efficacité théorique de cet encouragement à la dénonciation dans la lutte contre les accords anti-concurrentiels est pourtant ambigu. La première étude consacrée à cette question, Motta & Polo (2003), étudie l'effet des clauses de clémence lorsqu'elles sont limitées aux dénonciations qui suivent l'ouverture d'une enquête. Dans ce cas, les clauses de clémence réduisent la stabilité *ex post* des accords (i.e. après mise en oeuvre de la collusion), en accroissant l'incitation unilatérale à dévier. Elles renforcent en revanche l'incitation *ex ante* à s'y livrer en diminuant l'amende espérée en cas de contrôle. A ce premier effet pervers peut s'ajouter un renforcement de la stabilité des accords collusifs (*ex post*) lorsque les programmes de clémence sont étendus aux dénonciations qui précèdent l'ouverture d'une enquête. La dénonciation peut alors servir de stratégie de punition, et être ainsi utilisée comme une menace qui facilite la collusion (Spagnolo, 2003). Les programmes de clémence qui récompensent la dénonciation renversent cependant ce résultat (Aubert, Kovacic & Rey, 2005) et rendent plus difficile la coordination au sein du cartel (Brisset & Thomas, 2004). Dans le cas contraire, les clauses de clémence facilitent la collusion. Elles tendent alors à en accroître la durée (Motchenkova,

2004b), et peuvent même permettre leur mise en œuvre sur des marchés qui, en leur absence, auraient connu un fonctionnement concurrentiel (Spagnolo, 2002).

En dépit de leur intégration croissante dans les droits de la concurrence, peu de travaux empiriques ont été consacrés à l'effet des clauses de clémence sur les comportements anti-concurrentiels. Une première exception est l'expérience proposée par Hamaguchi & Kawagoe (2005), qui étudient la coordination des participants au sein de groupes de taille variable (2 ou 7) en présence de clauses de clémence destinées à rompre la collaboration. Conformément aux principes de la logique de l'action collective (Olson, 1978), il apparaît que l'augmentation de la taille du groupe rend instables les accords de collusion. L'absence de traitement exempt de clauses de clémence rend cependant difficile l'évaluation de leur effet à partir de ces résultats. Surtout, cette expérience laisse de côté les déterminants stratégiques liés à la concurrence, puisqu'elle utilise des fonctions de paiement indépendantes du nombre de firmes qui participent à l'accord de collusion. L'expérience réalisée par Apesteguia, Dufwenberg & Selten (2004) intègre explicitement cet aspect. Le protocole de l'expérience organise en effet une concurrence en prix, sans répétition, entre trois firmes dont la collusion est facilitée par une étape préalable de communication libre mais sans engagement crédible. L'effet des clauses de clémence sur la formation d'accords de collusion est en outre testé par quatre traitements qui diffèrent par le régime d'exonération accordé à la firme dénonciatrice. Le traitement de référence (appelé *Standard*) permet aux firmes de dénoncer la collusion mais n'offre aucune incitation financière en ce sens. Deux traitements récompensent toutes les firmes dénonciatrices, soit par une exonération de l'amende (*Leniency*) soit par une récompense à la dénonciation (*Bonus*). Enfin, l'effet de la communication est testé par une version du traitement *Standard* qui l'élimine (*Ideal*). La baisse significative du prix entre ces deux traitements témoigne du rôle moteur de la communication dans la formation d'accords de collusion.

Surtout, les comportements expérimentaux se démarquent assez nettement des résultats théoriques résumés plus haut. D'abord, l'introduction de clauses de clémence (*Le-*

niency VS *Standard*) tend à diminuer le nombre d'accords de collusion conclus comme leur stabilité (nombre d'accords dénoncés par l'une de ses parties prenantes). Ensuite, les clauses de clémence parviennent empiriquement à jouer le rôle qu'en attendent les régulateurs en durcissant la concurrence. Le prix d'équilibre dans les traitements avec clauses de clémence (*Leniency*, *Bonus*) est en effet significativement inférieur à celui qui émerge en leur absence (*Standard*, *Ideal*). Enfin, le versement de récompenses à la dénonciation (*Bonus*), plutôt que de simples exonérations (*Leniency*), laisse inchangé le prix de marché. Cette version des clauses de clémence, pourtant privilégiée par les analyses théoriques, tend même à être contre-productive *ex ante*. Encouragées par la perspective d'obtenir les gains de dénonciation, les firmes ont en effet tendance à conclure plus encore d'accords de collusion afin de pouvoir les révéler.

Outre les motivations individuelles à choisir l'illégalité, l'analyse de la demande de travail au noir doit également incorporer les stratégies de marché adoptées par les firmes, qui en déterminent le bénéfice. L'analyse de l'efficacité de la dénonciation, et d'éventuelles mesures d'encouragement tels que les programmes de clémence, pour lutter contre le travail au noir se trouvent par conséquent à l'intersection des travaux consacrés respectivement aux activités illégales en groupe et à la régulation de la concurrence. A cette fin, l'analyse proposée ici intègre simultanément les décisions individuelles d'embauche illégale et les stratégies concurrentielles des firmes. Loin de résoudre la malédiction de Bertrand, la dénonciation apparaît comme une menace crédible qui renforce l'incitation à recourir au travail au noir. Elle peut permettre, en effet, de dégager des profits positifs de l'évasion grâce à la mise en oeuvre d'un équilibre d'*évasion collusive*, dans lequel les firmes choisissent d'embaucher au noir mais parviennent à interrompre la guerre des prix. Comme le souligne Motta (2004, ch.4 p.138), la notion économique de collusion fait référence à un état du marché dans lequel les prix atteignent un niveau supérieur au prix de référence – correspondant en général au coût marginal – plutôt qu'aux moyens utilisés pour y arriver. De la même façon, l'analyse théorique met en évidence différentes stratégies capables de faire émerger l'équilibre d'*évasion collusive*.

La collusion tacite en est une première, largement analysée dans la littérature. En son absence, l'intensité de la concurrence conduit à la *malédiction de Bertrand*, en vertu de laquelle les firmes sont contraintes à une évasion qui, pourtant, n'améliore pas leur situation (Section 3.1). Lorsque les autorités s'appuient sur les dénonciations pour sanctionner le travail au noir, l'évasion collusive peut être mise en œuvre par une stratégie de *silence collusif*. La dénonciation, utilisée comme une menace crédible, élargit alors le champ des marchés capables de tirer des profits positifs de l'évasion fiscale (Section 3.2). Ces conditions d'émergence de l'évasion collusive – collusion tacite et silence collusif – sont testées par une expérience, intégrant explicitement la dénonciation (Section 3.3). Les comportements observés confirment la généralisation de l'évasion ainsi que l'élargissement des possibilités d'évasion collusive grâce à la dénonciation (Section 3.4). La perspective d'encourager la dénonciation ne semble donc pas de nature à pouvoir faire diminuer l'importance du travail au noir (Section 3.5).

3.1 Demande de travail au noir et concurrence à la Bertrand

Avant d'évaluer les effets possibles de la dénonciation sur la demande de travail au noir, il convient d'introduire explicitement l'environnement économique qui en détermine le bénéfice attendu. A cette fin, nous nous situons dans le cadre le plus propice à isoler les mécanismes issus de la concurrence en considérant une concurrence en prix. Si l'interaction entre les firmes est répétée selon un horizon temporel suffisamment long, le marché peut mettre en oeuvre un accord de collusion tacite générant des profits positifs. Dans le cas contraire, la malédiction de Bertrand entraîne le marché dans un état d'évasion généralisée qui détériore la situation des firmes.

3.1.1 Cadre d'analyse

Nous considérons un marché sur lequel n firmes identiques se livrent une concurrence en prix. Au sein de chaque firme, l'activité est déléguée à un agent dont l'effort, noté e, détermine le niveau de production selon la fonction : $q = f(e)$. La fonction de production inverse correspond donc au niveau d'effort nécessaire pour obtenir un niveau donné du produit : $e(q) = f^{-1}(q)$. La demande pour le bien homogène produite par les firmes dépend uniquement du prix, selon la relation $D = D(p)$.

Dans la relation qui lie les firmes à leurs employés, nous ignorons les problèmes liés à l'asymétrie d'information en supposant que l'effort est observable et vérifiable. Si les agents sont en outre supposés homogènes, l'effort fournit pour un niveau de salaire donné est le même pour tous les employés et dans toutes les firmes, égal à e_0. Au total, en notant W le coût unitaire de l'effort pour la firme, la fonction de coût est donc : $C(q) = e(q)\ W$. Le modèle de concurrence à la Bertrand requiert, entre autres hypothèses, que le coût marginal soit constant.[7] A cette fin, nous nous limitons au cas linéaire, dans lequel la fonction de production est $q = e$. La fonction de coût devient alors $C(q) = q\ W$, et le coût marginal correspond alors au coût unitaire de l'effort $C_m = W$.

Sous ces hypothèses, le paradoxe de Bertrand prédit que, quelle que soit la taille de l'industrie, l'équilibre non-coopératif de l'industrie correspond à l'équilibre concurrentiel $(p^c, Q^c/n)$, où $Q^c = D(p^c)$, tel que le profit s'annule : $\Pi^c = p^c\ \frac{Q^c}{n} - C\left(\frac{Q^c}{n}\right) = 0$. Afin de modéliser la demande de travail au noir de l'industrie à long terme, nous utilisons la version dynamique du jeu. Plus précisément, on suppose que le marché est susceptible de disparaître à chaque période selon la probabilité constante γ. Cette hypothèse est formellement équivalente à la version classique en horizon infini. La probabilité de destruction du marché peut en effet s'interpréter comme un taux d'escompte "objectif".

[7] Voir, par exemple, D'Aspremont, Dos Santos Ferreira, & Gérard-Varet (2003) pour une discussion exhaustive des hypothèses du modèle de base.

Si l'on note δ la préférence pour le présent des firmes, le modèle utilisé ici est en tout point équivalent à la formulation traditionnellement utilisée en substituant $1-\delta$ à γ. Cette version offre l'avantage d'un plus grand réalisme qui facilite le passage à l'analyse expérimentale.[8]

Dans ce cadre, le paradoxe de Bertrand est levé par la possibilité que les firmes mettent en oeuvre un accord de collusion tacite. Les conditions sous lesquelles cet accord collusif peut constituer un équilibre non-coopératif du marché dépendent de la stratégie de punition sur laquelle il s'appuie. Les deux stratégies de punition les plus largement utilisées dans la littérature sont la stratégie de cliquet (*trigger strategy*, Friedman, 1971) et la stratégie de carotte et bâton (*stick and carot strategy*, Hackner (1996) par exemple). Toutes deux consistent à maintenir le prix de collusion aussi longtemps que toutes les firmes présentes sur le marché s'y conforment. Leur différence essentielle tient à la durée de la punition. Tandis que la première consiste à punir toute déviation par le retour irrémédiable à l'équilibre de Nash du jeu, la seconde recourt en effet à une punition limitée dans le temps. Sous les hypothèses de base du modèle de Bertrand, cadre que nous adoptons, la stratégie de cliquet est la punition la plus sévère, et par conséquent

[8]Bien que cette formulation soit formellement équivalente à celle d'un jeu répété en horizon infini, l'équivalence ontologique de ces deux versions a suscité d'importants débats entre théoriciens de jeux. Selten, Mitzkewitz & Uhlich (1997) considèrent ainsi qu'il est impossible de reproduire en laboratoire les conditions qui correspondent à un horizon infini, pour la simple raison que le jeu ne saurait se répéter au-delà du temps imparti à l'expérience. Nous adoptons ici le point de vue défendu par A. Rubinstein, selon lequel les comportements observés correspondent au pendant empirique du modèle théorique en horizon infini tant que ces comportements sont indépendants de celui qui sera adopté à la dernière période : *[finite and infinite horizon] «models capture a very realistic feature of life, namely the fact that the existence of a prespecified finite period may crucially affect people's behavior (consider the last few months of a presidency or the fact that religions attempt to persuade their believers that there is "life after death")»* (Osborne & Rubinstein ; 1994, p.136). D'un point de vue empirique il semble en tout état de cause que les comportements de coopération – en début de jeu – soient assez peu sensibles à la règle de terminaison choisie. Dans le cadre d'un Dilemme du prisonnier répété, Normann & Wallace (2005) obtiennent en effet des profils de coopération semblables à l'effet de "fin de jeu" près, que la période finale soit connue des participants, qu'elle leur soit cachée ou qu'elle soit aléatoire.

la plus appropriée pour caractériser la stabilité des accords collusifs (Rey, 2003).[9] En s'appuyant sur cette stratégie de punition, un équilibre non-coopératif du marché $p > p_c$ peut assurer un niveau de profit quelconque $\Pi_m = \Pi(p) > 0$ si :

$$\sum_{t=0}^{\infty}(1-\gamma)^t.\frac{\Pi_m}{n} \geq \Pi_m + \sum_{t=1}^{\infty}(1-\gamma)^t\Pi^c \tag{3.1}$$

La collusion tacite est donc un équilibre sur les marchés dont la probabilité de destruction est suffisamment faible, telle que :

$$\gamma \leq \frac{1}{n} \equiv \gamma_c$$

Lorsque le travail au noir est introduit dans le modèle, la fraude est une source de profit supplémentaire qui s'ajoute à celles qu'offrent les possibilité de collusion. La demande de travail au noir est en effet prise en compte sous la forme d'un coût endogène. Avant de prendre sa décision de prix, chaque firme peut en effet décider de la légalité de l'emploi offert à l'employé. Si elle choisit un emploi légal, la firme doit s'acquitter des taxes, dont le taux est noté τ, en plus du salaire qui rémunère l'agent, w. Le coût marginal d'une firme qui choisit cette option est donc : $W = (1+\tau)\ w$. Si, à l'inverse, elle opte pour un emploi illégal (au noir) le coût marginal de la firme se réduit au salaire $W = w$. Dans ce modèle, le recourt au travail au noir introduit donc une source possible d'hétérogénéité du coût. Contrairement à la plupart des travaux consacrés à

[9]Une seconde différence tient à la sévérité de la punition. La punition étant limitée dans le temps, la stratégie de carotte et bâton peut en effet s'appuyer sur un code pénal qui fait tomber le prix de marché y compris en deçà du coût marginal. Abreu (1986) établit que cette stratégie est la punition optimale, au sens où elle permet de mettre en oeuvre les stratégies de collusion les plus rentables. Ces deux stratégies de punition doivent respecter la contrainte de rationalité individuelle, ou encore de *niveau de sécurité* (*security level*, Lambson, 1987), selon laquelle les profits inter-temporels le long du sentier de punition doivent être non-négatifs. La stratégie de cliquet impose que cette contrainte soit strictement vérifiée. En l'absence de différenciation des produits comme de contrainte de capacités, elle coïncide en outre avec la punition optimale (Lambertini & Sasaki, 2002). Elle apparaît enfin comme plus réaliste empiriquement. L'étude de Mason & Phillips (2002) montre ainsi que la stratégie de cliquet est largement privilégiée par les firmes expérimentales.

cette question, nous supposons cependant que le coût marginal est endogène et choisit par les firmes. L'hypothèse d'homogénéité des firmes assure alors que toutes les firmes choisiront le même coût à l'équilibre.[10]

La réduction de coût obtenue grâce au travail illégal se fait au prix du risque de détection. On note α la probabilité de détection choisie par les autorités, à travers la politique de contrôle. En cas de détection, la punition est composée de deux éléments. Les firmes doivent s'acquitter d'une part d'une amende F, choisie par les autorités, et d'autre part du remboursement de la fraude qui, pour un niveau donné de production individuelle q, s'élève à τwq. Cette dernière hypothèse se justifie par un souci de réalisme, puisque dans la plupart des pays industrialisés, en France en particulier, la détection de la fraude fiscale est suivie d'une évaluation du montant des taxes non payées (Feinstein, 1999).

En suivant l'analyse classique de l'économie du crime (Becker, 1968), une firme choisit alors le travail au noir si le bénéfice de l'illégalité excède celui de l'honnêteté. Au prix de l'équilibre concurrentiel légal[11] (*i.e.* état de profit nul en l'absence d'évasion), le bénéfice réalisé à chaque période grâce à l'illégalité correspond, sous hypothèse de neutralité au risque, à :

[10]Depuis le modèle fondateur de Rothschild (1999), un certain nombre de travaux récents ont étudié la robustesse des résultats présentés ci-dessus à l'hétérogénéité du coût marginal. Ils établissent que le paradoxe de Bertrand (Blume, 2003) comme le modèle de collusion tacite (Thal, 2004 ; Collie, 2004) continuent à s'appliquer dans les mêmes termes.

[11]Dans tout ce qui suit, le terme d'*équilibre concurrentiel* désigne tout état du marché où les firmes adoptent la stratégie qui conduit à l'annulation des profits. Lorsque le travail au noir est pris en compte, il existe deux états de profit nul (voir Proposition 3.1 ci-dessous), selon que les firmes choisissent le coût légal (*équilibre concurrentiel légal*) ou l'évasion (*équilibre concurrentiel illégal*).

$$\begin{aligned}
\Pi_F &= (1-\alpha)\left[p^c\ \frac{Q^c}{n} - w\ \frac{Q^c}{n}\right] - \alpha\ F \\
&= (1-\alpha)\left[p^c\ \frac{Q^c}{n} - (1+\tau)w\ \frac{Q^c}{n} + \tau\ w\ \frac{Q^c}{n}\right] - \alpha\ F \\
&= (1-\alpha)\left[\Pi_c + \tau\ w\ \frac{Q^c}{n}\right] - \alpha\ F \\
\Pi_F &= (1-\alpha)\tau\ w\ \frac{Q^c}{n} - \alpha\ F = \pi_F - \alpha\ F \qquad (3.2)
\end{aligned}$$

où $\pi_F = (1-\alpha)\ \tau\ w\ \dfrac{Q^c}{n}$ correspond au bénéfice brut (*i.e.* hors coût fixe lié à l'amende) de l'évasion.

Si un accord de collusion tacite pouvait constituer un équilibre du marché, le profit de l'évasion fiscale s'ajouterait à celui de la collusion. Dans les termes de la condition (3.1), l'évasion fiscale consisterait donc en un simple accroissement de Π_m. Dans ce cas, la possibilité de recourir au travail au noir ne fait par conséquent que renforcer la capacité du marché à mettre en oeuvre l'accord collusif, en accroissant la rentabilité de la production. Afin d'isoler l'effet propre de la concurrence sur la demande de travail au noir, nous nous restreignons donc aux marchés *robustes à la collusion tacite* (*collusion–proof markets*), c'est à dire tels que : $\gamma \geq \gamma_c$. Cette hypothèse découle de la volonté de séparer les effets de la collusion de ceux de la concurrence, non seulement en termes analytiques mais également en termes de politique de détection. La collusion en prix fait elle-même l'objet d'une répression par les autorités de la concurrence, qui est susceptible de révéler l'évasion fiscale. La question de la lutte contre le travail au noir se pose donc avec d'autant plus d'acuité que le marché présente toutes les apparences d'un fonctionnement concurrentiel.

Si les firmes recourent au travail légal, le seul équilibre qui subsiste sous cette hypothèse est l'équilibre concurrentiel légal, dans lequel les profits sont nuls. Ce gain

correspond donc au bénéfice de l'honnêteté sur un marché robuste à la collusion tacite, et une firme choisit alors d'opter pour le travail au noir si le profit d'évasion est positif, soit : $\pi_F \geq \alpha\ F$. Etudier la demande de travail au noir sur un marché robuste à la collusion tacite requiert que cette condition soit vérifiée.

***Hypothèse* 3.1.** *Nous considérons les marchés tels que :*

1. *Le marché est robuste à la collusion tacite :* $\gamma \geq \frac{1}{n}$ *;*
2. *L'évasion fiscale est rentable au prix de l'équilibre concurrentiel légal :* $\Pi_F > 0$.

L'Hypothèse 3.1.1 écarte toute possibilité de collusion tacite sur un prix supérieur au prix de l'équilibre concurrentiel légal p^c. Cette situation conduit traditionnellement à des profits nuls pour les firmes présentes sur le marché. Dans le cas étudié ici, les firmes ont cependant la possibilité d'abaisser leur coût marginal en recourant au travail au noir. A cet égard, la décision isolée d'une firme est décrite par l'Hypothèse 3.1.2, qui garantit qu'une firme peut accroître ses profits en choisissant le travail au noir à partir de l'équilibre concurrentiel légal. Pour connaître la demande de travail de l'industrie à l'équilibre, il faut y ajouter la dynamique de prix liée à la concurrence.

3.1.2 La malédiction de Bertrand

L'incitation à choisir le travail au noir qui découle de l'Hypothèse 3.1.2 est commune à toutes les firmes. Partant de l'équilibre concurrentiel légal, l'ensemble de l'industrie choisirait donc d'assumer le risque de détection en optant pour le coût marginal inférieur offert par l'évasion. En transposant à cette situation les résultats discutés dans la section précédente, deux états du marché s'avèrent candidats à en constituer un équilibre non-coopératif. La réduction du coût marginal peut d'abord être utilisée pour réaliser des marges positives. Dans cet état, les firmes parviennent à maintenir durablement un prix supérieur au coût marginal, en conservant le prix concurrentiel légal malgré l'évasion

fiscale. Pour cette raison, cet état est appelé *évasion collusive*. Dans le second état, une nouvelle guerre des prix s'engage.

La réduction du coût marginal peut en effet être utilisée par les firmes pour élargir leurs parts de marchés. Dans ce scénario de guerre des prix, les firmes abaissent alors leur prix tant que cette stratégie permet de dégager des profits positifs. La dynamique du marché conduit alors l'industrie à un état où les profits espérés sont nuls conditionnellement à l'évasion, appelé *équilibre concurrentiel illégal*.

Preuve La dynamique de guerre des prix avec évasion est formellement équivalente au modèle de concurrence à la Bertrand avec coût d'entrée homogène analysé par Sharkey & Sibley (1993). L'amende espérée associée à l'évasion, $-\alpha\ F$, correspond en effet à un coût fixe (puisqu'indépendant du niveau de production) inhérent à l'activité de la firme lorsqu'elle choisit l'évasion. Il s'agit cependant d'un coût fixe *récupérable* (*avoidable*, voir Wang & Yang (2001) pour une discussion), puisqu'une firme peut immédiatement s'y soustraire en choisissant d'adopter le coût légal. Dans ce dernier cas, la stratégie de tarification est, comme nous l'avons vu, d'adopter le prix de l'équilibre concurrentiel légal, p^c.

Lorsqu'une guerre des prix s'ouvre, l'équilibre du marché correspond donc aux prédictions théoriques de l'analyse de Sharkey & Sibley (1993). Nous nous contentons par conséquent, ici, d'en résumer les conclusions, renvoyant le lecteur à l'article original pour les démonstrations formelles.

D'abord les auteurs montrent qu'il n'existe pas d'équilibre en stratégie pure dans cette situation (Sharkey & Sibley, *Théorème 1*). L'intuition de ce résultat peut être comprise en remarquant que tous les états du marché présentent une incitation unilatérale à dévier. D'une part, comme nous l'avons signalé, l'équilibre concurrentiel légal ne peut être un équilibre du marché, puisque l'évasion est rentable lorsque le prix de marché est p^c. Si, d'autre part, une guerre des prix s'engage conditionnellement à l'évasion, les firmes abaissent leur prix afin de s'approprier l'intégralité du marché. Cette dynamique se poursuit jusqu'à ce que le prix atteigne le coût marginal (illégal). Dans ce cas, le profit des firmes est négatif, égal à l'amende espérée. Les firmes ont alors une préférence stricte pour la stratégie sans évasion, procurant des profits nuls.

Ce premier résultat conduit à étudier les équilibres du marché en stratégie mixte. La décision d'évasion est alors aléatoire, décrite par la probabilité β. De la même façon, la stratégie de tarification prend la forme d'une distribution $F(p)$ définie sur l'ensemble des prix rationnels $p \in [w; p^m]$ où p^m désigne le prix de monopole (*Théorème 2*). Dans ce cadre, l'équilibre symétrique du marché correspond au couple $\{\beta; F(p)\}$ tel que les profits espérés sont nuls (*Equations (3.3)* à *(3.5)*). Le comportement de tarification en cas d'évasion est donc choisi de façon à ce que les profits réalisés soient juste suffisants

à couvrir le coût fixe d'évasion.

Au total, l'équilibre non-coopératif du marché qui résulte d'une guerre des prix est donc l'équilibre concurrentiel illégal, dans lequel le profit espéré conditionnel à l'évasion est nul pour toutes les firmes du marché. ■

Par définition, l'état d'évasion collusive permet de dégager des profits positifs et est donc strictement préféré par les firmes à l'équilibre concurrentiel illégal. Il constitue par conséquent l'équilibre non-coopératif du marché s'il est exempt d'incitations individuelles à dévier. Cette condition est vérifiée si le bénéfice inter-temporel de la collusion est supérieur au profit de la déviation. Comme précédemment, la collusion consiste pour les firmes à se partager équitablement la demande qui s'adresse à elle, et d'obtenir par là le profit d'évasion Π_F définit en (3.2), tandis que la déviation permet à la firme qui fixe un prix infinitésimalement inférieur d'obtenir le bénéfice de l'ensemble du marché, $n\pi_F - \alpha.F$. Dans ce dernier cas, les autres firmes obtiennent les profits espérés négatifs issus de la politique de détection : $-\alpha\ F$. La stratégie de cliquet – qui consiste à adopter à jamais la stratégie de l'équilibre concurrentiel illégal, dont les profits sont nuls – reste la punition la plus sévère. Elle est donc utilisée pour évaluer la capacité de l'évasion collusive à constituer un équilibre du marché. Au total, l'état d'évasion collusive est donc un équilibre du marché si :

$$\begin{aligned} (1-\alpha)\tau\ w\ Q^c - \alpha\ F &\leq \frac{1}{\gamma}\left[(1-\alpha)\frac{\tau\ w\ Q^c}{n} - \alpha\ F\right] \\ &\Leftrightarrow \\ \gamma &\leq \frac{\pi_F - \alpha\ F}{n\ \pi_F - \alpha\ F} \end{aligned} \tag{3.3}$$

Sous cette condition, la tarification au prix concurrentiel légal n'est qu'une apparence de légalité dans le fonctionnement du marché puisqu'elle permet de dégager des profits positifs, offerts par l'évasion fiscale. Sur les marchés qui retiennent notre attention, cependant, la concurrence est suffisante à éviter cette situation. Un marché robuste à la collusion tacite s'avère en effet robuste à l'évasion collusive. L'équilibre concurrentiel illégal devient alors le seul équilibre stable du marché.

Proposition 3.1. *Sous les Hypothèses 3.1, les firmes choisissent l'évasion avec une probabilité strictement positive et obtiennent des profits espérés nuls.*

Preuve La robustesse à la collusion tacite contient la robustesse à l'évasion collusive. Cette conclusion se déduit directement de la comparaison entre les seuils des conditions (3.1) et (3.3). Pour tout profit brut d'évasion, π, positif on a :

$$\frac{1}{n} - \frac{\pi - \alpha\ F}{n\ \pi - \alpha\ F} = \frac{1}{n} - \frac{1}{n}\left[\frac{\pi - \alpha\ F}{\pi - \dfrac{\alpha\ F}{n}}\right] = \frac{\alpha\ F\ (n-1)}{n} \tag{3.4}$$

Pour tout marché non monopolistique ($n > 1$), cette quantité est positive. En remplaçant π par sa valeur au prix concurrentiel légal p^c, on a en particulier : $\gamma \geq \dfrac{\pi_F - \alpha.F}{n.\pi_F - \alpha.F}$ dès lors que $\gamma \geq \gamma_c$.

Puisque l'évasion collusive n'est pas un équilibre du marché, la dynamique de la concurrence conduit donc les firmes à l'équilibre concurrentiel illégal, dans lequel les profits sont nuls. ■

La Proposition 3.1 confirme la validité du paradoxe de Bertrand lorsque le coût marginal est endogène : dès lors que le marché est robuste à la collusion tacite, les firmes se trouvent dans l'impossibilité d'obtenir des profit positifs, qu'ils proviennent d'un accroissement du prix ou d'une diminution du coût. La guerre des prix conduit donc nécessairement à des profits nuls dans la mesure où l'intensité de la concurrence entraîne, avec ou sans possibilité d'évasion, une diminution du prix jusqu'à annulation des profits.

Par définition, les firmes sont donc indifférentes entre l'équilibre de profits espérés nuls auquel aboutit l'évasion – équilibre concurrentiel illégal – et l'état de profit nul originel, sans évasion – équilibre concurrentiel légal. Comme l'indique la seconde partie de la proposition, les firmes sont pourtant conduites à choisir l'évasion fiscale. La dynamique qui sous-tend ce résultat peut-être décomposée en deux phases. A partir de l'équilibre concurrentiel légal, les firmes sont d'abord incitées à dévier du coût légal par la rentabilité de l'évasion. Ensuite, cette réduction du coût marginal ouvre de nouvelles possibilités de guerre de prix, qui sont exploitées en raison de la robustesse du marché

à la collusion tacite. La rentabilité de l'évasion est donc la cause de la déviation initiale qui entraîne le marché vers l'équilibre concurrentiel illégal. Pour y remédier, la solution naturelle consiste par conséquent à se tourner vers les instruments qui pourraient contrecarrer cette rentabilité.

A cet égard, la dénonciation est du point de vue des autorités un instrument à la fois non coûteux et efficace pour lutter contre l'évasion : non coûteux parce qu'elle fait porter sur les firmes la responsabilité de la surveillance (Kaplow & Shavell, 1994) ; efficace puisque la dénonciation, lorsqu'elle est utilisée, tend accroître la probabilité de détection en révélant l'information dont disposent les acteurs du marchés. La capacité de dissuasion de la politique de répression est alors parfaite, le profit de l'évasion s'élevant $-F$. La prochaine section s'intéresse à l'équilibre du marché qui résulte de l'addition de cette surveillance endogène à celle – exogène – qu'exercent les autorités.

3.2 Dénonciation : l'équilibre de silence collusif

La dénonciation n'est possible que si l'évasion fiscale est connue des firmes présentes sur le marché. A cette fin, nous supposons que les firmes reçoivent à chaque période, après leur décision de prix, un signal parfait sur le coût choisit par chaque firme de l'industrie. Formellement, chaque firme i reçoit donc un vecteur d'informations $I_i = \left\{I_i^j : j \neq i, j = 1, ..., n\right\}$, où $I_i^j = 1$ si la firme j a choisit le travail au noir, 0 sinon. Dans ce cadre, la dénonciation introduit une nouvelle variable de décision fondée sur ce signal. Elle correspond en effet à une décision binaire, par laquelle la firme i décide de transmettre (1) ou non (0) le signal reçu sur la firme j aux autorités. La décision de dénonciation est donc une fonction de I_i dans $\{0,1\}^{n-1}$: $D_i(I_i) = \left\{D_i^j(I_i^j) : j \neq i, j = 1, ..., n\right\}$, $D_i^j(I_i^j) = \{0,1\}$. Lorsque la décision de dénonciation est prise en compte, une stratégie de la firme i à la période t est donc le triplet formé du prix, du coût et du vecteur de dénonciations : $\{p_{i,t}; W_{i,t}; D_{i,t}\left(I_{i,t}\right)\}$. L'évasion

de la firme i est alors *dénoncée* si : $\sum_{j \neq i} D_j^i(1) > 0$; et la firme i est *dénonciatrice* de l'évasion lorsque : $\sum_{j \neq i} D_i^j(1) > 0$.

Pour la clarté de la présentation, cette hypothèse d'information parfaite a été écartée dans la section précédente. On peut cependant interpréter le modèle de la Section 3.1 comme la situation de référence dans laquelle la dénonciation n'a aucun effet sur la probabilité de détection encourue par les firmes. Dans ce cas, tout se passe comme si la stratégie de dénonciation était contrainte : $D_i^j(I_i^j) = 0, \forall I_i^j = 0, 1; \forall i, j; i \neq j$. De plus, le signal n'est reçu qu'après la décision en matière de prix. Cette information n'a donc pas d'autre effet sur le comportement des firmes que celui qui passe par la dénonciation. En particulier il n'a aucune influence sur la coordination, puisque l'information n'est fournie qu'*ex post*. L'introduction du signal dans le modèle, sous hypothèse de dénonciation contrainte, laisserait donc inchangés les résultats de la Section 3.1.2.[12]

Dans cette section, la dénonciation est supposée efficace, au sens où elle est utilisée par les autorités pour mettre en oeuvre les sanctions. Une firme détectée grâce à la dénonciation se voit donc infliger la sanction associée à l'évasion, c'est à dire la somme du montant de l'évasion et de l'amende F.

3.2.1 Dénonciation et concurrence

Conditionnellement à l'absence d'évasion, le prix de l'équilibre concurrentiel légal est par définition le seul prix stable sur un marché robuste à la collusion tacite. Comme nous l'avons indiqué plus haut, l'efficacité de la dénonciation renforce cette stabilité puisqu'elle permet aux firmes de briser la rentabilité de l'évasion garantie par l'Hypo-

[12]Cette conclusion n'est valide qu'à condition que l'information soit fournie *ex post*, c'est à dire après le choix du prix. Dans le cas contraire, les stratégies de marché s'apparenteraient à une décision séquentielle dans laquelle les firmes choisissent le prix après avoir observé le coût de leurs concurrentes. Voir Elberfeld & Wolfstetter (1999) pour une analyse théorique.

thèse 3.1.2. Pour que cet état constitue un équilibre du marché, il faut encore, cependant, qu'il soit exempt de l'incitation individuelle à dévier offerte par l'évasion. En l'absence de dénonciation, la Proposition 3.1 a répondu par la négative à cette question. Cette section est consacrée à l'influence de la dénonciation sur ce résultat. Elle se concentre par conséquent sur la stratégie de dénonciation adoptée par les firmes lorsqu'elles choisissent l'évasion fiscale.

a) Silence collusif

A cette fin, nous supposons que la dénonciation révèle parfaitement aux autorités le coût choisi par la firme dénonciatrice. Une firme dénonciatrice qui s'adonne elle-même au travail au noir se voit alors infliger – en plus du remboursement de la taxe – l'amende notée F' ($\leq F$) avec certitude.[13] En l'absence de dénonciation – c'est à dire pour une firme qui n'est ni dénoncée ni dénonciatrice – la politique de détection est supposée inchangée et les sanctions sont appliquées avec la probabilité α.

[13]Cette hypothèse est conforme à la tradition des modèles de clémence appliqués à l'économie industrielle. Dans ce cadre, c'est en effet la participation à un accord de collusion qui permet aux firmes d'en connaître l'existence, et la dénonciation a alors simultanément valeur d'aveu. Dans notre contexte, cette situation correspondrait au cas où seules les firmes qui fraudent obtiennent de l'information sur l'évasion de leur concurrente. Dans le modèle, nous n'avons pas jugé pertinent de restreindre l'information obtenue par les firmes en fonction du coût choisi. L'hypothèse de révélation parfaite par la dénonciation n'apparaît cependant que comme une simplification, sans conséquence sur la portée des résultats. Une version plus générale du modèle consisterait à considérer que la dénonciation attire l'attention des autorités, mais ne donne pas lieu à une détection systématique. Dans cette version du modèle la dénonciation changerait non seulement l'amende (F') mais également la probabilité de détection (qui devient, par exemple, ϕ) de la firme dénonciatrice. L'amende espérée qu'elle encoure deviendrait alors $\phi\, F' < F'$. Au regard de la statique comparative présentée ci-dessous (Section 3.2.2), cette généralisation du modèle ne ferait que renforcer la crédibilité de la menace de dénonciation et faciliter par là la mise en oeuvre du silence collusif. Les mécanismes à l'oeuvre apparaissent donc plus clairement en imposant l'hypothèse que $\phi = 1$, sans toutefois changer les résultats qualitatifs.

Compte tenu de cette politique de sanction, la dénonciation annihile toute incitation à dévier de l'équilibre concurrentiel légal en choisissant l'évasion. L'évasion ne peut donc être rentable qu'à condition que les firmes n'aient pas intérêt à la révéler. La dénonciation peut en revanche être utilisée pour punir les baisses de prix, dans l'espoir de maintenir un prix qui assure un profit positif. Cet état, dans lequel les firmes n'utilisent la dénonciation qu'en cas d'abaissement du prix par l'une de leurs concurrentes, est appelé *silence collusif*.

Définition 3.1. *On appelle* ***silence collusif*** *l'état dans lequel la stratégie de la firme* i, $\forall i, j \neq i, t$ *est :*

$$p_{i,t}^{*} = p^{c}; W_{i,t}^{*} = w; D_{i,t}^{j*}(0) = 0; D_{i,t}^{j*}(1) = \begin{cases} 0 & si\ p_{j,t} \geq p^{c} \\ 1 & si\ p_{j,t} < p^{c} \end{cases}$$

Dans l'état de silence collusif, les firmes tiennent donc secrète l'évasion de leurs concurrentes tant que celles-ci maintiennent le prix de l'équilibre concurrentiel légal. Cette tarification permet à chacune d'obtenir le profit de l'évasion. Tout abaissement du prix se trouve, en revanche, sanctionné par une dénonciation. Cet état tient donc son nom de ce que la dénonciation est utilisée par les firmes comme un instrument de mise en oeuvre de la collusion en prix.

Il faut noter, cependant, que la dénonciation est coûteuse pour une firme qui a choisit l'évasion, puisqu'elle encourt alors avec certitude les sanctions infligées à une firme dénonciatrice. Le bénéfice de cette décision de dénonciation est, quant à lui, composé des profits réalisés grâce au maintient du prix concurrentiel légal malgré l'évasion. Pour que la stratégie adoptée dans l'état de silence collusif soit individuellement rationnelle, il faut donc que les profits espérés de la collusion excèdent le coût de la dénonciation, c'est à dire que : $-F' + \sum_{t=1}^{\infty}(1-\gamma)^{t}\Pi_{F} \geq 0$.[14] Dans ce cas – et dans ce cas seulement – la

[14]Pour la clarté de l'exposition, nous considérons le prix concurrentiel légal comme une référence naturelle sur un marché robuste à la collusion tacite. Par définition de l'équilibre concurrentiel, les profits associés au coût légal sont nuls lorsque ce prix est choisi. Les comportements sont alors guidés unique-

dénonciation est alors une menace crédible contre les diminutions de prix. La crédibilité de la menace fait, en retour, du silence collusif l'équilibre non-coopératif du marché.

Proposition 3.2. *La stratégie de dénonciation de l'état de silence collusif est une menace crédible si :*

$$\gamma \leq \frac{\Pi_F}{F' + \Pi_F} \equiv \gamma^F \tag{3.5}$$

Sous cette condition, l'état de silence collusif est l'équilibre non-coopératif du marché.

Preuve L'état de silence collusif est un équilibre du marché si la stratégie décrite dans la Définition 3.1 est la meilleure réponse à cette même stratégie. Il constitue le seul équilibre si les firmes ont intérêt à dévier de l'équilibre concurrentiel légal, c'est à dire si l'état de silence collusif est la meilleure réponse à l'équilibre concurrentiel légal.

Par définition de la crédibilité de la menace (*i.e.* sous (3.5)), la dénonciation est la meilleure réponse à un abaissement du prix dès lors qu'elle permet de revenir ultérieurement au prix du silence collusif. Le profit associé à un abaissement du prix est donc égal à la sanction infligée à une firme dénoncée, soit $-F$. Les firmes maintiennent par conséquent le prix de collusion afin d'en obtenir les profits positifs et la stratégie de l'état de silence collusif est exempte d'incitation individuelle à dévier.

Sur un marché robuste à la collusion tacite, les firmes ne sauraient maintenir un prix supérieur à celui de l'équilibre concurrentiel légal en choisissant le coût légal. Le profit de l'honnêteté est donc nul. La dénonciation est par ailleurs non coûteuse pour une firme qui choisit le coût légal. La stratégie de dénonciation optimale à l'équilibre concurrentiel légal consiste donc à dénoncer toute firme qui utilise l'évasion pour abaisser son prix : $D_{i,t}^{j*}(1) = \begin{cases} 0 & \text{si } p_{j,t} \geq p^c \\ 1 & \text{si } p_{j,t} < p^c \end{cases}$.

L'état de silence collusif ne diffère donc de la stratégie de l'équilibre concurrentiel légal que par l'évasion fiscale. Dans la mesure où celle-ci est rentable par définition, la meilleure réponse à l'équilibre concurrentiel légal est donc de choisir le coût illégal. ■

---

ment par les profits d'évasion. Limiter la présentation du modèle au prix de l'équilibre concurrentiel permet donc en outre d'en faire apparaître plus clairement les mécanismes essentiels, liés à l'évasion. La multiplicité des équilibres est cependant une propriété commune à tous les modèles de collusion, à laquelle l'évasion collusive ne fait pas exception. Si, en particulier, la condition de crédibilité de la menace est vérifiée pour un niveau de prix quelconque p, elle le sera également pour tout autre prix $p' > p$. Cet aspect est pris en compte dans la Section 3.4.1, qui généralise l'analyse à tout prix d'évasion collusive.

Plutôt qu'un rempart permettant aux firmes de se protéger d'une évasion contrainte par l'intensité de la concurrence, la dénonciation agit donc comme une menace qui rend durablement rentable l'illégalité. Lorsque les autorités décident de rendre efficace la dénonciation, elles instaurent un instrument de coordination au service d'une collusion en prix assise sur le silence.

b) Silence collusif et collusion tacite

L'Hypothèse 3.1.1 a été introduite dans l'objectif d'évacuer la possibilité que les firmes recourent à un accord de collusion. Pour ce faire, nous nous restreignons aux marchés dont la probabilité de destruction est suffisamment forte, $\gamma > \gamma_c$. Comme l'indique la Proposition 3.2, le silence collusif est par ailleurs l'équilibre d'un marché sur lequel la dénonciation est efficace si la probabilité de détection est suffisamment faible, telle que : $\gamma < \gamma^F$. L'écart entre ces seuils mesure donc la propension d'un marché robuste à la collusion tacite à mettre en oeuvre le silence collusif. Cette quantité, appelée *intervalle de silence collusif* d'un marché robuste à la collusion tacite, est mesurée par :

$$R = \tau \; w \; Q^c - \frac{n}{n-1} \; \frac{F' + \alpha \; F}{1-\alpha} \tag{3.6}$$

Les marchés robustes à la collusion tacite dont l'intervalle de silence collusif est positif offrent alors l'apparence d'un fonctionnement concurrentiel (tarification au prix p^c) qui masque une évasion fiscale gardée secrète :

Proposition 3.3. *Si la dénonciation est une menace crédible ($\gamma < \gamma^F$), l'état de silence collusif est l'équilibre de tout marché robuste à la collusion tacite tel que* $(n-1)\Pi_F > F'$.

Preuve La dénonciation est une menace crédible sur un marché robuste à la collusion tacite si sa probabilité de destruction est telle que : $\gamma_c \leq \gamma \leq \gamma^F$. En utilisant la définition de γ^F en (3.5) et celle

du profit d'évasion (3.2), on a :

$$\gamma^F \geq \frac{1}{n} \Leftrightarrow \frac{\Pi_F}{F' + \Pi_F} \geq \frac{1}{n} \Leftrightarrow \Pi_F \geq \frac{F'}{n-1}$$
$$\Rightarrow R \equiv \tau\ w\ Q^c - \frac{n}{n-1}\ \frac{F' + \alpha\ F}{1-\alpha} > 0$$

L'intervalle de silence collusif R – mesuré ici comme la différence entre le profit global de l'évasion ($\tau\ w\ Q^c$) et la sanction espérée qui découle de la taille de l'industrie et de la probabilité de détection – est donc positif si : $\Pi_F \geq \frac{F'}{n-1}$, soit $(n-1)\Pi_F > F'$. ■

La condition de validité présentée dans la Proposition 3.3 peut s'interpréter en termes simples d'arbitrage. Le terme de droite, F', représente le coût en valeur absolue que doit supporter un dénonciateur. Π_F désigne le profit individuel d'évasion dans l'état de silence collusif. Le terme de gauche de l'inégalité, $(n-1)\Pi_F$, mesure par conséquent le bénéfice de la dénonciation, c'est à dire les profits réalisés par l'ensemble des firmes présentes sur le marché à l'exception de la firme qui dévie (*i.e.* qui poste un prix inférieur à p^c). La condition établit donc qu'un marché robuste à la collusion tacite peut mettre en oeuvre l'état de silence collusif si le bénéfice instantané de la dénonciation en excède le coût.

Au total, les résultats du modèle mettent en évidence l'utilisation par les firmes de la dénonciation comme une barrière non pas contre l'évasion fiscale mais contre le déclenchement d'une guerre des prix. Sur un marché robuste à la collusion tacite, cet effet de la dénonciation dépend des caractéristiques de l'environnement par deux canaux. D'une part, la dénonciation constitue une menace de moins en moins coûteuse à mesure que γ^F augmente. L'état de silence collusif s'en trouve d'autant plus facilement mis en oeuvre. Au fur et à mesure de cette augmentation de γ^F, d'autre part, l'intervalle de silence collusif s'accroît et de plus en plus de marchés robustes à la collusion tacite sont aptes à maintenir secrète l'évasion. La prochaine section évalue l'influence des caractéristiques du marché et de la politique de détection sur chacun de ces effets.

3.2.2 Statique comparative du modèle : l'effet des clauses de clémence

Le Tableau 3.1 résume la sensibilité des conditions décrites ci-dessus aux variables exogènes du modèle.

Preuve D'après l'expression (3.5), γ^F est inférieur à 1 tant que $F' > 0$. R est donc positif ou négatif selon que : $(n-1)\Pi_F > F'$ (voir Proposition 3.3). A l'inverse, dès que F' devient négative on a : $\Pi_F > F' + \Pi_F$ et donc $\gamma^F > 1$. L'intervalle de silence collusif, qui est une mesure de l'écart entre γ^F et γ_c, est donc toujours positif (*i.e.* $(n-1)\Pi_F > F'$ est trivialement vérifiée tant que l'évasion est rentable).

La statique comparative est obtenue par simple différentiation des conditions définissant R (3.6) et γ^F. Les caractéristiques du marché ($\beta \in \{\tau; w; Q^c\}$) ont une influence sans ambiguïté sur les variables d'intérêt :

$$\frac{\partial \gamma^F}{\partial \beta} = (1-\alpha)\,\frac{F'}{(F'+\Pi_F)^2} > 0\,;\; \frac{\partial R}{\partial \beta} = 1 > 0$$

L'effet de la probabilité de détection est obtenu par manipulation des expressions :

$$\begin{aligned} \frac{\partial \gamma^F}{\partial \alpha} &= -\tau\,\frac{Q^c}{n} - F < 0 \\ \frac{\partial R}{\partial \alpha} &= -\frac{n}{1-\alpha}\left(\alpha F + \frac{F'}{(1-\alpha)(n-1)}\right) \end{aligned} \tag{3.7}$$

Lorsque F' est positive, R est donc strictement décroissant de la probabilité de détection. Dans le cas contraire, R est positif si (3.7) l'est soit : $F' > n(n-1)\alpha$. L'effet de l'amende, quant à lui, provient de :

$$\frac{\partial \gamma^F}{\partial F'} = -\frac{\alpha F'}{(\Pi_F + F')^2}\,;\; \frac{\partial R}{\partial F'} = -\frac{\alpha n}{1-\alpha} < 0$$

L'effet de la taille de l'industrie est décrit par :

$$\begin{aligned} \frac{\partial \gamma^F}{\partial n} &= \frac{\partial \Pi_F}{\partial n}.\frac{\Pi_F + F' - \Pi_F}{(\Pi_F + F')^2} = -(1-\alpha)\frac{\tau\; w\; Q^c}{n^2}\;\frac{F'}{(\Pi_F + F')^2} \\ \frac{\partial R}{\partial n} &= \frac{\alpha(n-1)^2 F - F'}{(1-\alpha)(n-1)^2} \end{aligned} \tag{3.8}$$

En conséquence, $\dfrac{\partial \gamma^F}{\partial n}$ est du signe opposé à celui de F'. En l'absence de programme de clémence (ou lorsque celui-ci se limite à une réduction d'amende), R est croissant de n tant que $\dfrac{F'}{F} < \dfrac{1}{\alpha(n-1)^2}$,

décroissant sinon. Il est monotone croissant de n lorsque le programme de clémence récompense la dénonciation. Outre leur influence sur la sensibilité des variables aux paramètres exogènes, les programmes de clémence ont un effet univoque sur R et γ^F :

$$\frac{\partial \gamma^F}{\partial F'} = -\frac{\Pi_F}{(\Pi_F + F')^2} < 0\,;\; \frac{\partial R}{\partial F'} = -\frac{n}{(n-1)(1-\alpha)} < 0$$

L'ensemble de ces éléments est synthétisé dans le Tableau 3.1. ■

Tableau 3.1 – Statique comparative de l'équilibre de silence collusif

	Exemption ($F' > 0$)							**Clémence**	**Bonus** ($F' < 0$)						
	Signe	τ	w	Q^c	α	F	n	(F')	Signe	τ	w	Q^c	α	F	n
R	$+/-^a$	$+$	$+$	$+$	$-$	$-$	$+/-^b$	$-$	$+$	$+$	$+$	$+$	$+/-^c$	$-$	$+$
γ_F	< 1	$+$	$+$	$+$	$-$	$-$	$-$	$-$	> 1	$-$	$-$	$-$	$-$	$+$	$+$

[a] Positif si : $(n-1)\Pi_F > F'$ [b] Positif si : $\frac{F'}{F} < \frac{1}{\alpha(n-1)^2}$ [c] Positif si : $F' > n(n-1)\alpha$

Les colonnes intitulées "signe" présentent l'intervalle de valeurs possible pour R et γ^F. Les programmes de clémence, instaurés pour encourager la dénonciation et ainsi lutter contre les activités illégales, consistent en une réduction de l'amende infligée au dénonciateur. Dans les termes du modèle, une mesure de clémence se traduit donc par une diminution de F'. En raison de leurs différences considérables d'efficacité, il est devenu classique (Spagnolo, 2002 ; 2003) de distinguer les mesures de clémence selon qu'elles offrent une exemption partielle d'amende ($F > F' > 0$) ou une récompense ($F > 0 > F'$) à la dénonciation. Au regard de l'influence des variables exogènes selon que règne un programme d'exemption (partie gauche du Tableau 3.1) ou de récompense (partie droite), comme de l'impact des clauses de clémence elles-mêmes (partie centrale), la statique comparative présentée ici confirme une différence marquée.

a) Absence de clauses de clémence

En l'absence de clause de clémence, ou lorsque le programme prend la forme d'exemptions ($F' > 0$), la condition de crédibilité de la menace (3.5) peut ne pas être vérifiée

($\gamma^F < 1$). L'intervalle de silence collusif peut en conséquence être ou positif ou négatif, selon l'importance de l'amende infligée à un dénonciateur (première colonne du Tableau 3.1).

La capacité du marché à mettre en oeuvre l'état de silence collusif, comme l'intervalle de silence collusif d'un marché robuste à la collusion tacite, sont croissants avec le profit de l'évasion puisque, à coût (F') constant, la dénonciation devient de plus en plus rentable. En conséquence, toutes les variables qui accroissent le profit d'évasion (taux d'imposition, salaire, niveau de la demande) accroissent la capacité du marché à mettre en oeuvre le silence collusif et élargissent l'intervalle de silence collusif. Pour la même raison, les composantes de la politique de détection (amende, probabilité de détection), qui diminuent le profit de l'évasion, diminuent la stabilité du silence collusif.

Le résultat, traditionnel en économie industrielle, selon lequel la taille de l'industrie constitue une barrière naturelle à la collusion, reste valide dans le cas du silence collusif. La taille de l'industrie agit comme une diminution indirecte du profit d'évasion, puisqu'elle détermine le partage du profit global. La condition (3.5) est donc d'autant plus difficilement vérifiée que la taille de l'industrie est importante. Son effet sur l'intervalle de silence collusif est, en revanche, ambigu : à mesure que la taille de l'industrie s'accroît, γ^F et γ_c augmentent simultanément. L'intervalle de silence collusif ne se réduit avec la taille de l'industrie que si le pourcentage d'exemption ($1-F'/F$) est suffisamment faible (note b, Tableau 3.1). Lorsque l'exemption devient importante, la taille de l'industrie tend à l'inverse à élargir l'intervalle de silence collusif. Grâce au programme de clémence, un marché robuste à la collusion tacite est donc d'autant plus probablement à même de mettre en oeuvre le silence collusif que sa taille est importante.

b) Programme actif de clémence

Au delà de cet effet pervers, les programmes de clémence influencent considérablement la capacité d'un marché à mettre en oeuvre le silence collusif. Puisque la dénonciation joue le rôle de menace mise au service de la collusion, encourager la menace conduit en effet à faciliter sa mise en oeuvre, et les programmes de clémence conduisent alors à un renforcement du silence collusif. D'abord, l'effet de la réduction d'amende offerte au dénonciateur est de faciliter l'accès à la menace, à travers l'augmentation de γ^F (partie centrale du Tableau 3.1). Les clauses de clémence sont donc contre-productives *ex ante*, au sens où elles encouragent l'entrée dans l'accord de silence collusif. S'y ajoute un second échec, puisque les clauses de clémence étendent également l'intervalle de silence collusif. La proportion de marchés robustes à la collusion tacite mais pour lesquels le silence collusif est un équilibre tend donc à s'accroître.

De plus, si les autorités mettent en place un système de bonus – par lequel la dénonciation est récompensée – le coût de la dénonciation disparaît et seuls en subsistent les bénéfices, qui s'élèvent alors à la somme de la récompense et du profit inter-temporel de la collusion. Dans ce cas, la menace de dénonciation est crédible en toute circonstance ($\gamma^F > 1$) et tout marché robuste à la collusion tacite peut mettre en oeuvre le silence collusif ($R > 0$).

Enfin, nous avons vu ci-dessus que la taille de l'industrie constitue une barrière naturelle contre la mise en oeuvre du silence collusif : lorsque la taille de l'industrie s'accroît, la condition qui assure la crédibilité de la menace est de plus en plus contraignante et de moins en moins de marchés sont susceptibles d'y recourir. L'introduction de clauses de clémence tend à affaiblir cet effet : $\dfrac{\partial^2 \gamma^F}{\partial n \partial F'} \leq 0$. A mesure que l'amende offerte au dénonciateur s'accroît, la limite imposée au silence collusif par la taille de l'industrie tend donc à s'affaiblir.

Preuve Par différenciation de (3.8), on a :

$$\begin{aligned}\frac{\partial^2\gamma^F}{\partial n\partial F'} &= \frac{\partial^2\Pi_F}{\partial n\partial F'}.\frac{F'}{(\Pi_F+F')^2}+\frac{\partial\Pi_F}{\partial n}.\frac{(\Pi_F+F')\,(\Pi_F+F'-F')}{(\Pi_F+F')^4}\\ &= \frac{\partial^2\Pi_F}{\partial n\partial F'}.\frac{F'}{(\Pi_F+F')^2}+\frac{\partial\Pi_F}{\partial n}.\frac{(\Pi_F-F')}{(\Pi_F+F')^3}\end{aligned}$$

Sachant que $\frac{\partial^2\Pi_F}{\partial n\partial F'}=0$, il vient : $\frac{\partial^2\gamma^F}{\partial n\partial F'}=\underbrace{\frac{\partial\Pi_F}{\partial n}}_{\leq 0}.\underbrace{\frac{(\Pi_F-F')}{(\Pi_F+F')^3}}_{\geq 0}$. Tant que $|F'|$ reste inférieur au profit de collusion, on a donc $\frac{\partial^2\gamma^F}{\partial n\partial F'}<0$. ■

Proposition 3.4. *Les programmes de clémence tendent à encourager la mise en oeuvre du silence collusif. Lorsque la clause de clémence prend la forme de bonus, le silence collusif est toujours un équilibre.*

Le modèle présenté ici a permis d'isoler l'influence de la concurrence sur la propension des firmes à recourir au travail au noir. En l'absence de dénonciation, l'intensité de la concurrence contraint les firmes à choisir le coût minimum. La guerre des prix qui s'en suit annule cependant les profits afférents à l'évasion. Loin d'offrir une protection contre ce mécanisme, la dénonciation permet d'imposer une discipline qui ouvre de nouvelles possibilités de collusion : elle fournit une menace de punition contre les baisses de prix, qui permet aux firmes de dégager le profit d'une évasion tenue secrète. Dans ce cadre, les programmes de clémence apparaissent comme un encouragement au silence plutôt qu'à la dénonciation : facilitant l'accès à la menace crédible qui permet de mettre en oeuvre le silence collusif, elles renforcent la capacité du marché à le pratiquer. La pertinence empirique de ces prédictions est testée en reproduisant les hypothèses du modèle dans le cadre de marchés expérimentaux, présentés dans les prochaines sections.

3.3 Présentation des marchés expérimentaux

Comme nous l'avons déja souligné (Chapitre 1, Section 1.2), la méthode expérimentale permet par bien des aspects de résoudre les difficultés inhérentes à l'analyse empirique des activités illégales. A ces avantages s'ajoute ici celui de pouvoir reproduire l'environnement économique considéré dans le modèle, afin de tester ses prédictions théoriques. Il est, par exemple, particulièrement délicat de juger de la robustesse d'un marché réel à la collusion tacite. L'expérience en laboratoire que nous réalisons offre, quant à elle, la possibilité de contraindre les marchés à respecter les conditions de l'Hypothèse 3.1 et ainsi d'étudier la mise en oeuvre du silence collusif sur des marchés robustes à la collusion tacite.

3.3.1 Cadre de l'expérience

L'objectif de l'expérience est d'offrir un test des prédictions du modèle que sont la malédiction de Bertrand et les conditions d'émergence de l'évasion collusive. A cette fin, le protocole est conçu de façon à créer un environnement économique qui reproduit le plus fidèlement possible les hypothèses essentielles du modèle. La nécessité de garantir la cohérence entre cet environnement et les hypothèses du modèle nous a cependant conduit à retenir un ajustement important.

Les résultats théoriques reposent sur l'hypothèse que les firmes sont homogènes en tout point, en termes de fonction de profit en particulier. Cette hypothèse implique notamment que les firmes n'ont ni préférence pour l'honnêteté ni aversion au risque, qui introduiraient autant de sources d'hétérogénéité dans les décisions de coût.[15] Même si la neutralité des instructions (décrites dans les prochaines sections) est destinée à traiter

[15]La Section 3.5 propose une discussion plus complète des perspectives de recherche ouvertes par la prise en compte de cette hétérogénéité.

en partie cette question, il est à l'évidence impossible d'imposer le strict respect de cette hypothèse d'homogénéité dans l'expérience. Contrairement aux hypothèses du modèle, il est donc possible que les participants manifestent, en raison de leur hétérogénéité inobservable, des préférences vis-à-vis de l'évasion qui diffèrent du seul calcul de profit. La dénonciation, non coûteuse pour une firme qui choisit le coût légal, pourrait alors être utilisée par des participants qui désapprouvent l'évasion pour contraindre le marché à adopter l'équilibre concurrentiel légal.

Dans le cas particulier qui est le nôtre, faisant intervenir la dénonciation, cette hétérogénéité peut en outre être à l'origine de phénomènes de "pression des pairs", au sens où les décisions individuelles peuvent être influencées par la perception qu'en a le groupe d'appartenance (Fehr & Gächter, 2000a). La dénonciation constituerait alors un instrument de punition monétaire pour les firmes qui manifestent une préférence intrinsèque pour la légalité et condamnent le recours au travail au noir (Falk, Fehr & Fischbacher, 2005). En l'absence même d'instrument de punition contre les comportements jugés blâmables, la seule connaissance du comportement des membres du groupe peut également être à l'origine d'un phénomène de pression des pairs (Falk & Ichino, 2005) et de mimétisme social (Falk & Fischbacher, 2002).

Afin de limiter l'incidence de ces effets – hétérogénéité inobservable vis-à-vis de la légalité et pression des pairs – nous avons choisi de restreindre la possibilité de dénoncer aux seules firmes qui ont choisi l'évasion. Etant donnée la très forte prédominance empirique de cette décision, en l'absence même de possibilités de dénonciation (voir ci-dessous, Section 3.4.2), cette restriction ne devrait pas altérer nos résultats, tout en les préservant des déterminants "hors-modèle" du comportement.

3.3.2 Protocole expérimental

A cette exception près, le protocole de l'expérience reproduit l'environnement économique décrit dans les Sections 3.1.1 et 3.2. Les prédictions théoriques sont testées par trois traitements, accroissant progressivement la facilité de la dénonciation.

a) Description de l'expérience

Les participants se voient attribuer le rôle de firmes et sont regroupés aléatoirement pour former les marchés. La répétition du jeu est assurée en maintenant constante la taille comme la composition de ces groupes pendant toute l'expérience. L'horizon de répétition est reproduit par le biais d'une interruption aléatoire du jeu, susceptible de survenir à chaque période avec la probabilité γ.[16] Pour des raisons pratiques, un seul tirage est effectué à chaque période pour l'ensemble des marchés.

Au sein d'une période, la concurrence à la Bertrand avec coût endogène constitue un jeu en deux étapes : les firmes doivent choisir individuellement et de façon privée d'abord un coût (élevé ou faible), puis un prix (parmi l'ensemble des valeurs discrètes réalisables).

Le prix de marché détermine les quantités vendues selon la spécification linéaire : $Q = d - lp$. Plutôt que par sa forme algébrique, qui fait appel à leurs compétences mathématiques, cette fonction de demande est exposée aux participants par l'intermédiaire d'un tableau (reproduit dans l'Annexe, Tableau 3.A). Chaque cellule contient les quantités individuelles vendues, lorsque le prix de marché est celui qui apparaît en ligne, par chacune des firmes actives dont le nombre est indiqué en colonne. Le gain pour la période d'une firme active est donc calculé selon la formule $Q/n\ (p - W)$, tandis que le

[16] Les valeurs numériques utilisées dans l'expérience sont décrites en Annexe, Section 3.B.

gain d'une firme inactive est nul. Les pénalités inhérentes à l'évasion sont retranchées de ces gains. A cette fin, un tirage au sort est réalisé pour chaque firme ayant choisi le coût faible. En vertu de ce tirage, le gain de la firme pour la période est négatif, égal à $-F$, avec la probabilité α.

L'ensemble de ces éléments permet de reproduire l'environnement concurrentiel dans lequel s'inscrivent nos résultats théoriques. L'effet de la dénonciation et des clauses de clémence sur les décisions des participants est observé en introduisant plusieurs variations dans l'environnement.

b) Traitements

Un premier traitement reproduit les conditions qui conduisent à la malédiction de Bertrand et sert ainsi de traitement de référence. Dans ce traitement de Contrôle, les participants n'ont donc pas la possibilité de dénoncer l'évasion fiscale. La dénonciation est introduite dans les deux autres traitements, intitulés Dénonciation et Clémence. L'objectif de ces traitements est d'apprécier l'influence de la dénonciation sur les décisions d'évasion comme de tarification, par comparaison avec les comportements observés dans le traitement de Contrôle. L'introduction de la dénonciation nécessite en particulier d'informer les participants des décisions prises par les autres firmes du marché. Pour éviter que l'effet de l'information s'ajoute à celui de la dénonciation dans cette comparaison, l'information est fournie aux participants dès le traitement de Contrôle.

Pour ce faire, une fenêtre contenant la liste des décisions prises par l'ensemble des firmes apparaît sur chaque écran, avant l'annonce des gains pour la période. Chaque ligne de cette liste se rapporte à une firme du marché, et comporte le prix qu'elle a choisi. Si la firme sur l'écran de laquelle la fenêtre apparaît a choisi l'évasion (coût faible), le coût sélectionné est ajouté, à côté du prix, sur cette ligne. L'ordre d'apparition des firmes sur la liste est modifié aléatoirement à chaque période pour limiter l'apparition

de réputations individuelles. Le jeu ne se poursuit qu'après que chaque participant a fermé la fenêtre d'information, assurant ainsi qu'elle a – au moins – été consultée.

Dans les traitements Dénonciation et Clémence, les participants peuvent utiliser cette information pour dénoncer les firmes qui ont choisi le coût faible. La possibilité de dénoncer consiste concrètement en une case à cocher dans la fenêtre d'information, qui s'ajoute à côté du coût lorsqu'il est fourni. Conformément au choix méthodologique discuté ci-dessus, l'usage de la dénonciation est donc réservé aux participants qui ont choisi le coût faible.[17] Avant de clore la fenêtre d'information, les participants peuvent cocher les firmes de leur choix, sans limitation de nombre. Toute firme qui a été dénoncée au moins une fois se voit attribuer le profit négatif $(-F)$ associé à la détection. Les firmes qui se sont mises en position de dénonciatrice – en dénonçant au moins une autre firme – obtiennent quant à elles un gain pour la période égal au profit de dénonciation, $-F'$. Ce profit est fixé à $-F' = -F$ dans le traitement Dénonciation, à $-F' > -F$ dans le traitement Clémence.

Outre ces trois traitements, des variations sont également introduites dans les tailles des groupes, formés de 3 à 6 participants.[18] La taille de l'industrie, mesurée par le nombre de firmes actives, peut différer du nombre de firmes présentes – la taille du groupe – en raison des prix qu'elles choisissent. La taille du groupe constitue donc un plafond sur la taille que le marché est susceptible d'atteindre. Ces variations permettent néanmoins d'observer l'influence du nombre de firmes présentes sur leur capacité à se coordonner. La probabilité d'interruption du jeu (γ) étant maintenue constante, les variations du nombre de firmes actives constituent en outre autant de variations dans

[17] Compte tenu des hypothèses retenues, la dénonciation d'un participant ayant choisi le coût élevé serait sans influence sur les gains comme les stratégies des participants. Par souci de simplification, le protocole de l'expérience impose donc que seuls les participants qui ont choisi le coût faible peuvent être dénoncés.

[18] Ces tailles des groupes ont été choisies pour leur capacité à représenter l'éventail des tailles de marchés, de la plus petite à la plus grande, conformément au résultat célèbre de Selten (1973) : «*four are few and six are many*».

la capacité du marché à mettre en oeuvre le silence collusif (condition (3.5)).

c) Déroulement des sessions

A leur arrivée, les participants procèdent à un tirage au sort qui sélectionne l'ordinateur qui leur est attribué. Ce tirage au sort détermine simultanément le marché auquel ils appartiennent, ainsi que la taille de leur groupe. Ces groupes sont maintenus inchangés – en termes de tailles comme de composition – pendant l'ensemble de l'expérience (protocole en *partners*).

Les participants jouent successivement chacun des trois traitements, selon l'ordre de présentation adopté ci-dessus : Contrôle, puis Dénonciation et enfin Clémence. Les instructions afférentes à chaque traitement sont distribuées et lues juste avant son déroulement.[19] S'ils connaissent le nombre de traitements, les participants ignorent par conséquent la nature des traitements ultérieurs. Cette procédure permet d'éviter que les comportements dans un traitement soient influencés par le comportement anticipé dans le(s) traitement(s) subséquent(s). En comparaison d'autres jeux expérimentaux, les institutions de marché font intervenir des fonctions de paiements et des interactions assez complexes (Holt, 1995). Pour permettre aux participants de se familiariser avec ces règles, il est donc devenu classique d'ajouter une phase préliminaire d'entraînement, au cours de laquelle les décisions sont sans conséquence sur la rémunération (Fouraker & Siegel, 1963). Contrairement aux traitements, dont l'interruption est aléatoire, la durée de la phase d'entraînement est certaine et fixée à trois périodes. Cet élément est de connaissance commune entre les participants, qui sont encouragés à utiliser ces périodes pour tester leur compréhension du jeu. A la fin de l'expérience, un questionnaire est proposé au participants afin de recueillir des informations sur leurs caractéristiques individuelles (sexe, niveau d'éducation, ...).

[19] Les instructions de l'ensemble de l'expérience sont reproduites en Annexe, Section 3.A.

Les instructions lues aux participants sont rédigées en termes neutres, éliminant toute référence à un contexte de concurrence comme d'illégalité.[20] Le prix est désigné comme un *nombre*, le coût comme une option (coût élevé pour l'option A, correspondant au coût légal ; faible pour l'option B, correspondant à l'évasion). Les firmes sont appelées *participants* et les marchés des *groupes*. Le terme de punition n'est jamais employé pour désigner l'amende, qualifiée d'*annulation des gains* assortie d'*une perte fixe*. Dans les deux derniers traitements, la dénonciation est désignée par sa manifestation physique en parlant de *cocher la case d'un participant* pour une firme dénonciatrice et d'*être cochée* pour une firme dénoncée. Le script informatique des expériences a été développé en utilisant le logiciel Regate (Zeiliger, 2000), en collaboration avec Romain Zeiliger.

Les gains obtenus par chaque participant sont calculés à partir de la somme des ECU accumulés pendant l'ensemble des périodes de l'expérience (à l'exception des trois périodes d'entraînement), selon le taux de conversion de 1 Euro pour 15 ECU. Ces gains sont versés de façon privée, à la fin de l'expérience. Le taux horaire de rémunération atteint une moyenne de 12 Euros. Au total, 5 sessions ont été conduites dans le laboratoire d'économie expérimentale du GATE. Ces sessions ont réuni un total de 76 participants, constitués d'étudiants inscrits en premier cycle à l'ITECH (Institut TExtile et Chimique de Lyon), à l'EM Lyon (Ecole de Management de Lyon) et à l'Ecole

[20] Dans le cadre des marchés expérimentaux, la légitimité de la contextualisation des instructions a suscité d'intenses débats méthodologiques. Bien que les travaux originels privilégient des instructions neutres (Plott, 1982), la complexité des institutions de marché étudiées a rapidement conduit à ancrer les instructions dans une réalité connue pour en faciliter la compréhension. De fait, le critère de la complexité semble l'avoir emporté (Holt, 1995, p.356). Quoique communément admise, cette contextualisation ne semble pas indifférente sur les comportements. Divers travaux comparent les comportements de marchés obtenus selon que les instructions font référence ou pas à une situation économique. Ces études concluent à une plus grande agressivité des comportements lorsque les instructions sont rédigées en termes neutres (Huck, Normann & Oechssler, 2004) ou contextualisées (Franciosi, Kujal, Michelitsch & *al.*, 1995), mais rejettent l'hypothèse que les comportements restent inchangés. La relative simplicité du marché que nous étudions comme l'intervention de comportements illégaux (voir Chapitre 1, Note (45) pour une discussion de cet aspect) militent, dans notre cas, en faveur de la neutralité des instructions.

Centrale de Lyon.

Après formation des groupes, ces participants ont permis de constituer 22 marchés expérimentaux. La répétition aléatoire du jeu fournit 1357 observations sur les décisions individuelles des firmes expérimentales (et 396 observations de marchés expérimentaux). Parmi ces observations, 27.5% (373 firmes et 110 marchés) concernent le traitement Contrôle, 33.5% (454 firmes et 132 marchés) le traitement Dénonciation et 39% (530 firmes et 154 marchés) le traitement Clémence.

Ces observations reflètent la réaction des firmes expérimentales, en termes d'évasion et de prix choisis, à des variations de l'environnement telles que le coût de la dénonciation ou la taille de l'industrie. Elles permettent donc d'apprécier la validité du modèle théorique par comparaison avec les réactions prédites.

3.4 Malédiction de Bertrand et évasion collusive : résultats empiriques

Comme nous l'avons souligné à cette occasion (Note (14)), le modèle a été présenté en considérant un équilibre particulier, dans lequel le prix est celui de l'équilibre concurrentiel légal. Dans le cadre des expériences, il n'est pas assuré que cet équilibre soit sélectionné par les firmes. Il convient donc de généraliser le modèle à l'ensemble des prix qui constituent un équilibre de silence collusif. Cette généralisation permet de définir les expressions utilisées pour mesurer les variables considérées dans le modèle théorique et d'en dériver les prédictions testables. Les comportements observés confirment la malédiction de Bertrand (Section 3.4.2), et valident les conditions d'émergence de l'évasion collusive identifiées par l'analyse théorique (Section 3.4.3).

3.4.1 Contrepartie empirique du cadre théorique

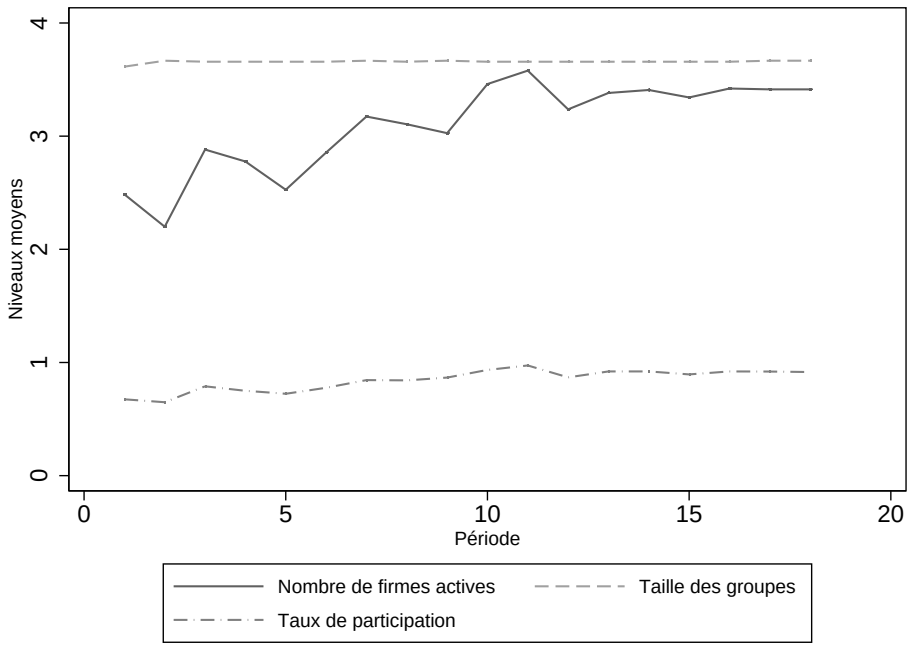
Avant de procéder à l'analyse statistique des comportements observés, cette section présente la transcription empirique de l'analyse théorique. La définition des variables utilisées permet en particulier de dériver les prédictions du modèle quant aux décisions des firmes expérimentales.

D'abord, la taille de l'industrie détermine les profits potentiels de la collusion et est à ce titre une variable centrale de notre analyse. Le déroulement de l'expérience crée cependant une ambiguïté quant à la mesure adéquate de la taille de l'industrie. Le protocole conduit en effet à une déconnection entre la taille *a priori* du marché (*i.e.* taille des groupes de participants formés au début de l'expérience) et le nombre de firmes actives sur le marché, correspondant au nombre de firmes qui ont choisit le prix minimum. Ces mesures offrent des appréciations différentes quand aux possibilités de collusion sur le marché.

D'une part, la taille du groupe fournit une appréciation *ex ante* de l'intensité de la concurrence, au sens où elle reflète le nombre potentiel de concurrentes . Le nombre de firmes actives en fournit d'autre part une mesure *ex post*, puisqu'elle détermine les quantités individuelles effectivement vendues par les firmes au prix de marché. La taille du groupe permet donc d'anticiper l'intensité potentielle de la concurrence tandis que le nombre de firmes actives mesure son intensité effective. L'objectif de l'analyse empirique étant de tester la capacité du modèle à expliquer les comportements observés, la mesure pertinente de la taille de l'industrie est celle que prennent en compte les participants pour apprécier l'intensité de la concurrence dans leurs décisions. A cet égard, le nombre de firmes actives semble être le candidat naturel.

Le Graphique 3.1 propose un aperçu de la pertinence de cet indicateur. Il représente en effet l'évolution du nombre de firmes actives moyen (intensité concurrentielle *ex post*) et du taux de participation moyen (mesuré à chaque période par le rapport

Graphique 3.1 – Evolution des mesures d'intensité



entre le nombre de firmes actives et la taille du groupe) au cours de l'expérience. Par construction, la taille des groupes (intensité concurrentielle *ex ante*), également représentée, est constante au cours de l'expérience. Le taux de participation n'est que très légèrement croissant au cours du temps. Pour une taille de groupe donnée, le nombre de firmes actives à une période tend donc à être très proche du niveau qu'elle a atteint à la période précédente. L'évolution du nombre actives reflète cette tendance, et suit un sentier de croissance régulière au cours du temps qui ne subit pas de sursaut notable.

Le nombre de firmes actives à une période apparaît ainsi comme le meilleur indicateur de l'intensité effective de la concurrence à la période suivante. La taille de l'industrie (variable n) est donc mesurée à chaque période par le nombre de firmes actives à la période précédente.[21] La distribution des tailles d'industrie empiriques (nombre de firmes

[21]Cette mesure nous oblige à abandonner la première observation de chaque firme. La première période de jeu du premier traitement (76 observations de firmes et 22 de marchés) est donc éliminée au même titre que les périodes d'essai. La robustesse des résultats à un changement de mesure est discutée

Tableau 3.2 – Distribution de l'intensité de la concurrence

	Intensité de la concurrence						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Contrôle	23.57	23.91	33.33	10.10	7.41	1.68	100.00
Dénonciation	9.69	22.25	40.97	14.10	9.03	3.96	100.00
Clémence	7.36	7.36	51.32	12.83	14.34	6.79	100.00
Total	11.94	16.47	43.48	12.65	10.85	4.61	100.00

Note. Pourcentage de firmes soumises à une intensité concurrentielle donnée (mesurée par le nombre de firmes actives à la période précédente) au sein de chaque traitement. En %.

actives sur les marchés expérimentaux) dans chaque traitement est présentée dans le Tableau 3.2.

Outre la taille de l'industrie, l'environnement économique considéré dans le modèle est également décrit par l'Hypothèse 3.1, qui circonscrit le profit d'évasion et la robustesse à la collusion tacite. Le profit d'évasion défini en (3.2) évalue les bénéfices tirés de l'évasion lorsque le prix de marché est celui de l'équilibre concurrentiel légal. Cette expression se généralise aisément à tout prix $p \neq p^c$:

$$\Pi_F(p) = (1-\alpha)\frac{Q(p)}{n}\ (p-w) - \alpha\ F = \pi_F(p) - \alpha\ F \tag{3.9}$$

Cette expression décrit le bénéfice de l'évasion pour tout prix de marché p. Le paramètrage de l'expérience (présenté en détail dans l'Annexe de ce chapitre, Section 3.B) a été conçu de façon à ce que ce profit soit positif aussi souvent que possible, afin de respecter l'Hypothèse 3.1.2. De fait, le profit d'évasion est positif tant que $p > w$ (correspondant au coût marginal en cas d'évasion, fixé à 5) pour tous les marchés considérés

en Annexe, Section 3.C. La Section 3.C.1 montre en particulier que les comportements observés tendent à confirmer la pertinence de la mesure par le nombre de firmes actives. Les résultats présentés ici sont, dans l'ensemble, maintenus en termes de statistiques descriptives, Section 3.4.2 (discutés dans la Section 3.C.2). Les difficultés d'identification du modèle sont à l'origine de changements notables dans les résultats qui concernent la mise en œuvre de l'évasion collusive, Section 3.4.3 (Section 3.C.3). Le pouvoir explicatif de la crédibilité de la menace de dénonciation est néanmoins confirmé.

dans l'expérience. En l'absence de dénonciation, cette condition doit théoriquement conduire les firmes à choisir l'évasion. Si, par surcroît, le marché considéré est robuste à la collusion tacite, la concurrence devrait en outre conduire à une baisse du prix jusqu'à annulation des profits.

Prédiction* 3.1. *(Malédiction de Bertrand) *En l'absence de dénonciation, les firmes choisissent d'assumer le risque de la détection en optant pour l'évasion fiscale. L'intensité de la concurrence tend à annuler les profits de l'évasion si le marché est robuste à la collusion tacite.*

L'Hypothèse 3.1.1 est, quand à elle, destinée à apprécier la capacité du marché à mettre un oeuvre un accord de collusion tacite. Elle établit plus précisément qu'un marché est robuste à la collusion tacite si $\gamma > \gamma^c \equiv \frac{1}{n}$. Cette valeur seuil est mesurée par l'inverse du nombre de firmes actives sur le marché.

L'équilibre de silence collusif (Section 3.2) a été étudié en considérant la stratégie de la Définition 3.1, où le prix collusif est celui de l'équilibre concurrentiel légal. Le raisonnement qui a conduit à ce résultat se généralise sans grande difficulté à tout prix de marché $p \neq p^c$. La condition qui assure la crédibilité de la menace (3.2) compare le profit de dénonciation au profit inter-temporel de la coopération. Si la coopération se fait au prix p plutôt qu'au prix p^c, le profit de coopération correspond alors au profit d'évasion (3.9). En utilisant cette définition de Π_F, la condition (3.2) permet donc d'évaluer la crédibilité de la menace pour tout prix de collusion p. Par définition, un prix de collusion p donné peut théoriquement constituer un équilibre de silence collusif si cette condition est vérifiée. La capacité du marché à mettre en oeuvre le silence collusif est par conséquent mesurée par la valeur prise par γ^F au prix de marché sélectionné : $\gamma^F = \dfrac{\Pi_F(p)}{F' + \Pi_F(p)}$. Lorsque la dénonciation est impossible (*i.e.* dans le traitement Contrôle), la capacité du marché à mettre en oeuvre une évasion collusive est décrite par (3.3). Cette expression – adaptée à l'équilibre sélectionné – est donc utilisée pour mesurer γ^F dans ce cas $\gamma^F = \dfrac{\pi_F(p) - \alpha\ F}{n\ \pi_F(p) - \alpha\ F}$. Compte tenu de ces définitions empiriques, l'intervalle de silence collusif est alors mesuré par $R = \gamma^F - \gamma^c$.

Tableau 3.3 – Possibilités de collusion dans chaque traitement

		Moyenne	Ecart-Type	Minimum	Maximum
Contrôle	$\gamma - \gamma^c$	-0.259	0.288	-0.750	0.083
	$\gamma_F - \gamma$	0.225	0.304	-0.250	0.750
	$R = \gamma_F - \gamma^c$	-0.034	0.109	-1.000	0.000
Dénonciation	$\gamma - \gamma^c$	-0.155	0.219	-0.750	0.083
	$\gamma_F - \gamma$	0.073	0.141	-0.250	0.575
	$R = \gamma_F - \gamma^c$	-0.081	0.129	-1.000	0.272
Clémence	$\gamma - \gamma^c$	-0.104	0.199	-0.750	0.083
	$\gamma_F - \gamma$	0.204	0.162	-0.250	0.663
	$R = \gamma_F - \gamma^c$	0.100	0.123	-0.357	0.442
Total	$\gamma - \gamma^c$	-0.158	0.237	-0.750	0.083
	$\gamma_F - \gamma$	0.162	0.209	-0.250	0.750
	$R = \gamma_F - \gamma^c$	0.005	0.147	-1.000	0.442

Note. Statistiques descriptives (moyenne et écart-types entre les firmes, minimum et maximum) des mesures de possibilités de collusion au sein de chaque traitement : collusion tacite (première ligne dans chaque traitement), silence collusif (deuxième ligne) et intervalle de silence collusif (troisième ligne).

La propension à la collusion des marchés expérimentaux, mesurée par ces indicateurs, est résumée dans le Tableau 3.3. Par définition, la robustesse à la collusion tacite (première ligne pour chaque traitement) ne dépend que de la taille de l'industrie. Les variations entre traitements reflètent donc uniquement les variations dans le nombre de firmes actives à la période précédente. La mise en oeuvre du silence collusif (deuxième ligne) est, quant à elle, d'autant plus facile que la dénonciation est peu coûteuse. L'intervalle de silence collusif (troisième ligne) reflète la compatibilité entre ces deux conditions. Comme le prévoit le modèle théorique, l'intervalle de silence collusif tend à s'élargir (le maximum devient positif, la moyenne augmente) à mesure que la dénonciation est facilitée.

En vertu du modèle théorique, ces conditions – robustesse à la collusion tacite et crédibilité de la dénonciation – décrivent l'ensemble des possibilités de collusion dans l'expérience. D'abord, si $\gamma - \gamma^c < 0$ la capacité du marché à mettre en oeuvre un accord

de collusion tacite s'applique à tout profit positif, qu'il contienne ou non un bénéfice lié à l'évasion. Lorsqu'il n'est pas robuste à la collusion tacite, le marché est donc trivialement capable de dégager des profits positifs grâce à l'évasion (Section 3.1.1). A l'inverse, le bénéfice de la déviation domine celui de la coopération si $\gamma - \gamma^c > 0$ et le marché devrait se trouver dans l'impossibilité de mettre en oeuvre un accord de collusion tacite. La dénonciation élargit cependant les possibilités de collusion. Si la dénonciation est une menace crédible ($\gamma^F - \gamma > 0$), le marché peut mettre en oeuvre l'équilibre de silence collusif et dégager des profits positifs grâce à l'évasion. Enfin, si le marché est robuste à la collusion tacite et que la menace de dénonciation n'est pas crédible, l'environnement qui prévaut conduit à la malédiction de Bertrand : l'évasion, choisie en raison de sa rentabilité, conduit à une nouvelle guerre des prix qui élimine les profits d'évasion.

Prédiction 3.2. (Evasion collusive) *Dans tous les traitements, la rentabilité de l'évasion conduit les firmes à l'adopter. Cette évasion permet de dégager des profits positifs dés lors que le marché n'est pas robuste à la collusion tacite ($\gamma > \gamma^c$) ou que la dénonciation est une menace crédible ($\gamma < \gamma^F$).*

Grâce à la combinaison des traitements et tailles de groupe instituée dans l'expérience, les marchés que nous observons présentent une importante variabilité dans les variables d'intérêt, que sont : la taille de l'industrie, la robustesse à la collusion tacite et la crédibilité de la menace de dénonciation. La réaction des firmes à ces variations permettent de tester les Prédictions 3.1 et 3.2, qui résument l'essentiel de l'analyse théorique proposée ci-dessus. Leur validité permet donc d'apprécier la capacité du modèle à expliquer les comportements observés.

Tableau 3.4 – Taux d'évasion sous le traitement Contrôle

	Intensité de la concurrence						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Evasion	95.04	92.21	97.17	97.62	100.00	90.91	95.43
Coût légal	4.96	7.79	2.83	2.38	0.00	9.09	4.57
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Note. Pour chaque intensité de la concurrence, pourcentage de firmes ayant choisi le coût minimum dans le traitement Contrôle. En %.

3.4.2 Comportements observés

Les trois traitements considérés dans l'expérience introduisent progressivement la dénonciation. En son absence (traitement Contrôle), les comportements prédits par le modèle correspondent à la malédiction de Bertrand. L'évasion collusive, quant à elle, correspond à une stratégie de tarification et d'évasion dont le modèle théorique a identifié plusieurs déterminants (robustesse à la collusion tacite et silence collusif). Avant d'étudier la pertinence empirique de ces déterminants (Section 3.4.3), nous proposons ici un premier aperçu de l'ampleur de l'évasion collusive dans les expériences.

a) Traitement de contrôle : malédiction de Bertrand[22]

La malédiction de Bertrand, résumée dans la Prédiction 3.1, prédit que l'intensité de la concurrence devrait empêcher les firmes de choisir un coût légal, et les contraindre ainsi à l'évasion. Comme le montre le Tableau 3.4 cette prédiction est largement cor-

[22]Dans cette section, et en particulier dans les Tableaux 3.4 et 3.5, 4 sessions pendant lesquelles 10 marchés sont observés s'ajoutent aux expériences présentées ci-dessus. Ces sessions ont été conduites dans les conditions décrites dans la Section 3.3, mais sous le traitement Contrôle uniquement. Ces données ne sont donc plus utilisées dans la suite de la présentation. Compte tenu de la répétition aléatoire du jeu, l'ensemble des sessions considérées dans ces tableaux fournissent 528 observations du comportement des firmes et 150 observations de marchés expérimentaux.

Tableau 3.5 – Distribution des prix de marché sous le traitement Contrôle

Prix de Marché	Intensité de la concurrence 1	2	3	4	5	6	Total
5	25.71	8.00	5.71	0.00	14.29	20.00	12.71
6	74.29	88.00	94.29	100.00	85.71	80.00	86.44
7	0.00	4.00	0.00	0.00	0.00	0.00	0.85
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Note. Pourcentage de marchés sur lesquels le prix minimum correspond à celui qui est indiqué en ligne, pour une intensité de la concurrence donnée, dans le traitement Contrôle. En %.

roborée par les comportements observés dans le traitement Contrôle, où l'évasion représente plus de 95% des décisions observées.

Cette observation est encore renforcée si l'on se limite au comportement des firmes actives. Parmi les 32 marchés observés, aucune firme ne parvient en effet à être active en choisissant le coût légal. Le Tableau 3.5 fait en outre apparaître la distribution des prix de marché en fonction de l'intensité de la concurrence. Il convient de rappeler que le paramètrage fait en sorte qu'un prix de 6 est le prix minimum garantissant des profits d'évasion non négatifs (voir Section 3.4.1). Un prix de 5 correspond en effet à une tarification au coût marginal (illégal), pour laquelle les profits sont en conséquence négatifs et égaux à l'amende espérée. Malgré cette propriété, l'intensité de la concurrence fait diminuer le prix jusqu'au coût marginal dans près de 16% des cas. Dans la très grande majorité des cas (83%), le prix butte sur sa limite inférieure garantissant des profits non-négatifs. Une infime minorité des firmes parviennent, enfin, à maintenir un prix collusif, immédiatement supérieur à 6. Cette situation ne concerne que 3 (2%) des 150 marchés observés et n'apparaît que lorsque le marché est en duopole.

Au total, le comportement observé sous le traitement de contrôle confirme donc tant la généralisation de l'évasion que la tendance à l'annulation des profits en l'absence de dénonciation.

Tableau 3.6 – Taux d'évasion

	Intensité de la concurrence						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Contrôle	94.29	91.55	96.97	100.00	100.00	80.00	95.29
Dénonciation	88.64	93.07	96.77	95.31	97.56	100.00	95.15
Clémence	74.36	89.74	97.79	98.53	100.00	97.22	95.85
Total	87.58	91.94	97.31	97.53	99.28	96.61	95.47

Note. Pour chaque intensité de la concurrence, pourcentage de firmes ayant choisi l'évasion dans chaque traitement (trois premières lignes) et dans l'ensemble (dernière ligne). En %.

Observation 3.1. *L'intensité de la concurrence contraint les firmes à choisir le travail au noir et élimine les profits d'évasion.*

b) Statistiques descriptives

Cette généralisation de l'évasion est une propriété assez largement partagée par les comportements sous l'ensemble des traitements de l'expérience. Le Tableau 3.6 présente les taux d'évasion (*i.e.* pourcentage d'observations ayant choisi le coût illégal) dans chaque traitement en fonction de l'intensité de la concurrence. Tous traitements confondus, l'évasion est choisie dans plus de 95% des cas (4.43% des observations choisissent le coût légal). A l'exception notable des marchés où l'intensité concurrentielle est de 6, l'évasion apparaît en outre d'autant plus fréquente que l'intensité de la concurrence est importante. Le taux moyen d'évasion croît ainsi régulièrement avec l'intensité de la concurrence, passant de près de 88% sur les marchés en monopole à 99.3% sur les marchés où l'intensité est de 5 (et 96.6% lorsqu'elle est de 6). Entre les traitements, l'introduction progressive de la dénonciation tend, en moyenne, à accroître légèrement le taux d'évasion. Cet effet est cependant très ambigu lorsque l'évolution du taux d'évasion est désagrégée en fonction de l'intensité de la concurrence.

Cette diversité, entre traitements comme en fonction de l'intensité de la concurrence,

se retrouve en partie dans les comportements de tarification (présenté dans le Tableau 3.7). Le prix individuel posté par les firmes tend, en moyenne, à diminuer lorsque l'intensité de la concurrence augmente et à s'accroître lorsque la dénonciation est facilitée. A traitement donné, le prix est d'autant plus faible que l'intensité de la concurrence est forte. Au regard de l'effet de la dénonciation sur le prix choisi, en revanche, deux sous-ensembles se distinguent selon que le nombre de firmes est inférieur ou supérieur à 3. Dans le premier cas (concurrence entre 3 firmes ou moins), le prix choisi est d'autant plus élevé en moyenne, à intensité de la concurrence donnée, que la dénonciation est aisée. La dénonciation tend donc à contre-carrer la baisse du prix engendrée par l'intensité de la concurrence. Dans le second cas (4 firmes ou plus), il semble au contraire que l'effet de la concurrence l'emporte, et le prix ne subit que de légères variations à la baisse lorsque la dénonciation est facilitée.

Les prix qui émergent à l'équilibre (présentés dans le Tableau 3.8 et représentés dans le Graphique 3.2) rendent plus univoques ces tendances. A l'exception du traitement de contrôle – où le prix minimum moyen fluctue autour de 6 sans lien apparent avec l'intensité de la concurrence – le prix de marché est en effet décroissant de l'intensité de la concurrence à traitement donné.

A mesure que la dénonciation est rendue possible, et facilitée, le prix minimum s'accroît pour chaque intensité concurrentielle. Le Graphique 3.2 propose une représen-

Tableau 3.7 – Prix choisi moyen

	Intensité de la concurrence						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Contrôle	6.57	6.77	6.28	6.20	6.00	7.00	6.45
Dénonciation	7.09	6.65	6.22	6.25	6.12	5.94	6.38
Clémence	9.49	7.49	6.47	6.15	6.00	6.11	6.63
Total	7.46	6.85	6.35	6.20	6.04	6.14	6.50

Note. Pour chaque intensité de la concurrence, prix choisi en moyenne par les firmes dans chaque traitement (trois premières lignes) et dans l'ensemble (dernière ligne).

Tableau 3.8 – Prix d'équilibre moyen

	Intensité de la concurrence						
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Contrôle	5.89	6.00	5.94	6.00	5.75	5.00	5.93
Dénonciation	6.50	6.12	6.05	6.00	6.00	5.67	6.09
Clémence	8.43	6.69	6.25	6.00	6.00	6.00	6.42
Total	6.84	6.19	6.13	6.00	5.96	5.80	6.19

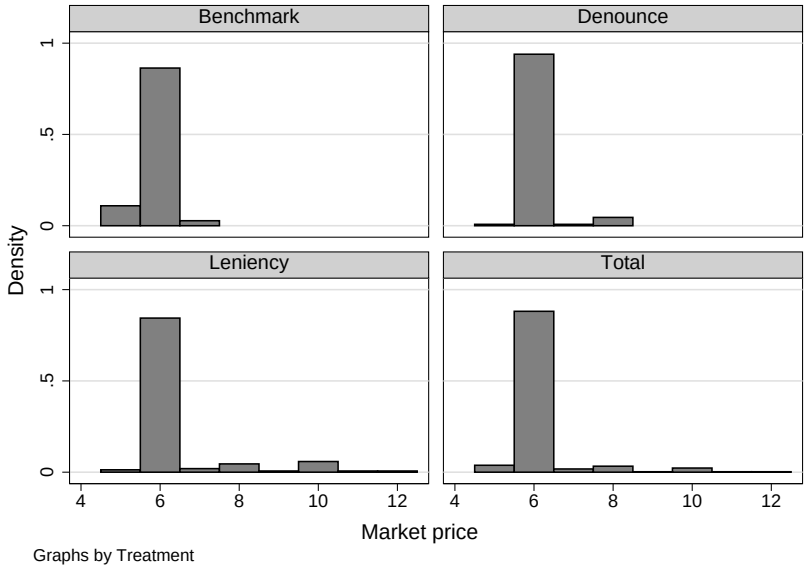
Note. Pour chaque intensité de la concurrence, prix minimum moyen sur les marchés dans chaque traitement (trois premières lignes) et dans l'ensemble (dernière ligne).

tation de la fonction de densité du prix d'équilibre en fonction des traitements. Bien que le prix minimum de profit non nul (égal à 6) reste très largement modal dans tous les traitements, la distribution tend à s'étaler vers la droite à mesure que la dénonciation est facilitée.[23]

Comme le montre le Tableau 3.9, la dénonciation est pourtant peu utilisée dans les expériences. Ainsi, dans l'ensemble des deux traitements où elle est autorisée, la dénonciation n'est utilisée que par 4.88% des observations. En raison des dénonciations multiples, 8.13% des participants se voient infliger une amende suite à ces dénonciations. C'est donc plus de la menace qu'elle constitue que de son pouvoir de sanction que la dénonciation tire son efficacité. L'usage qui en est fait reflète la stratégie de silence collusif. La dénonciation vise en effet très majoritairement (plus de 78% des dénonciations) des participants qui choisissent le prix minimum de profits non négatifs. Ce comportement de tarification est considéré par les dénonciateurs comme une déviation par rapport au prix collusif, correspondant au prix qu'eux-même choisissent, supérieur

[23]Ces variations ne semblent pas devoir être imputées à un effet d'apprentissage au cours de l'expérience. Le taux d'évasion, le prix choisi et le prix de marché sont en effet remarquablement stables en moyenne entre les périodes. Leurs écart-types entre les périodes de jeu sont ainsi de 0.217 pour le premier (pour un taux d'évasion moyen de 0.954), 0.169 pour le deuxième (pour un prix choisi moyen de 6.50) et 0.220 pour le troisième (pour un prix de marché moyen de 6.19). La représentation graphique du profil d'évolution de ces variables est fournie en Annexe, Graphique 3.B.

Graphique 3.2 – Prix d'équilibre par traitement



à 6 pour 75% des dénonciateurs.

Au total, la réaction des comportements d'évasion et de tarification au variations de l'environnement semble assez cohérente avec l'analyse théorique. D'une part, la rentabilité de l'évasion porte une écrasante majorité de firmes à s'y livrer. Les profits qui en sont tirés – par l'intermédiaire du prix – sont, d'autre part, la résultante de deux forces opposées : toutes choses égales par ailleurs, l'intensité de la concurrence tend à abaisser le prix choisi par les firmes comme le prix d'équilibre sur le marché ; la possibilité de

Tableau 3.9 – Comportement de dénonciation

	Prix choisi									
Participant	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**Total**	**Ensemble**
Dénonciateur	25.00	12.50	4.17	2.08	39.58	4.17	10.42	2.08	100	4.88
Dénoncé	78.75	2.50	2.50	1.25	8.75	2.50	3.75	0.00	100	8.13

Note. *Moitié supérieure* : Pourcentage de firmes, parmi les firmes qui ont choisi de dénoncer un ou plusieurs autres participants, ayant choisi le prix indiqué en colonne. En %. *Moitié inférieure* : Pourcentage de firmes, parmi les firmes qui ont été dénoncées par au moins un autre participant, ayant choisi le prix indiqué en colonne. En %. *Dernière colonne* : Pourcentage de firmes ayant choisi de dénoncer un ou plusieurs autres participants, ou ayant été dénoncées par au moins au autre participant, parmi l'ensemble des observations.

dénoncer l'évasion s'oppose à ce premier effet et encourage des hausses de prix à travers l'usage d'une stratégie de silence collusif. La faiblesse de la concurrence et la menace de dénonciation tendent ainsi à encourager l'évasion collusive.

Ces deux effets sont pris en compte dans le modèle par les variables de robustesse à la collusion tacite et de crédibilité de la menace de dénonciation. Comme le résume la Prédiction 3.2, le modèle théorique prédit en effet que, conditionnellement à l'évasion, le prix choisi peut être collusif si le marché n'est pas robuste à la collusion tacite ($\gamma < \gamma^c$) ou lorsque la dénonciation de l'évasion est une menace crédible ($\gamma < \gamma^F$). Quand seule la seconde condition est vérifiée, la dénonciation permet au marché de mettre en oeuvre un prix de collusion, fondé sur l'évasion, malgré sa robustesse à la collusion tacite. L'intervalle de silence collusif mesure cet élargissement de l'évantail des accords de collusion pouvant être mis en œuvre.

Chacune de ces deux caractéristiques permet au marché de mettre en œuvre une évasion collusive, par laquelle maintenir un prix de collusion permet de dégager des profits positifs de l'évasion. Le recours à cette stratégie est mesuré par une variable indicatrice, EC (*Evasion Collusive*), prenant la valeur 1 lorsqu'une firme choisit simultanément l'évasion et un prix collusif. Pour construire cette variable, un prix est considéré comme collusif s'il est supérieur au prix minimum garantissant des profits non-nuls, égal à 6 dans l'expérience (voir (3.10) ci-dessous pour une définition formelle).

Le Tableau 3.10 propose un premier aperçu de la performance des variables identifiées dans le modèle comme pertinentes pour expliquer le recours au silence collusif. Chaque cellule du tableau correspond à la proportion d'observations pour lesquelles $EC = 1$, en fonction du degré de robustesse à la collusion tacite et de crédibilité de la menace, telles que décrites dans le Tableau 3.3. Le taux d'évasion collusive apparaît d'abord fortement décroissant de la robustesse du marché à la collusion tacite, passant de 15% à 7% lorsque le marché devient robuste. La crédibilité de la menace de dénonciation – et par conséquent l'intervalle de silence collusif – tend elle aussi, ensuite, à

Tableau 3.10 – Evasion collusive observée

	Silence Collusif				
Collusion	Crédibilité		Intervalle		**Total**
Tacite	$\gamma > \gamma^F$	$\gamma < \gamma^F$	$R < 0$	$R > 0$	
$\gamma < \gamma^c$	4.6	15.7	11.2	20.3	14.6
$\gamma > \gamma^c$	1.2	10.1	1.2	10.1	7.3
Total	3.1	14.8	10.1	17.0	13.1

Note. Pourcentage d'observations pour lesquelles la variable *EC* est égale à 1 (évasion et prix choisi supérieur à 6). En %. *En ligne* : Robustesse à la collusion tacite ; *Colonne de gauche* : Crédibilité de la menace de dénonciation ; *Colonne de droite* : Intervalle de silence collusif.

encourager l'évasion collusive qui s'élève à moins de 3% lorsque la dénonciation n'est pas une menace crédible à près de 15% lorsqu'elle le devient. Ces deux effets tendent, enfin, à se renforcer mutuellement, le taux d'évasion collusive étant d'autant plus élevé lorsque le marché est robuste à la collusion tacite (respectivement lorsque la menace de dénonciation est crédible) que la dénonciation est par surcroit une menace crédible (resp. que le marché est non robuste à la collusuion tacite).

Le comportement observé dans les expériences semble être conforme aux attentes, en termes d'évasion fiscale (rentable en toutes circonstances tant que le prix reste supérieur ou égal à 6) comme de facteurs facilitant l'évasion collusive. La prochaine section propose une analyse formelle du lien entre ces tendances et l'analyse théorique.

3.4.3 Conditions de mise en œuvre de l'évasion collusive

Lorsque l'évasion collusive constitue un équilibre du marché, la multiplicité des prix collusifs qui peuvent lui être associés laisse aux marchés expérimentaux le libre choix du prix sélectionné. Si le modèle prédit les conditions favorisant l'émergence de l'évasion collusive, le choix du prix qui accompagne l'évasion dans cette stratégie est donc essentiellement un problème empirique, qui découle de la coordination entre les

firmes. En conséquence, le modèle économétrique est spécifié selon deux dimensions, qui distinguent l'émergence de l'évasion et la sélection du prix collusif. L'évasion collusive est modélisée comme un choix binaire, conditionnellement auquel les firmes choisissent le niveau du prix associé.

a) Emergence de l'évasion collusive

Le recours à une stratégie d'évasion collusive est mesuré par la variable binaire EC, présentée ci-dessus (Section 3.4.2), décrivant les décisions de la firme i à la période t selon :

$$EC_i^t = \begin{cases} 1 & \text{si } \{p_i^t > 6; W_i^t = w\} \\ 0 & \text{sinon} \end{cases} \tag{3.10}$$

En notant $I[C]$ la variable indicatrice de la validité de la condition C, le modèle théorique prédit que la probabilité de recourir à l'évasion collusive est décroissante de $I[\gamma > \gamma^c]$ (robustesse à la collusion tacite) et croissante de $I[\gamma^F > \gamma]$ (crédibilité de la menace de dénonciation). La propension à utiliser la stratégie d'évasion collusive est modélisée comme une variable latente, EC_i^{t*}, déterminée par ces conditions théoriques et un certain nombre de caractéristiques observables (période de jeu, sexe, âge,...) réunies dans le vecteur X. La propension à recourir au silence collusif est alors expliquée par l'équation latente : $EC_i^{t*} = \beta_0 + \beta_c I[\gamma > \gamma^c] + \beta_F I[\gamma^F > \gamma] + \delta_{EC}\ X_{i,t}$. Comme le résume la Prédiction 3.2, le modèle théorique se traduit dans cette équation par l'espace de paramètres : $\{\beta_c < 0; \beta_F > 0\}$.

Les paramètres sont estimés en s'appuyant sur le comportement d'évasion collusive

observé, selon la relation :

$$\begin{aligned} EC_i^t &= \begin{cases} 1 & \text{si } EC_i^{t*} \geq 0 \\ 0 & \text{sinon} \end{cases} \\ EC_i^{t*} &= \beta_0 + \beta_c I[\gamma > \gamma^c] + \beta_F I[\gamma^F > \gamma] + \delta_{EC} X_{i,t} + \epsilon_{i,t} \end{aligned} \tag{3.11}$$

Si le terme d'erreur du modèle, $\epsilon_{i,t}$, est supposé i.i.d. entre les observations et de loi normale centrée réduite, l'équation (3.11) définit un Probit dichotomique. L'hétérogénéité inobservable est cependant incorporée en considérant un modèle à erreurs composées :

$$\epsilon_i^t = u_i + \omega_{i,t} \equiv N(\mathbf{0}, \Sigma) \ , \ \Sigma = \begin{pmatrix} \sigma_u & \rho \\ \rho & 1 \end{pmatrix} \tag{3.12}$$

Dans cette expression, le terme $\omega_{i,t}$ est un terme d'erreur i.i.d de loi normale, supposé centré et réduit pour assurer l'identification du modèle. Le terme aléatoire u_i, également de loi normale, représente la distribution de l'hétérogénéité inobservable entre les individus (Hausman & Taylor, 1981). Outre les caractéristiques observables contenues dans le vecteur X, des variables de contrôle sont également incluses pour tenir compte de la dynamique propre aux marchés.[24] Le modèle est estimé par la méthode du maximum de vraisemblance en information complète.

Les variables incluses dans le vecteur X reflètent les caractéristiques individuelles des participants (*Age* ; variable indicatrice de *Sexe*, valant 1 pour un homme ; *Education* post-bac en années), les caractéristiques du marché auquel ils appartiennent (*Taille du groupe*) ainsi que des variables qui reflètent la dynamique de l'interaction. L'effet du

[24] L'inclusion d'effets fixes dans les modèles non-linéaires peut poser d'importants problèmes à la fois pratiques (nombre de paramètres estimés) et théoriques (convergence des estimateurs) en fonction de la taille de la dimension individuelle et temporelle des données (Greene, 2004). Le premier problème est très limité dans notre cas, où seuls 22 effets fixes de marché sont à estimer. La dimension temporelle (18 périodes) excède en outre la limite inférieure (8) généralement admise comme suffisante à rendre négligeable le biais éventuel (Heckman, 1981).

temps est ainsi pris en compte par deux variables, égales au numéro de la période au sein d'un traitement (*Round*) et dans l'ensemble de l'expérience (*Période*). L'usage effectif de la dénonciation est incorporé en incluant des variables indiquant qu'un participant a été dénoncé (*Participant dénoncé*) ou dénonciateur (*Participant dénonciateur*) à la période précédente.

Les résultats de l'estimation du probit à effets individuels aléatoires définit en (3.11) et (3.12) sont présentés dans le Tableau 3.11. Comme le laissait présager l'analyse descriptive présentée ci-dessus, la robustesse à la collusion tacite comme la crédibilité de la menace de dénonciation influencent significativement la probabilité de recourir à l'évasion collusive : les paramètres estimés correspondants (respectivement $\hat{\beta}_c$ et $\hat{\beta}_F$) sont significatifs et ont le signe attendu. Lorsqu'elle est utilisée, la dénonciation conduit les firmes qui en sont la cible (*Participant dénoncé*) à adopter la stratégie d'évasion collusive. Cet effet est anticipé par les participants, puisqu'un participant dénonciateur a une forte probabilité d'adopter l'évasion collusive suite à la mise en œuvre de la menace.

Observation 3.2. *Un marché peut mettre en œuvre l'évasion collusive d'autant plus facilement qu'il n'est pas robuste à la collusion tacite et que la dénonciation est une menace crédible.*

Cette première spécification apporte une confirmation supplémentaire du rôle joué par les mécanismes identifiés dans le modèle. Elle laisse cependant inexpliquée la coordination qui conduit à sélectionner le prix incorporé par les firmes dans leur stratégie d'évasion collusive.

Tableau 3.11 – Evasion collusive

Variable	Coefficient	(Ecart-type)
Probabilité d'évasion collusive (Probit, variable endogène : EC)		
$I[\gamma > \gamma^c]$	-1.222***	(0.419)
$I[\gamma^F > \gamma]$	0.494*	(0.281)
Participant dénonciateur	0.328***	(0.113)
Participant dénoncé	0.321*	(0.194)
Age	0.179	(0.146)
Sexe	0.353	(0.235)
Education	-0.264	(0.258)
Période	-0.005	(0.015)
Round	-0.067*	(0.040)
Taille du groupe	-0.004	(0.217)
Constante	-4.310	(2.746)
Contrôles fixes Marchés	*oui*	
Distributions Estimées		
$\hat{\sigma}$	0.433	(0.105)
$\hat{\rho}$	0.158***	(0.0643)

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Probit à effets individuels aléatoires. La variable endogène (EC) vaut 1 lorsqu'une observation a choisit l'évasion collusive (évasion et prix choisi supérieur à 6). L'hétérogénéité inobservable individuelle est incorporée par un effet aléatoire, propre aux participants. La variable $I[\gamma > \gamma^c]$ vaut 1 si le marché est robuste à la collusion tacite, la variable $I[\gamma^F > \gamma]$ indique que la menace de dénonciation est crédible. *Participant dénonciateur* indique que l'observation a dénoncé au moins un autre participant à la période précédente, *Participant dénoncé* qu'elle a été dénoncée par au moins un autre participant à la période précédente. L'*Age* est mesuré en années ; la variable *Sexe* indique que le participant est un homme ; la variable *Education* mesure le nombre d'années d'études après bac. La variable *Période* mesure le passage du temps dans l'ensemble de l'expérience ; la variable *Round* le passage du temps dans chaque traitement (réinitialisée en début de traitement). La *Taille du groupe* est pour chaque participant, égal à la taille de son groupe d'appartenance dans l'expérience.

b) Evasion collusive et coordination

La stratégie d'évasion collusive consiste à associer un prix collusif à l'évasion. Le niveau du prix choisi par une firme dans cette stratégie est donc par définition borné inférieurement par la valeur seuil qui permet de qualifier un prix de collusif (fixé à 6 dans le modèle économétrique). Le prix choisi conditionnellement à l'évasion collusive p_i^t est donc à son tour modélisé comme une variable latente, notée p_i^{t*}, qui n'est observée que lorsque la firme recourt au silence collusif. La robustesse à la collusion tacite et la menace de dénonciation sont théoriquement neutres sur ce choix. Les variables correspondantes (respectivement $I[\gamma^F > \gamma]$ et $I[\gamma^F > \gamma]$) sont cependant incluses dans l'équation latente à titre de contrôle. S'y ajoutent des variables observables décrivant les caractéristiques des firmes, du marché et de l'environnement, regroupées dans la matrice Z_i^t. L'équation qui modélise le choix du niveau du prix par les firmes s'écrit donc :

$$\begin{aligned} p_i^t &= \begin{cases} p_i^{t*} & \text{si } EC_i^{t*} = 1 \\ 0 & \text{sinon} \end{cases} \\ p_i^{t*} &= \mu_0 + \mu_c\ I[\gamma^F > \gamma] + \mu_F\ I[\gamma^F > \gamma] + \delta_p\ Z_{i,t} + v_{i,t} \end{aligned} \tag{3.13}$$

où $v_{i,t}$ est une variable aléatoire de loi normale centrée et d'écart-type σ. Si des facteurs inobservables expliquent simultanément la propension au silence collusif et la coordination sur le niveau du prix, les termes d'erreur des modèles (3.11) et (3.13) sont corrélés. On suppose donc que les termes $v_{i,t}$ et $\epsilon_{i,t}$ suivent une loi normale bivariée : $N(\mathbf{0}, \Omega)$, $\Omega = \begin{pmatrix} \sigma & \rho_p \\ \rho_p & 1 \end{pmatrix}$ où ρ_p mesure la corrélation entre l'équation de sélection (3.11) et l'équation d'intensité (3.13).

En toute généralité, les équations (3.11) et (3.13) définissent un Tobit type II (Amemiya, 1984, initialement développé par Heckman, 1978) qui permet de distinguer les déterminants du processus de sélection (choix de l'évasion collusive), contenus dans X, et les facteurs qui influencent le choix d'intensité conditionnellement à la participation

(prix collusif) notés Z. Le modèle théorique reste cependant silencieux sur le processus de coordination lui-même. L'ensemble des variables considérées dans l'estimation de la section précédente, X, sont donc *a priori* candidates à être incluses dans Z. Dans ce cas, toutes les variables sont communes aux deux équations. C'est alors la non-linéarité du modèle qui permet d'en identifier les paramètres. Bien que valide, la non linéarité produit une identification faible en l'absence de restrictions d'exclusion (Honoré, Vella & Verbeek, 2005). A cet égard la variable *Period*, mesurant le passage du temps dans l'ensemble de l'expérience, paraît être le choix d'exclusion le plus naturel. Alors qu'elle semble n'apporter aucune information sur la propension à utiliser une stratégie d'évasion collusive (Tableau 3.11), cette variable décrit en effet le rythme de répétition du jeu, connu pour son effet important sur la capacité des joueurs à se coordonner (Berninghaus & Ehrhart, 1998). Cette variable est donc incluse uniquement dans les explicatives Z afin de faciliter l'identification du modèle.

Les résultats issus de différentes spécifications du modèle sont présentés dans le Tableau 3.12. Le modèle est estimé à partir des données produites par les expériences, comprenant 22 marchés expérimentaux composés dans leur ensemble de 72 individus. Cette taille d'échantillon relativement faible contraint de façon importante l'identification empirique du modèle. En particulier, le modèle n'est empiriquement identifié qu'à condition d'exclure l'hétérogénéité inobservable – entre individus et entre marchés – de l'équation de sélection. Cet aspect a cependant été intégré dans la section précédente. Au regard de la comparaison entre les Tableaux 3.11 et 3.12, les paramètres estimés semblent assez peu sensibles à cette exclusion de l'hétérogénéité inobservable.

Les paramètres de l'équation consacrée à l'évasion collusive (moitié supérieure du Tableau 3.12) sont en effet très similaires, tant en termes de signes que de significativité, à ceux du Probit commentés plus haut. Négliger l'hétérogénéité inobservable reporte cependant une partie de la variation de l'endogène sur les caractéristiques observables des individus, telles que le sexe.

Tableau 3.12 – Evasion collusive et coordination

	Coefficient	t	Coefficient	t	Coefficient	t
Probabilité d'évasion collusive (Probit, variable endogène : EC)						
$I[\gamma > \gamma^c]$	-1.136**	-2.44	-1.132**	-2.45	-1.133**	-2.45
$I[\gamma^F > \gamma]$	0.410*	1.88	0.410*	1.88	0.410*	1.88
Participant dénonciateur	0.458***	4.66	0.456***	4.70	0.456***	4.70
Participant dénoncé	0.560***	3.60	0.560***	3.60	0.560***	3.60
Age	-0.050	-0.79	-0.050	-0.78	-0.050	-0.78
Sexe	0.573***	5.52	0.572***	5.51	0.572***	5.51
Education	0.097	0.80	0.097	0.79	0.097	0.79
Round	-0.051*	-1.79	-0.050*	-1.78	-.050*	-1.78
Taille du groupe	-0.113	-1.14	-0.113	-1.14	-0.113	-1.14
Constante	-0.584	-0.52	-0.582	-0.52	-0.582	-0.52
Sélection de l'équilibre (Tobit, variable endogène : p)						
$I[\gamma > \gamma^c]$	-0.673*	-1.66	-0.610	-1.12	-0.610	-1.12
$I[\gamma^F > \gamma]$	0.161	0.35	0.353	0.61	0.353	0.61
Participant dénonciateur	0.211	0.91	0.142	0.95	0.142	0.95
Participant dénoncé	0.901***	3.34	0.561***	2.89	0.561***	2.89
Age	0.266*	1.85	0.140	0.28	0.134	1.02
Sexe	-0.338	-1.08	-0.041	-0.16	0.098	0.39
Education	-0.762***	-3.01	-0.206	-0.28	-0.304	-1.22
Période	0.067*	1.74	0.064	1.34	0.064	1.34
Round	0.078	1.00	0.033	0.45	0.033	0.45
Taille du groupe	-0.128	-0.63	0.353	1.13	0.701***	2.72
Constante	5.997***	2.62	3.038	0.35	2.373	1.07
Contrôles Individuels	-	-	oui		oui	
Contrôles Marchés	-	-	-	-	oui	
			Distributions estimées			
$\hat{\sigma}$	1.535	-	1.192	-	1.206	-
$\hat{\rho}$	-0.193**	3.85[a]	-0.030	0.300	-0.018	0.06

Niveaux de signification : *** 10%, ** 5%, *** 1%.
[a] Test de Wald d'indépendance des équations estimées.

Note. Tobit Type II. *Moitié supérieure* : Variable endogène (EC) valant 1 lorsqu'une observation a choisit l'évasion collusive (évasion et prix choisi supérieur à 6). *Moitié inférieure* : Variable endogène égale au prix choisi, conditionnellement à $EC = 1$. La variable $I[\gamma > \gamma^c]$ vaut 1 si le marché est robuste à la collusion tacite, la variable $I[\gamma^F > \gamma]$ indique que la menace de dénonciation est crédible. *Participant dénonciateur* indique que l'observation a dénoncé au moins un autre participant à la période précédente, *Participant dénoncé* qu'elle a été dénoncée par au moins un autre participant à la période précédente. L'*Age* est mesuré en années ; la variable *Sexe* indique que le participant est un homme ; la variable *Education* mesure le nombre d'années d'études après bac. La variable *Période* mesure le passage du temps dans l'ensemble de l'expérience ; la variable *Round* le passage du temps dans chaque traitement (réinitialisée en début de traitement). La *Taille du groupe* est égale, pour chaque participant, à la taille de son groupe d'appartenance dans l'expérience.

La moitié inférieure du tableau présente les paramètres estimés de l'équation de prix. L'introduction progressive de l'hétérogénéité inobservable (de gauche à droite, sous forme de variables de contrôle) rend de moins en moins performantes les variables d'hétérogénéité observables. De même, l'influence de la période de jeu tend à s'effacer lorsque les effets fixes individuels et de marchés sont inclus.

Dans l'ensemble, une très faible proportion des variables utilisées s'avère capable d'expliquer le niveau du prix choisi. La taille de l'industrie et l'usage que font les participants de la dénonciation constituent une exception importante. L'effet de la taille de l'industrie sur le niveau du prix est très instable, et devient fortement significatif lorsque les effets fixes de marché sont ajoutés au modèle. Le paramètre estimé suggère une corrélation positive entre la taille de l'industrie et le niveau du prix choisi par une firme conditionnellement à la stratégie d'évasion collusive. A mesure que l'effet de la taille de l'industrie devient positif, il faut cependant remarquer que la constante du modèle diminue régulièrement, jusqu'à devenir non significative. Compte tenu de la très faible variabilité de la taille du groupe parmi les observations qui ont choisi l'évasion collusive[25] le fait que la taille de l'industrie joue progressivement le rôle de constante pourrait expliquer ce résultat contre-intuitif. L'effet de la dénonciation effective est, quant à elle, conforme aux attentes. Toutes choses égales par ailleurs, un participant qui a été dénoncé à la période précédente tend à accroître le prix collusif choisi. Cette observation confirme que la dénonciation est un instrument puissant de coordination, mis au service de la collusion. Les résultats de l'équation de sélection ont en effet montré que l'usage effectif de la dénonciation – comme la menace qu'elle constitue – encourage le recours à l'évasion collusive non seulement pour le participant dénoncé mais également pour le participant dénonciateur. Ce comportement peut s'interpréter comme une anticipation de l'effet de la dénonciation sur le participant dénoncé, et cette observation confirme donc la validité de cette anticipation.

[25]La taille moyenne des marchés qui sont dans ce cas est de 3.2, sa variance de 0.60. Avec un taux d'évasion collusive moyen de 14%, seules 23 observations d'évasion collusive dans des groupes de taille supérieure à 3 subsistent.

Au total, les résultats d'estimation quant au choix du prix collusif sont très sensibles à l'inclusion de l'hétérogénéité. Le choix du niveau du prix conditionnellement à l'évasion semble donc principalement dépendre des caractéristiques des individus qui interagissent plutôt que de l'environnement économique. Au regard de la comparaison entre les deuxième et troisième colonne du Tableau 3.12, l'introduction de l'hétérogénéité individuelle affecte peu les résultats. Il semble donc, plus précisément, que le niveau du prix dépende essentiellement de la dynamique du marché qui résulte de l'interaction.

Surtout, la robustesse à la collusion tacite comme la crédibilité de la menace de dénonciation sont, elles-aussi, fortement non significatives dès que l'hétérogénéité des marchés est prise en compte. Ce dernier résultat confirme que la portée du modèle théorique proposé se limite aux conditions d'émergence de l'évasion collusive, faisant du choix du niveau de prix un problème de coordination entre les firmes qui composent le marché. De ce point de vue, les firmes tendent à outre-passer la référence naturelle que constitue l'équilibre concurrentiel légal en se coordonnant sur un prix qui lui est inférieur. L'équilibre sélectionné est donc différent de l'état d'évasion collusive qui a servi d'illustration à la présentation du modèle théorique.

Observation 3.3. *L'interaction entre les firmes conduit à choisir un prix supérieur au prix de profit nul avec évasion, mais inférieur au prix de l'équilibre concurrentiel légal.*

Les comportements observés sur les marchés expérimentaux confirment donc la portée empirique de la malédiction de Bertrand ainsi que les conditions d'émergence de l'évasion collusive. Ils montrent également, cependant, la difficulté à modéliser la coordination entre les firmes, d'où résulte le choix du prix de marché.[26] Le prix de l'équilibre concurrentiel légal, en particulier, ne constitue pas le point focal naturel que l'intuition suggérait.

[26] Dans un environnement expérimental de concurrence à la Bertrand, Abbink & Brandts (2004) rencontrent des difficultés similaires. Ils proposent une discussion approfondie des explications possibles, en termes de point focal ou d'imitation.

3.5 Conclusion

Afin d'évaluer l'efficacité potentielle de l'introduction de dispositions facilitant la dénonciation du travail illégal, ce chapitre a proposé une analyse théorique et expérimentale des déterminants de la demande de travail au noir qui émane des producteurs. Dans les termes de l'analyse économique du crime, la spécificité de la demande de travail au noir tient à ce que le bénéfice de l'illégalité dépend du comportement des firmes concurrentes. Exploitant cet aspect, l'analyse théorique a d'abord établi que l'intensité de la concurrence encourage le recours au travail au noir mais tend à en éliminer tout bénéfice pour les producteurs.

Cet effet pervers de la concurrence (appelé *malédiction de Betrand*) ne peut être combattu que partiellement. La rentabilité de l'évasion reste en effet, en toutes circonstances, un motif suffisant pour conduire les producteurs à s'y livrer. Le modèle théorique a cependant isolé deux conditions qui affectent la dynamique du marché et peuvent faire en sorte que cette rentabilité résiste aux forces de la concurrence. D'une part, conformément aux résultats classiques de l'économie industrielle, la collusion tacite permet aux firmes de maintenir durablement un prix supérieur à celui qui annule les profits. Loin de rétablir la légalité du travail sur le marché, la dénonciation renforce d'autre part les possibilités de collusion tacite. Lorsque son usage est crédible (c'est à dire suffisamment peu coûteux) la dénonciation constitue une menace qui facilite la conclusion d'un accord entre les firmes et s'ajoute aux stratégies de punition traditionnellement considérées dans les analyses consacrées à la collusion. Les producteurs s'appuient alors sur une stratégie de *silence collusif*, ne faisant usage de la dénonciation que lorsqu'un concurrent cesse de coopérer. Les programmes de clémence, qui rendent moins coûteux l'exercice de la dénonciation, apparaissent dans ce cadre comme un instrument contre-productif offrant aux fraudeurs un moyen d'améliorer la rentabilité de la fraude.

Chacun de ces deux mécanismes assurent aux producteurs qui choisissent la fraude des profits positifs. Dans ce scénario d'*évasion collusive*, le fonctionnement du marché à un prix relativement faible (*i.e.* proche des coûts légaux) n'est alors qu'une illusion de légalité. L'observation des marchés expérimentaux corrobore ces prédictions théoriques. La rentabilité de la fraude conduit dans un premier temps à une évasion très largement répandue. La concurrence tend cependant, dans un second temps, à contre-carrer cette rentabilité en forçant la diminution du prix jusqu'à annulation des profits espérés. La collusion tacite et le silence collusif s'opposent à cette dynamique et permettent de maintenir la rentabilité de l'évasion. L'intensité de la concurrence reste pourtant une force puissante, qui conduit les producteurs à choisir un prix de vente inférieur au coût légal. La portée de ces résultats est en partie restreinte par la taille de l'échantillon utilisé. De nouvelles expériences semblent donc nécessaires pour confirmer ces résultats encourageants. Accroître le nombre d'observations pourrait permettre, en outre, d'approfondir l'analyse quant aux déterminants de la coordination par laquelle les firmes sélectionnent un prix de collusion.

L'analyse théorique a par ailleurs été conduite en retenant l'hypothèse d'homogénéité des firmes. L'aversion au risque ou, plus encore, le coût moral de la fraude (Cummings, Martinez-Vazquez, McKee & *al.*, 2005) constituent pourtant autant de sources potentielles d'hétérogénéité, que des développements ultérieurs devraient prendre en compte. Dans notre cadre, la principale conséquence de cette hypothèse réside dans la symétrie des stratégies d'évasion sur le marché. Le degré d'information des firmes sur les coûts de leurs concurrentes est alors non pertinent dans l'analyse. Si les firmes sont susceptibles de faire des choix d'évasion différents, au contraire, la demande de travail au noir peut conduire à un marché où les firmes sont confrontées à une incertitude sur le coût de leurs concurrentes.

Cette propriété modifierait de façon importante l'analyse du modèle et, en particulier, les possibilités de collusion offertes par le travail au noir. Spulber (1995) montre ainsi que l'incertitude sur les coûts facilite considérablement l'obtention de profits po-

sitifs. Dans le cadre de notre analyse, ce résultat, qui a reçu récemment une première confirmation expérimentale (Abbink & Brandts, 2005), impliquerait donc que l'hétérogénéité des stratégies d'évasion est une force supplémentaire facilitant le maintient de la rentabilité de la fraude. L'hétérogénéité des choix d'évasion écarterait donc l'analyse du cadre traditionnel du modèle de Bertrand. La stratégie de cliquet ne concorde plus, dans ce cas, avec la stratégie optimale. Une généralisation naturelle de notre analyse consisterait donc également à considérer les possibilités de collusion offertes par des stratégies de punition de carotte et bâton. Dans ce cadre, Billette de Villemeur, Flochel & Versaevel (2004) montrent que le coût marginal et la durée de la punition sont des substituts dans la mise en oeuvre des accords de collusion. Ils sont alors d'autant plus faciles à mettre en œuvre que le coût marginal est élevé. Le choix du travail au noir tendrait donc à réduire la capacité du marché à maintenir la rentabilité de l'évasion. Ce dernier effet s'oppose au premier et rend nécessaire une nouvelle analyse pour lever les ambiguïtés quant à l'effet attendu d'une hétérogénéité des coûts sur les déterminants de la demande de travail au noir.

Annexes

3.A Instructions de l'expérience

Les instructions présentées ci-dessous sont lues aux participants au début de l'expérience. Ils disposent pendant la lecture du tableau résumant la fonction de demande, présenté dans le Tableau 3.A. La prise de décision au cours de l'expérience se fait à l'aide d'une interface graphique dont le Graphique 3.A propose une capture d'écran.

Instructions

Vous allez participer à une expérimentation qui s'inscrit dans un programme de recherche scientifique soutenu conjointement par l'Agence Centrale des Organismes de Sécurité Sociale et le Centre National de la Recherche Scientifique. Lors de cette session, vous allez gagner une certaine somme d'argent. Vos gains dépendent de vos décisions et des décisions des autres participants avec lesquels vous interagirez.

Le déroulement de la session expérimentale

L'expérimentation se déroulera en trois parties. Chaque partie contient plusieurs périodes. Au début de l'expérimentation, des groupes de tailles différentes sont formés au hasard. Vous êtes alors informés de la taille de votre groupe d'appartenance ; la composition et la taille de votre groupe restent inchangées tout au long de la session expérimentale.

Votre gain s'exprime en ECU (Experimental Currency Unit) ; à la fin de la session expérimentale, votre gain en ECU sera transformé en Euro (€). Vous pourrez éventuellement subir des pertes en ECU sur certaines périodes mais vous ne pourrez pas achever la session expérimentale avec des gains négatifs en ECU et a fortiori en Euros.

Les règles qui suivent décrivent le déroulement de chaque période de la première partie. De nouvelles règles vous seront successivement présentées pour les parties 2 et 3.

Le déroulement d'une période

Au début de chaque période, chaque participant doit prendre successivement deux décisions :

1. Tout d'abord une décision sur **la forme des coûts** qu'il devra supporter.
2. Puis une décision qui lui permettra de **gagner des points**.

1. En ce qui concerne la forme des coûts que vous devrez supporter, vous avez le choix entre deux options :

- **L'option A** : le **coût de chaque point** est de 5.(1 + 0,8) = 9
- **L'option B** : le niveau des coûts est aléatoire :
  - Vous avez 95 chances sur 100 que le **coût de chaque point** soit de 5
  - Vous avez 5 chances sur 100 de subir directement **un coût en ECU de 20 ECU** (sans aucun gain en points).

2. En ce qui concerne vos gains en points, vous devez choisir un nombre n compris entre 5 et 19.

- Soit le nombre **n** que vous avez choisi **n'est pas le plus petit** parmi la liste des nombres choisis par les participants de votre groupe : **votre gain en point est nul**.
- Soit le nombre **n** que vous avez choisi **est le plus petit** parmi la liste des nombres choisis par les participants de votre groupe : **votre gain en point est donné par le tableau en annexe**.

Les deux décisions précédentes (sur les coûts par points et les gains en points), vous permettent dés lors de déterminer **le nombre d'ECU que vous avez gagné au cours de cette période** : à l'exception du cas où vous subissez directement un coût de 20 ECU (avec un gain en points nul), il vous suffit dans un premier temps de calculer l'écart entre la valeur ***n*** et ***le coût de chaque point*** ; puis de multiplier cet écart par ***votre gain en points***.

Exemple 1 :

Vous choisissez l'option **A** : votre coût pour chaque point est donc de **9**.

Vous choisissez **n = 11**.

Ce nombre **n = 11** que vous avez choisi **n'est pas le plus petit** parmi la liste des nombres choisis par les participants de votre groupe : **votre gain en points est nul**.

Les ECU que vous gagnez au cours de cette période sont alors de ***(11 - 9).0 = 0****.*

Exemple 2 :

Vous choisissez **l'option A** : votre coût pour chaque point est donc de **9**.

Vous choisissez **n = 11**.

Ce nombre **n = 11** que vous avez choisi **est le plus petit** parmi la liste des nombres choisis par les participants de votre groupe et 3 participants (dont vous-même) l'ont choisi. Les gains en points pour chacun de ces participants sont donc : **6** comme l'indique le tableau.

Les ECU que vous gagnez au cours de cette période sont alors de ***(11 - 9).6 = 12****.*

Exemple 3 :

Vous choisissez **l'option B**.

Vous choisissez **n = 11**.

Ce nombre **n = 11** que vous avez choisi **est le plus petit** parmi la liste des nombres choisis par les participants de votre groupe et 2 participants (dont vous-même) l'ont choisi. Les gains en points pour chacun de ces participants sont donc : **9** comme l'indique le tableau.

- Si le tirage au sort vous indique que votre coût pour chaque point est de **5**. *Les ECU que vous gagnez au cours de cette période sont alors de* ***(11 - 5).9= 54****.*

- Si le tirage au sort vous attribue directement un coût de 20 ECU. *Les ECU que vous perdez au cours de cette période sont alors de* ***-20****.*

Quelle information vous avez sur les décisions des autres ?

A l'issue de la deuxième décision, chacun voit d'abord sur son écran la liste des nombres n que chaque participant de son groupe a choisi. Votre propre décision apparaît grisée sur l'écran.

De plus, si lors de la première décision vous avez choisi **l'option** B pour déterminer la forme de vos coûts, vous voyez sur votre écran à côté du nombre **n** de chaque participant de votre groupe, son choix entre **les options A ou B**.

L'ordre dans lequel la liste des participants de votre groupe apparaît change à chaque période. Ainsi le participant qui apparaît au début de la liste au cours de la première période peut, au cours de la deuxième période, apparaître à la fin ou en deuxième position, etc.

A la fin de chaque période, vous êtes informé du nombre d'ECU acquis au cours de la période.

Quelles informations sur votre écran concernant les périodes ?

Sur l'écran de votre ordinateur, trois zones seront présentes :
- La première vous informe de la progression de la période en cours ;
- La deuxième vous permet de prendre vos décisions ;
- La troisième vous rappelle les décisions des périodes précédentes.

Le passage d'une période à une autre

A la fin de chaque période, un tirage au sort permet ou pas de passer à une nouvelle période. Avec 75 chances sur 100 une nouvelle période commence et donc avec 25 chances sur 100 la partie s'arrête (pour tout le monde). Lorsque la partie 1 (ou la partie 2) s'achève, vous passez à la partie 2 (ou la partie 3).

Le calcul de vos gains en ECU et de sa valeur en €

A la fin de l'expérimentation (fin de la partie 3), nous calculerons le total de vos ECU sur toutes les périodes des trois parties. Ce total augmente si votre gain pour la période a été positif et diminue si votre gain pour la période a été négatif (le gain de la période qui s'affiche à l'écran apparaît alors avec un " - "). Ce total sera converti en Euro (€) sur la base de

15 ECU = 1 €

A cela s'ajoute une indemnité forfaitaire de participation de **2 €**. Cette somme vous sera payée individuellement en espèce et de façon privée juste avant de quitter la salle d'expérimentation. **Quels que soient vos gains, vous ne pouvez pas perdre d'argent**.

Nous vous encourageons à prendre quelques minutes pour relire les instructions. Si vous avez des questions, s'il vous plaît, levez votre main, nous viendrons répondre à vos questions.

Pendant le déroulement de cette session expérimentale, il vous est demandé de ne pas poser de questions et de ne pas communiquer entre vous.

Merci de bien vouloir respecter ces consignes.

Avant de commencer la première partie, nous allons vous proposer de faire trois périodes d'essais sans aucun enjeu (avec les règles de la première partie décritent ci-dessus). **Pendant les périodes d'essai, les ECU ne sont pas comptabilisés : vous pouvez donc essayer toutes les décisions que vous souhaitez pour vérifier que vous avez bien compris ces instructions.**

***

Le déroulement de la deuxième partie

Qu'est ce qui change par rapport à la première partie ?

La deuxième partie est **identique** à la première, **sauf** en ce qui concerne l'élément suivant :

Après avoir pris votre décision sur n vous voyez sur votre écran, comme lors de la première partie, la liste des décisions des participants de votre groupe, vos propres décisions étant grisées. Durant la deuxième partie, sur cette liste, en face de chaque participant de votre groupe ayant choisi B, vous avez la possibilité de mettre **une coche**. Dans ce cas, **vous-même** et **chaque participant** de votre groupe que vous avez coché subissent directement **un coût de 20 ECU**, (sans aucun gain en points).

Dans tous les cas, chaque participant du groupe qui a choisi B et qui se voit attribuer directement **un coût de 20 ECU** sait si cela est dû

- Soit au hasard (il a en effet 5 chances sur 100 que cela survienne si personne ne l'a coché et qu'il n'a coché personne) ;
- Soit à la décision anonyme d'un autre participant du groupe qui a lui-même retenu l'option B ;
- Soit à sa propre décision de cocher un autre participant.

Exemple 1 :
Vous choisissez **l'option A** : vous ne pouvez pas voir les décisions entre A et B des autres, aucune case à cocher n'apparaît et personne ne peut vous cocher.

Exemple 2 :
Vous choisissez **l'option B** : vous voyez sur l'écran les décisions des autres : si tous ont choisi A, le seul B qui apparaisse est grisé (c'est votre décision). Dans ce cas vous ne pouvez cocher personne.

Exemple 3 :
Vous choisissez **l'option B** : vous voyez sur l'écran les décisions des autres participants de votre groupe et une case à cocher apparaît à côté de chaque participant qui a choisi l'option B.

Vous pouvez décider de ne cocher aucun participant. Si tous les participants de votre groupe qui ont choisi B font le même choix, chaque participant qui a choisi B :

- a 95 chances sur 100 que **le coût de chaque point** soit de 5 ;
- a 5 chances sur 100 de subir directement **un coût de 20 ECU**, (sans aucun gain en points).

Exemple 4 :

Vous choisissez **l'option B** : vous voyez sur l'écran toutes les décisions des participants de votre groupe, votre propre décision étant grisée. Vous décidez de cocher deux B. Dans ce cas, les deux participants ainsi cochés subissent directement **un coût de 20 ECU** (sans aucun gain en points). Vous-même subissez également directement **un coût de 20 ECU** (sans aucun gain en points). Si d'autres B non cochés subsistent, les participants correspondant sont soumis au hasard décrit dans l'exemple 3.

Exemple 5 :

Vous choisissez **l'option B**.

Vous choisissez **n = 11**.

Ce nombre **n=11** que vous avez choisi **est le plus petit** parmi la liste des nombres choisis par les participants de votre groupe et 2 participants (dont vous-même) l'ont choisi. Les gains en points pour chacun de ces participants sont donc : **9** comme l'indique le tableau. **Ces gains en points sont déterminés avant toute décision de cocher d'autres participants**.

Vous voyez apparaître la liste des décisions des participants de votre groupe et une case à cocher apparaît devant chaque participant qui a choisi l'option B. Si vous décidez :

- De ne cocher aucun participant et aucun participant ne vous coche. Si le tirage au sort vous indique que votre coût pour chaque point est de **5**.

Les ECU que vous gagnez au cours de cette période sont alors de ***(11 - 5).9= 54****.*

- De cocher un autre participant :

Les ECU que vous perdez au cours de cette période sont alors de ***-20****.*

- De ne cocher aucun participant mais un autre participant vous coche :

Les ECU que vous perdez au cours de cette période sont alors de ***-20****.*

***

Le déroulement de la troisième partie

Qu'est ce qui change par rapport à la deuxième partie ?

La troisième partie est **identique** à la deuxième partie :

En particulier, si lors de la première décision vous avez choisi B, vous voyez sur l'écran toutes les décisions des participants de votre groupe, votre propre décision étant grisée. Vous avez toujours la possibilité de mettre **une coche**. Dans ce cas, chaque participant de votre

groupe que vous avez coché subit **un coût de 20 ECU** (sans aucun gain en points).

Mais **ce qui change** par rapport à la deuxième partie est l'élément suivant :
En cochant au moins un B, vous ne subissez plus **un coût de 20 ECU**, mais seulement **un coût de 10 ECU**.

Exemple :
Vous choisissez **l'option B**.
Vous choisissez **n = 11**.
Ce nombre **n = 11** que vous avez choisi **est le plus petit** parmi la liste des nombres choisis par les participants de votre groupe et 2 participants (dont vous-même) l'ont choisi. Les gains en points pour chacun de ces participants sont donc : **9** comme l'indique le tableau.
Vous voyez apparaître la liste des décisions des participants de votre groupe et vous décidez :

- De ne cocher aucun participant et aucun participant ne vous coche.
Si le tirage au sort vous indique que votre coût pour chaque point est de **5**.
Les ECU que vous gagnez au cours de cette période sont alors de ***(11 - 5).9= 54****.*

- De cocher un autre participant et personne ne vous coche :
Les ECU que vous perdez au cours de cette période sont alors de ***-10****.*

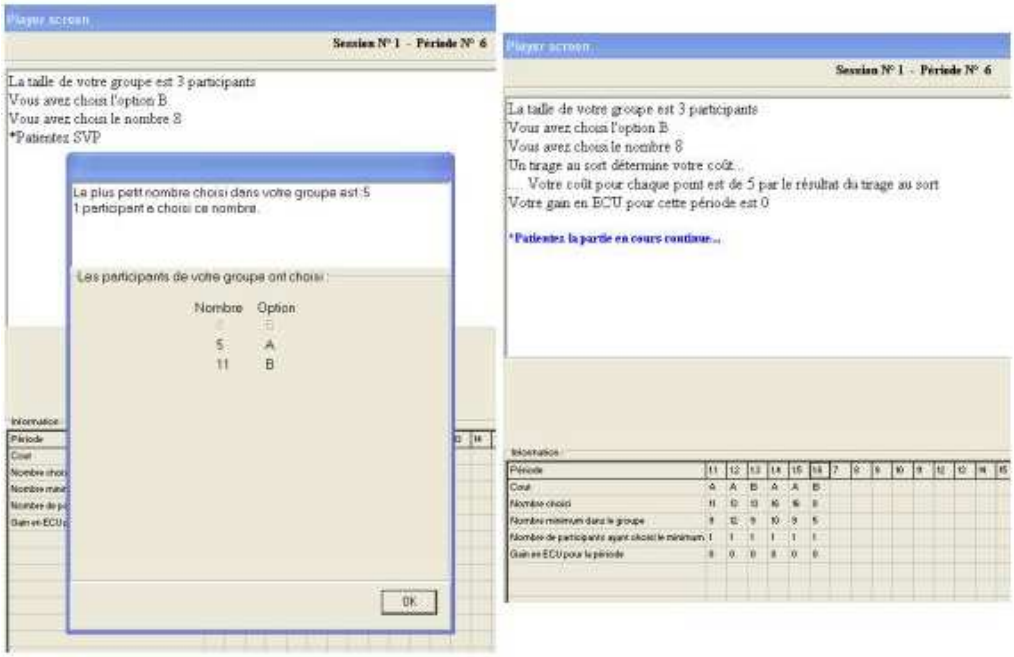
- De cocher un autre participant et un autre participant vous coche :
Les ECU que vous perdez au cours de cette période sont alors de **-10**.

- De ne cocher aucun participant mais un autre participant vous coche :
Les ECU que vous perdez au cours de cette période sont alors de ***-20****.*

Tableau 3.A – Tableau remis aux participants (Groupe de 6 ici)

p n	1	2	3	4	5	6
5	30.0	15.0	10.0	7.5	6.0	28.0
6	28.0	14.0	9.3	7.0	5.6	4.7
7	26.0	13.0	8.7	6.5	5.2	4.3
8	24.0	12.0	8.0	6.0	4.8	4.0
9	22.0	11.0	7.3	5.5	4.4	3.7
10	20.0	10.0	6.7	5.0	4.0	3.3
11	18.0	9.0	6.0	4.5	3.6	3.0
12	16.0	8.0	5.3	4.0	3.2	2.7
13	14.0	7.0	4.7	3.5	2.8	2.3
14	12.0	6.0	4.0	3.0	2.4	2.0
15	10.0	5.0	3.3	2.5	2.0	1.7
16	8.0	4.0	2.7	2.0	1.6	1.3
17	6.0	3.0	2.0	1.5	1.2	1.0
18	4.0	2.0	1.3	1.0	0.8	0.7
19	2.0	1.0	0.7	0.5	0.4	0.3
20	0.0	0.0	0.0	0.0	0.0	0.0

Graphique 3.A – Ecran de contrôle d'une firme (traitement Dénonciation)



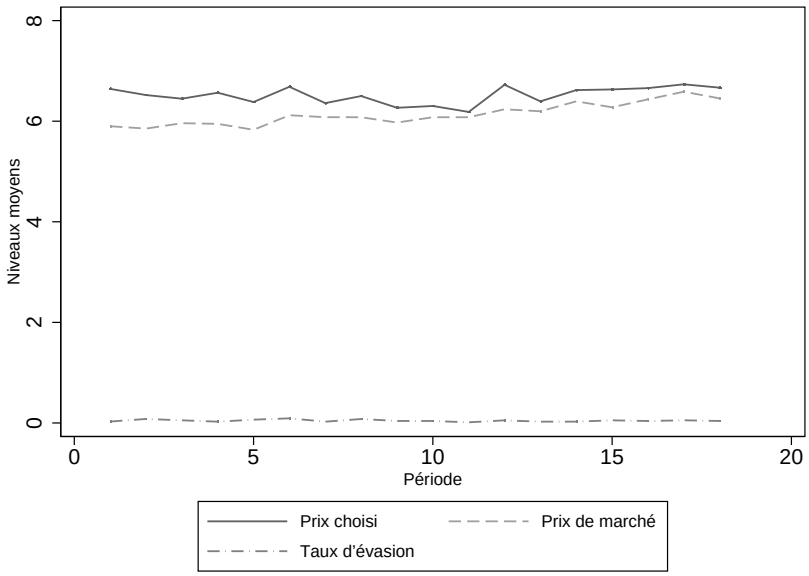
3.B Paramètres de l'expérience

La forme analytique de la fonction de demande associe la quantité demandée Q au prix de marché p selon l'expression : $Q = d - lp$. La fonction de coût reprend la forme fonctionnelle présentée dans la Section 3.1.1 : $C(q) = W.q$, $W = \{c, (1+\tau)c\}$. Les valeurs utilisées dans l'expérience sont :

Paramètres de marché : $d = 40$; $l = 2$; $\gamma = 0.25$;
Composantes du coût : $w = 5$, $\tau = 0.8$;
Politique de détection : $\alpha = 0.05$; $F = 20$; $F' = 10$.

En conséquence de ces paramètres, le prix de l'équilibre concurrentiel est $p^c = 9$ et les quantités individuelles totales écoulées sur le marché à ce prix sont $Q^c = 22$. Le profit d'évasion à l'équilibre concurrentiel qui en résulte est $\pi_F = 83.6$.

Graphique 3.B – Evolution des variables de décisions moyennes



Le Graphique 3.B propose une représentation de l'évolution du taux d'évasion (en pourcentage), du prix choisi en moyenne individuellement par les firmes et du prix de marché moyen entre les périodes de l'expérience.

3.C Robustesse des résultats à une mesure alternative de la taille du marché

Cette section reproduit les résultats empiriques présentés dans la Section 3.4 en utilisant la taille des groupes comme mesure de l'intensité de la concurrence. La prochaine section propose auparavant une évaluation des caractéristiques de l'environnement qui apparaissent pertinentes pour expliquer la dynamique des comportements dans l'expérience.

Tableau 3.B – Comportements experimentaux et variables d'environnement

Tobit (variable endogène : Prix choisi)		
Variable	**Coefficient**	**(Ecart-type)**
Taille du groupe	-0.207	(0.142)
Firmes actives à la période précédente	-0.198***	(0.038)
Round	-0.017	(0.030)
Traitement	0.227***	(0.080)
Age	0.148	(0.131)
Sexe	-0.052	(0.272)
Education	-0.301	(0.214)
Constante	5.717***	(1.869)
Effets fixes Marchés	oui	
Effets fixes Individuels	oui	
Effets fixes temporels	oui	
Effets Altéatoires temporels	oui	

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Tobit Type I. La variable endogène (prix choisi par la firme à chaque période) est censurée à gauche (le prix ne peut pas être choisi en-dessous de 5 dans l'expérience). Les effets temporels sont introduits en utilisant la variable *Période*, qui mesure le passage du temps dans l'ensemble de l'expérience. La variable *Round* mesure le passage du temps dans chaque traitement (réinitialisée en début de traitement) ; la variable *Traitement* mesure l'accès à la dénonciation, valant 1 pour le traitement Contrôle, 2 pour le traitement Dénonciation et 3 pour le traitement Clémence ; l'*Age* est mesuré en années ; la variable *Sexe* indique que le participant est un homme ; la variable *Education* mesure le nombre d'années d'études après bac.

3.C.1 Comportement des firmes et mesures d'intensité

La mesure d'intensité cohérente avec l'analyse théorique correspond en effet à la variable de taille de l'industrie qui influence le comportement des participants. L'évasion étant choisie par une écrasante majorité de participants, la variable de décision principale est le prix. Afin de discriminer entre les mesures d'intensité de la concurrence, le Tableau 3.B présente donc les résultats d'une régression "naïve", où la variable de choix (*Prix*) est expliquée par l'ensemble des variables exogènes disponibles. Afin de limiter les risques de gains négatifs, le protocole de l'expérience limite à un minimum de 5 le prix choisi par les participants. La variable endogène est donc mécaniquement censurée

à gauche et le modèle estimé est un Tobit (type I). L'hétérogénéité inobservable entre individus, entre marchés et entre les périodes est prise en compte par des variables de contrôle. La dimension temporelle (périodes) est également modélisée comme un effet aléatoire. La régression inclut les deux variables de mesure d'intensité de la concurrence discutée dans la Section 3.4.1 : la taille du groupe et le nombre de firmes actives à la période précédente.

Le nombre de firmes actives à la période précédente semble beaucoup plus déterminant sur les comportements que la taille des groupes. Les participants semblent donc prendre leurs décisions au regard de la concurrence effective à la période précédente plutôt qu'en fonction de la concurrence potentielle. Les trois prochaines sections élargissent néanmoins les résultats présentés dans ce chapitre à cette mesure alternative de l'intensité de la concurrence.

3.C.2 Description des marchés expérimentaux

La taille des groupes est pré-déterminée par le protocole de l'expérience. La distribution de tailles de groupes instaurées dans les sessions est présentée dans le Tableau 3.C.

Tableau 3.C – Distribution de l'intensité de la concurrence au sein des traitements

	Intensité de la concurrence				
	3	**4**	**5**	**6**	**Total**
Contrôle	63.54	15.55	13.14	7.77	100.00
Dénonciation	63.00	15.86	13.22	7.93	100.00
Clémence	63.02	15.85	13.21	7.92	100.00
Total	63.15	15.77	13.19	7.89	100.00

Note. Pourcentage de firmes soumises à une intensité concurrentielle donnée (mesurée par la taille du groupe d'appartenance) au sein de chaque traitement. En %.

La mesure d'intensité de la concurrence modifie les variables d'intérêts que sont la robustesse à la collusion tacite et la crédibilité de la menace, puisqu'elle détermine les profits de collusion. Les valeurs prises par ces variables lorsque l'intensité de la concurrence est mesurée par la concurrence potentielle sont présentées dans le Tableau 3.D.

Tableau 3.D – Possibilités de collusion dans chaque traitement

		Moyenne	**Ecart-Type**	**Minimum**	**Maximum**
Contrôle	$\gamma - \gamma^c$	-0.039	0.061	-0.083	0.083
	$\gamma_F - \gamma$	0.624	0.485	0.000	1.000
	$R = \gamma_F - \gamma^c$	-0.033	0.041	-0.333	-0.007
Dénonciation	$\gamma - \gamma^c$	-0.039	0.061	-0.083	0.083
	$\gamma_F - \gamma$	0.617	0.487	0.000	1.000
	$R = \gamma_F - \gamma^c$	-0.024	0.072	-0.333	0.272
Clémence	$\gamma - \gamma^c$	-0.039	0.061	-0.083	0.083
	$\gamma_F - \gamma$	0.992	0.087	0.000	1.000
	$R = \gamma_F - \gamma^c$	0.139	0.107	-0.333	0.442
Total	$\gamma - \gamma^c$	-0.039	0.061	-0.083	0.083
	$\gamma_F - \gamma$	0.773	0.419	0.000	1.000
	$R = \gamma_F - \gamma^c$	0.041	0.117	-0.333	0.442

Note. Statistiques descriptives (moyenne et écart-types entre les firmes, minimum et maximum) des mesures de possibilités de collusion au sein de chaque traitement : collusion tacite (première ligne dans chaque traitement), silence collusif (deuxième ligne) et intervalle de silence collusif (troisième ligne).

En utilisant cette mesure, la condition de robustesse à la collusion tacite est indépendante des traitements et reste donc constante. L'intervalle de silence collusif, quant à lui, est d'autant plus large que la dénonciation est facilitée.

Tableau 3.E – Taux d'évasion sous le traitement Contrôle

	Intensité de la concurrence				
	3	**4**	**5**	**6**	**Total**
Evasion	97.03	81.08	97.85	95.65	94.70
Coût légal	2.97	18.92	2.15	4.35	5.30
Total	100.00	100.00	100.00	100.00	100.00

Note. Pour chaque intensité de la concurrence, pourcentage de firmes ayant choisi le coût minimum dans le traitement Contrôle. En %.

Dans le traitement contrôle, l'évasion reste très largement majoritaire pour tous les niveaux d'intensité de la concurrence (Tableau 3.E). Le prix de marché décroît fortement avec la taille du groupe, confirmant la tendance à l'annulation des profits d'évasion en conséquence de la concurrence (Tableau 3.F).

Tableau 3.F – Distribution des prix de marché sous le traitement Contrôle

Prix de Marché	Intensité de la concurrence				
	3	**4**	**5**	**6**	**Total**
5	8.51	26.32	26.32	33.33	16.00
6	89.36	68.42	73.68	66.67	82.00
7	2.13	5.26	0.00	0.00	2.00
Total	100.00	100.00	100.00	100.00	100.00

Note. Pourcentage de marchés sur lesquels le prix minimum correspond à celui qui est indiqué en ligne, pour une intensité de la concurrence donnée, dans le traitement Contrôle. En %.

Cette généralisation de l'évasion s'étend à l'ensemble des traitements (Tableau 3.G).

Tableau 3.G – Taux d'évasion

	Intensité de la concurrence				
	3	**4**	**5**	**6**	**Total**
Contrôle	97.05	81.03	100.00	96.55	94.91
Dénonciation	94.41	95.83	95.00	100.00	95.15
Clémence	94.01	98.81	100.00	97.62	95.85
Total	94.98	92.99	98.32	98.13	95.36

Note. Pour chaque intensité de la concurrence, pourcentage de firmes ayant choisi l'évasion dans chaque traitement (trois premières lignes) et dans l'ensemble (dernière ligne). En %.

En termes de prix choisi (Tableau 3.H), deux sous-groupes apparaissent à nouveau. Lorsque la taille du marché est de 3, le prix apparaît fortement croissant de la facilité de la dénonciation. Pour les groupes de taille supérieure, l'effet de la concurrence l'emporte et le prix tend à décroître entre les traitements. Les prix d'équilibre (Tableau 3.I), quant à eux, sont décroissants de l'intensité de la concurrence à traitement donné ; et croissants de la facilité de la dénonciation à intensité donnée.

Tableau 3.H – Prix choisi moyen

	Intensité de la concurrence				
	3	**4**	**5**	**6**	**Total**
Contrôle	6.54	6.83	6.14	6.28	6.51
Dénonciation	6.49	6.24	6.27	6.03	6.38
Clémence	6.96	6.12	6.00	6.10	6.63
Total	6.69	6.35	6.13	6.12	6.52

Note. Pour chaque intensité de la concurrence, prix choisi en moyenne par les firmes dans chaque traitement (trois premières lignes) et dans l'ensemble (dernière ligne).

Le comportement de dénonciation décrit dans le Tableau 3.9 n'est pas distingué selon les tailles de groupe, et reste donc inchangé. Les caractéristiques du marché en termes de robustesse à la collusion tacite et de crédibilité de la menace de dénonciation en sont en revanche affectés. La distribution de l'évasion collusive en fonction de ces propriétés est

Tableau 3.I – Prix d'équilibre moyen

	Intensité de la concurrence				
	3	**4**	**5**	**6**	**Total**
Contrôle	5.97	5.81	5.82	5.59	5.90
Dénonciation	6.13	6.00	6.00	5.83	6.07
Clémence	6.58	6.00	6.00	6.00	6.37
Total	6.26	5.95	5.95	5.83	6.14

Note. Pour chaque intensité de la concurrence, prix minimum moyen sur les marchés dans chaque traitement (trois premières lignes) et dans l'ensemble (dernière ligne).

présentée dans le Tableau 3.J. La robustesse à la collusion tacite semble influencer très faiblement le recours à l'évasion collusive. La robustesse à la collusion tacite garantit cependant, en toute circonstance, un niveau très faible d'évasion collusive. Lorsque le marché n'est pas robuste à la collusion tacite, à l'inverse, la crédibilité de la menace de dénonciation influence considérablement la proportion d'évasion collusive.

Tableau 3.J – Evasion collusive observée

	Silence Collusif				
Collusion	Crédibilité		Intervalle		**Total**
Tacite	$\gamma > \gamma^F$	$\gamma < \gamma^F$	$R < 0$	$R > 0$	
$\gamma < \gamma^c$	3.7	16.6	6.5	25.1	14.9
$\gamma > \gamma^c$	4.4	2.6	4.5	2.6	3.7
Total	3.1	14.8	10.1	17.0	13.1

Note. Pourcentage d'observations pour lesquelles la variable *EC* est égale à 1 (évasion et prix choisi supérieur à 6). En %. *En ligne* : Robustesse à la collusion tacite ; *Colonne de gauche* : Crédibilité de la menace de dénonciation ; *Colonne de droite* : Intervalle de silence collusif.

3.C.3 Mise en œuvre de l'évasion collusive

Le Tableau 3.K présente les résultats de régression de la variable mesurant l'évasion collusive (*EC*, indicatrice de ce que la firme choisi l'évasion et un prix supérieur à 6)

Tableau 3.K – Evasion collusive

Variable	**Coefficient**	**(Ecart-type)**
Probabilité d'évasion collusive (Probit, variable endogène : EC)		
$I[\gamma > \gamma^c]$	-0.152	(0.646)
$I[\gamma^F > \gamma]$	0.475*	(0.252)
Participant dénonciateur	0.269**	(0.115)
Participant dénoncé	0.200	(0.200)
Age	0.171	(0.145)
Sexe	0.363	(0.232)
Education	-0.254	(0.255)
Période	0.004	(0.016)
Round	-0.068*	(0.040)
Nombre de firmes actives	-0.326***	(0.076)
Constante	-3.389	(2.459)
Contrôles fixes Marchés	*oui*	
Distributions Estimées		
$\hat{\sigma}$	0.422	(0.105)
$\hat{\rho}$	0.151***	(0.0638)

Niveaux de signification : *** 10%, ** 5%, *** 1%.

Note. Probit à effets individuels aléatoires. La variable endogène (EC) vaut 1 lorsqu'une observation a choisit l'évasion collusive (évasion et prix choisi supérieur à 6). L'hétérogénéité inobservable individuelle est incorporée par un effet aléatoire, propre aux participants. La variable $I[\gamma > \gamma^c]$ vaut 1 si le marché est robuste à la collusion tacite, la variable $I[\gamma^F > \gamma]$ indique que la menace de dénonciation est crédible. *Participant dénonciateur* indique que l'observation a dénoncé au moins un autre participant à la période précédente, *Participant dénoncé* qu'elle a été dénoncée par au moins un autre participant à la période précédente. L'*Age* est mesuré en années ; la variable *Sexe* indique que le participant est un homme ; la variable *Education* mesure le nombre d'années d'études après bac. La variable *Période* mesure le passage du temps dans l'ensemble de l'expérience ; la variable *Round* le passage du temps dans chaque traitement (réinitialisée en début de traitement). Le *Nombre de firmes actives* correspond au nombre de firmes qui ont choisi le prix minimum à la période précédente et mesure donc la concurrence effective passée.

sur les mesures de robustesse à la collusion tacite et de crédibilité de la menace de dénonciation, évaluées en utilisant la taille des groupes pour mesurer l'intensité de la concurrence. Les signes restent inchangés et la crédibilité de la menace de dénonciation explique significativement le recours au silence collusif. La robustesse à la collusion tacite n'est plus, ici, significative. Il faut cependant rappeler que la robustesse à la collusion tacite est mesurée comme l'inverse de l'intensité de la concurrence. Lorsque cette dernière est évaluée par la taille du groupe, la robustesse à la collusion tacite présente donc une variabilité très faible. Elle reste inchangée, en particulier, pour un participant donné et est donc incapable de recouvrir les variations de comportement au cours de l'expérience. A cette exception – notable – près, les résultats sont très similaires à ceux qui ont été présentés dans la Section 3.4.3.

Le Tableau 3.L présente les résultats du Tobit type II expliquant simultanément le recours à l'évasion collusive et la coordination entre les firmes pour choisir le prix de collusion. Le comportement de dénonciation affecte très différemment les variables d'intérêt. Comme l'indique la première colonne du tableau, le fait d'avoir été dénoncé (respectivement d'avoir utilisé la dénonciation) semble en effet ne pas expliquer le recours au silence collusif (resp. le choix du prix) mais est fortement significatif dans l'équation de choix du prix (resp. de recours au silence collusif). Ces variables sont donc respectivement exclues des équations dans lesquelles elles sont non pertinentes afin de permettre l'identification du modèle.[27] Ces tendances sont conformes au résultats de la Section 3.4.3 où ces variables s'avéraient avoir un pouvoir d'explication assez faible dans les équations correspondantes. Une différence importante avec ces résultats, en revanche, concerne le passage du temps (*Période*), qui influence ici significativement la coordination des firmes. Les périodes de jeu sont utilisées comme un moyen de communication permettant d'accroître progressivement le prix de collusion.

[27]En l'absence de cette restriction, le modèle avec hétérogénéité n'est pas identifié en raison de problèmes de colinéarité entre les variables.

Tableau 3.L – Evasion collusive et coordination

	Coefficient	t	Coefficient	t	Coefficient	t
Probabilité d'évasion collusive (Probit, variable endogène : EC)						
$I[\gamma > \gamma^c]$	0.220	1.01	0.225	1.03	0.225	1.03
$I[\gamma^F > \gamma]$	0.624***	3.37	0.643***	3.44	0.644***	3.44
Participant dénonciateur	0.382***	4.06	0.407***	4.42	0.407***	4.42
Participant dénoncé	0.360**	2.14	-	-	-	-
Age	-0.075	-1.16	-0.060	-0.93	-0.059	-0.93
Sexe	0.565***	5.08	0.551***	4.97	0.551***	4.97
Education	0.183	1.59	0.164	1.43	0.164	1.43
Round	-0.036	-1.25	-0.031	-1.08	-0.031	-1.08
Nombre de firmes actives	-0.428***	-7.93	-0.448***	-8.38	-0.448***	-8.38
Constante	0.0438	0.04	-0.150	-0.14	-0.150	-0.14
Sélection de l'équilibre (Tobit, variable endogène : p)						
$I[\gamma > \gamma^c]$	0.125	0.18	0.552	0.40	0.656	0.40
$I[\gamma^F > \gamma]$	0.422	0.60	-0.220	-0.31	-0.220	-0.31
Participant dénonciateur	0.217	0.79	-	-	-	-
Participant dénoncé	0.880**	2.41	0.640**	2.11	0.640**	2.11
Age	0.273*	1.68	0.510	0.61	0.610	1.25
Sexe	-0.279	-0.69	-1.086	-0.99	-0.966	-1.07
Education	-0.711***	-2.72	-1.154	-0.81	-1.273	-1.43
Période	0.075**	2.33	0.062**	1.99	0.062**	1.99
Round	0.073	0.85	0.036	0.48	0.036	0.48
Nombre de firmes actives	-0.090	-0.34	0.129	0.66	0.129	0.66
Constante	4.874	1.55	1.645	0.14	-0.169	-0.02
Contrôles Individuels	-	-	oui		oui	
Contrôles Marchés	-	-	-	-	oui	
			Distributions estimées			
$\hat{\sigma}$	1.530	-	1.1217	-	1.217	-
$\hat{\rho}$	-0.107	0.04[a]	-0.140	0.230	-0.170	0.230

Niveaux de signification : *** 10%, ** 5%, *** 1%.

[a] Test de Wald d'indépendance des équations estimées.

Note. Tobit Type II. *Moitié supérieure* : Variable endogène (EC) valant 1 lorsqu'une observation a choisit l'évasion collusive (évasion et prix choisi supérieur à 6). *Moitié inférieure* : Variable endogène égale au prix choisi, conditionnellement à $EC = 1$. La variable $I[\gamma > \gamma^c]$ vaut 1 si le marché est robuste à la collusion tacite, la variable $I[\gamma^F > \gamma]$ indique que la menace de dénonciation est crédible. *Participant dénoncé* indique que l'observation a été dénoncée par au moins un autre participant à la période précédente, *Participant dénonciateur* qu'il a dénoncé au moins un autre participant à la période précédente. L'*Age* est mesuré en années ; la variable *Sexe* indique que le participant est un homme ; la variable *Education* mesure le nombre d'années d'études après bac. La variable *Période* mesure le passage du temps dans l'ensemble de l'expérience ; la variable *Round* le passage du temps dans chaque traitement (réinitialisée en début de traitement). Le *Nombre de firmes actives* correspond au nombre de firmes qui ont choisi le prix minimum à la période précédente et mesure donc la concurrence effective passée.

Conclusion Générale

Les travaux présentés dans cette thèse constituent autant d'applications-types, destinées à évaluer la capacité des incitations à réconcilier les intérêts du principal et de l'agent lorsqu'est prise en compte l'influence d'agents économiques périphériques à cette relation contractuelle. Ils prolongent par ce biais l'objectif que s'est assigné la théorie de l'agence, en enrichissant la description du contexte institutionnel dans lequel s'inscrivent les transactions. Les applications retenues comprennent plus particulièrement l'intervention d'une tierce partie, affectant la relation établie entre le principal et l'agent, et se distinguent par la structure d'intérêts qu'entretiennent les joueurs.

Dans les situations de corruption (*Chapitre 1*), le corrupteur constitue un tiers affecté par les décisions que l'agent est appelé à prendre en vertu du contrat qui le lie au principal. Le pacte de corruption, conclut entre l'agent et le corrupteur, se greffe à ce contrat de délégation dans le but d'influencer la décision de l'agent dans le sens attendu par le corrupteur. L'agent se trouve en conséquence à l'intersection de deux engagements divergents – contrat de délégation et pacte de corruption.

A partir d'un survol de la littérature récente consacrée à la microéconomie de la corruption, nous avons d'abord montré que le comportement de corruption de l'agent découle des propriétés de chacun de ses engagements. Il apparaît, en particulier, que le principal et le corrupteur peuvent parfois recourir à des instruments identiques pour influencer le comportement de l'agent dans des directions opposées. L'agent fait face, par

exemple, à un conflit de réciprocités lorsque le principal opte pour un salaire d'efficience. L'influence de cet *effet de délégation* sur le comportement de corruption est testée par une investigation expérimentale fondée sur un jeu à trois joueurs. Les comportements observés confirment l'importance du conflit de réciprocités dans la décision de corruption. Les résultats permettent par là d'évaluer l'efficacité des instruments utilisés par le principal pour lutter contre la corruption.

La gestion de l'offre de soins de santé (*Chapitre 2*) est gouvernée par des intérêts contradictoires, qui président ensemble à la conception de la rémunération des médecins. Les autorités qui l'administrent, responsables de la maîtrise des coûts du système de santé, doivent en effet tenir compte de l'exigence de qualité des soins qui émane des patients. Les contrats de rémunération qui régissent l'offre de soins doivent donc répondre à ce double objectif, et promouvoir la santé tout en assurant l'efficacité de l'offre de soins.

L'analyse que nous proposons s'appuie sur une description des choix optimaux des médecins en termes de marges extensives (quantité de travail) et de marge intensive (temps consacré aux actes). Nous étudions un mode de rémunération original, instauré en 1999 au Québec, consistant à combiner une rémunération fixe et un paiement indexé sur la performance. L'analyse théorique met en évidence les ambiguïtés de l'effet des incitations sur ces décisions. Le modèle économétrique, estimé grâce à l'expérience naturelle fournie par la réforme, participe à lever ces indéterminations. Si elle permet d'accroître le temps consacré à chaque acte et d'encourager la diversification des activités, la combinaison de rémunérations adoptée par les autorités Québecoises apparaît surtout comme un instrument puissant de rééquilibrage des rémunérations entre des profils de pratique hétérogènes. Cet objectif est atteint au prix d'un accroissement important du coût du système de soins. Les résultats mettent cependant en évidence les gains d'efficacité permis par l'auto-sélection : en choisissant librement le mode de rémunération sous lequel ils exercent, les médecins révèlent leurs préférences à l'égard des choix de pratique et adoptent celui qui encourage les activités qu'ils sont portés à exercer.

Dans le cadre de la demande de travail au noir (*Chapitre 3*), le bénéfice que tire le principal du contrat – possiblement illégal – passé avec l'agent dépend du comportement des principaux qui interviennent sur le même marché. Mettant l'accent sur le rôle de la dénonciation dans la lutte contre le travail au noir, nous montrons qu'elle constitue un mécanisme de réconciliation permettant aux principaux d'accroître la rentabilité du contrat illégal.

Notre analyse théorique et expérimentale du comportement de demande de travail au noir prend explicitement en compte l'influence de la concurrence. En l'absence de dénonciation, l'analyse théorique établit d'abord que la concurrence tend à éliminer le bénéfice de la fraude mais encourage l'évasion fiscale. Dans ce cadre, la dénonciation constitue une menace crédible contre les baisses de prix et peut permettre aux fraudeurs d'éviter cette situation. Elle peut alors être utilisée pour maintenir durablement un niveau de prix garantissant des profits positifs malgré la concurrence. La dynamique observée sur les marchés expérimentaux est conforme à ces prédictions. La concurrence a donc un effet pervers sur la demande de travail au noir – encourageant la fraude mais en éliminant le bénéfice – que la dénonciation ne fait que renforcer en rétablissant la rentabilité de l'évasion – ce qui ne fait qu'accroître son attrait.

Ces applications constituent, dans leur ensemble, autant de configurations des intérêts des joueurs en présence – respectivement divergents, contradictoires et disposant d'un mécanisme de réconciliation – qui permettent d'élargir le champ d'investigation couvert par la théorie de l'agence et, singulièrement, l'analyse des incitations. S'appuyant sur sa propre expérience (Helpman & Laffont, 1975), Jean-Jacques Laffont remarquait dans un ouvrage récent (Laffont & Martimort, 2002 p.3) :

> *« [at the begininng of the seventies, general equilibrium theory (GE) met incentives]. The problems encountered were so serious that a whole generation of general equilibrium theorists momentarily gave up the grandiose framework of GE to reconsider the problem of exchange under asymmetric information in its simplest form, i.e., between two traders. In a sense, the theorists went back to basics.»*

Après plus de 30 ans de développements, la théorie de l'agence constitue désormais un outil d'analyse puissant des propriétés des transactions qui résultent des asymétries d'information. La question se pose aujourd'hui de savoir si les perspectives ouvertes à long terme résident dans le réexamen de l'analyse walrassienne à l'aune des avancées de l'économie de l'information (Magill & Quinzii (2005) ou Cooley, Marimon & Quadrini (2004) par exemple) ; ou dans la description exhaustive des variétés de contexte qui composent l'économie («*contemporary economic history may exhibit an interplay of local uniformity coupled with global institutional and behavioral diversity rather than global convergence and uniformity*», Bowles & Gintis, 2000 p.1433). Chacune de ces perspectives passe à moyen terme par l'agenda de recherche que Stiglitz (2000, p.1471) résumait en ces termes : *advances will entail new applications, showing the role that information considerations play in explaining a broader array of institutions and behavior*». En prolongeant l'analyse à l'influence de l'environnement dans lequel sont conclus les contrats entre le principal et l'agent, c'est cet horizon que les travaux présentés ici espèrent avoir participé à rapprocher.

La théorie de l'agence s'est en effet largement développée sous l'hypothèse que le principal possède tout pouvoir dans la conception du contrat. L'intervention d'une tierce partie peut pourtant, comme nous l'avons montré, imposer des limites importantes au contrôle du principal. Cette intervention provoque une redistribution du pouvoir de négociation, entre les joueurs affectés par le contrat, et modifie le rôle joué par les incitations. Ce prolongement de l'analyse conduit alors à une description plus fine des déterminants de l'interaction entre les joueurs. Il permet en particulier de réintégrer la complexité de la dynamique de concessions réciproques dont, dans le cadre plus large des théories de l'échange social, Simon (1978, p.3) avait déja souligné l'importance :

> *« [. . .] when two or more people interact, each expects to get something from the interaction that is valuable to him and is thereby motivated to give something up that is valuable to the others. Social exchange, in the form of "inducements–contrinutions balance" [. . .] was a central ingredient in sociological theories.»*

Bibliographie

Aas I. H. M. (1995). Incentives and financing methods, *Health Policy*, 34 (3), pp. 205-220.

Abbink K. (2002). Fair Salaries and the Moral Costs of Corruption, *Centre for Decision research and Experimental economics (CeDEx) WP*, 2002 (5).

Abbink K. (2004). Staff rotation as an anti-corruption policy : an experimental study, *European Journal of Political Economy*, 20 (4), pp. 887-906.

Abbink K., Brandts J. (2004). 24, *Centre for Decision research and Experimental economics (CeDEx) WP*, 2003-8.

Abbink K., Brandts J. (2005). Price Competition Under Cost Uncertainty : A Laboratory Analysis, *Economic Inquiry*, 43 (3), pp. 636-648.

Abbink K., Hennig-Schmidt H. (2005). Neutral versus Loaded Instructions in a Bribery Experiment, *Experimental Economics*, Forthcoming.

Abbink K., Irlenbusch B., Renner E. (2000). The moonlighting game : An experimental study on reciprocity and retribution, *Journal of Economic Behavior & Organization*, 42 (2), pp. 265-277.

Abbink K., Irlenbusch B., Renner E. (2002). An Experimental Bribery Game, *Journal of Law, Economics, & Organization*, 18 (2), pp. 428-454.

Abel-Smith B., Mossialos E. (1994). Cost containment and health care reform : a study of the European Union, *Health Policy*, 28 (2), pp. 89-132.

Abreu D. (1986). Extremal equilibria of oligopolistic supergames, *Journal of Economic Theory*, 39 (1), pp. 191-225.

Acemoglu D., Verdier T. (1998). Property Rights, Corruption and the Allocation of Talent : a General Equilibrium Approach, *Economic Journal*, 108 (450), pp. 1381-1403.

Ades A., Di Tella R. (1997). National Champions and Corruption : Some Unpleasant Interventionist Arithmetic, *Economic Journal*, 107 (443), pp. 1023-1042.

Ahrend R. (2002). Press Freedom, Human Capital and Corruption, *DELTA WP*, 2002-11.

Aidt T. S. (2003). Economic analysis of corruption : a survey, *Economic Journal*, 113 (491), pp. F632-F652.

Akerlof G. A. (1982). Labor Contracts as Partial Gift Exchange, *Quarterly Journal of Economics*, 97 (4), pp. 543-569.

Akerlof G. A. (1984). Gift Exchange and Efficiency-Wage Theory : Four Views, *American Economic Review*, 74 (2), pp. 79-83.

Akerlof G. A., Kranton R. E. (2005). Identity and the Economics of Organizations., *Journal of Economic Perspectives*, 19 (1), pp. 9-32.

Akerlof G. A., Yellen J. L. (1990). The Fair Wage-Effort Hypothesis and Unemployment, *Quarterly Journal of Economics*, 105 (2), pp. 255-283.

Akerlof G. A., Yellen J. L. (1994). Gang Behavior, Law Enforcement and Community Values, *in* H. Aaron, T. Mann and T. Taylor (Eds.), *Values and Public Policy*, pp. 173-209. Washington (D.C.) : Brookings Institution.

Amemiya T. (1984). Tobit models : A survey, *Journal of Econometrics*, 24 (1-2), pp. 3-61.

Anderson J. E., Bandiera O. (2000). Mafias as Enforcers, *Boston College WP*, 480.

Andrianova S. (2001). Corruption and Reputation, *Scottish Journal of Political Economy*, 48 (3), pp. 245-259.

Andvig J. C., Moene K. O. (1990). How corruption may corrupt, *Journal of Economic Behavior & Organization*, 13 (1), pp. 63-76.

Apesteguia J., Dufwenberg M., Selten R. (2005). Blowing the whistle, *Universidad Pública de Navarra WP*, 0303.

Arrow K. J. (1963). Uncertainty and the Welfare Economics of Medical Care, *American Economic Review*, 53 (5), pp. 941-973.

Arrow K. J. (2001a). The five most significant developments in economics of the twentieth century., *European Journal of the History of Economic Thought*, 8 (3), pp. 298-304.

Arrow K. J. (2001b). Reflections on the reflections, *Journal of Health Politics, Policy and Law*, 26 (5), pp. 1197-1203.

Aubert C., Kovacic W., Rey P. (2005). The Impact of Leniency Programs on Cartels, *International Journal of Industrial Organization*, Forthcoming.

Bac M. (1996a). Corruption and Supervision Costs in Hierarchies, *Journal of Comparative Economics*, 22 (2), pp. 99-118.

Bac M. (1996b). Corruption, supervision, and the structure of hierarchies, *Journal of Law, Economics, & Organization*, 12 (2), pp. 277-298.

Bac M. (2001). Corruption, Connections and Transparency : Does a Better Screen Imply a Better Scene ?, *Public Choice*, 107 (1-2), pp. 87-96.

Bac M., Bag P. K. (2000). Cost Effective Control of Corruption in Public Offices, *Bilckent University, Department of economics WP*.

Bag P. K. (1997). Controlling Corruption in Hierarchies, *Journal of Comparative Economics*, 25 (3), pp. 322-344.

Baltagi B. H., Bratberg E., Holmas T. H. (2005). A Panel Data Study of Physicians' Labor Supply : The Case of Norway, *Health Economics*, Forthcoming.

Bandiera O. (2003). Land Reform, the Market for Protection, and the Origins of the Sicilian Mafia : Theory and Evidence, *Journal of Law Economics & Organization*, 19 (1), pp. 218-244.

Banerjee A. V. (1997). A Theory of Misgovernance, *Quarterly Journal of Economics*, 112 (4), pp. 1289-1332.

Banfield E. C. (1975). Corruption as a Feature of Governmental Organization, *Journal of Law and Economics*, 18 (3), pp. 587-605.

Barankay I., Bandiera O., Rasul I. (2005). Managerial Incentives in Hierarchies : Evidence from a Field Experiment, *Mimeo*.

Bardhan P. (1997). Corruption and Development : A Review of Issues, *Journal of Economic Literature*, 35 (3), pp. 1320-1346.

Bardhan P. (2005). The economist's approach to the problem of corruption, *World Development*, Forthcoming.

Barro J., Beaulieu N. (2003). Selection and Improvement : Physician Responses to Financial Incentives, *NBER WP*, 10017.

Becker G. S. (1968). Crime and Punishment : An Economic Approach, *Journal of Political Economy*, 76 (2), pp. 169-217.

Becker G. S., Lewis H. G. (1973). On the Interaction between the Quantity and Quality of Children, *Journal of Political Economy*, 81 (2, Part 2), pp. S279-S288.

Becker G. S., Stigler G. J. (1974). Law Enforcement, Malfeasance, and Compensation of Enforcers, *Journal of Legal Studies*, 3 (1), pp. 1-18.

Benabou R., Tirole J. (2003). Intrinsic and Extrinsic Motivation, *Review of Economic Studies*, 70 (3), pp. 489-520.

Berentsen A., Brügger E., Lörtscher S. (2004). On Cheating and Whistle-Blowing, *University of Zürich, Institute for Empirical Research in Economics (IEW) WP*, 153.

Berg J., Dickhaut J., McCabe K. (1995). Trust, Reciprocity, and Social History, *Games and Economic Behavior*, 10 (1), pp. 122-142.

Bernheim B. D., Whinston M. D. (1986). Common Agency, *Econometrica*, 54 (4), pp. 923-942.

Berninghaus S. K., Ehrhart K.-M. (1998). Time horizon and equilibrium selection in tacit coordination games : Experimental results, *Journal of Economic Behavior & Organization*, 37 (2), pp. 231-248.

Bertrand J. (1883). Recherche sur la Théorie Mathématique de la Richesse, *Journal des Savants*, 48, pp. 499-508.

Bertrand M., Duflo E., Mullainathan S. (2004). How Much Should We Trust Differences-in-Differences Estimates ?, *Quarterly Journal of Economics*, 119 (1), pp. 249-275.

Besley T., Ghatak M. (2005). Competition and Incentives with Motivated Agents, *American Economic Review*, 95 (3), pp. 616-636.

Besley T., McLaren J. (1993). Taxes and Bribery : The Role of Wage Incentives, *Economic Journal*, 103 (416), pp. 119-141.

Billette de Villemeur E., Flochel L., Versaevel B. (2004). Optimal Collusion in Oligopoly Supergames : Marginal Costs Matter, *Mimeo*.

Blank R. M. (1988). Simultaneously Modeling the Supply of Weeks and Hours of Work among Female Household Heads, *Journal of Labor Economics*, 6 (2), pp. 177-204.

Blomquist N. S. (1989). Comparative Statics for Utility Maximization Models with Nonlinear Budget Constraints, *International Economic Review*, 30 (2), pp. 275-296.

Blume A. (2003). Bertrand without fudge, *Economics Letters*, 78 (2), pp. 167-168.

Blundell R., Duncan A., McCrae J., Meghir C. (2000). The labour market impact of the working families' tax credit, *Fiscal Studies*, 21 (1), pp. 75 - 104.

Blundell R., Macurdy T. (1999). Labor supply : A review of alternative approaches, *in* O. C. Ashenfelter and D. Card (Eds.), *Handbook of Labor Economics*, Vol. 3 (1), Ch. 27, pp. 1559-1695.

Bolduc D., Fortin B., Fournier M.-A. (1996). The Effect of Incentive Policies on the Practice Location of Doctors : A Multinomial Probit Analysis, *Journal of Labor Economics*, 14 (4), pp. 703-732.

Bolton P., Dewatripont M. (2005). *Contract Theory*. Cambridge (MA) : MIT Press.

Bowles R., Garoupa N. (1997). Casual police corruption and the economics of crime, *International Review of Law and Economics*, 17 (1), pp. 75-87.

Bowles S., Gintis H. (2000). Walrasian economics in retrospect, *Quarterly Journal of Economics*, 115 (4), pp. 1411-1439.

Boycko M., Shleifer A., Vishny R. W. (1996). A Theory of Privatisation, *Economic Journal*, 106 (435), pp. 309-319.

Brisset K., Thomas L. (2004). Leniency Program : A New Tool in Competition Policy to Deter Cartel Activity in Procurement Auctions, *European Journal of Law and Economics*, 17 (1), pp. 5-19.

Brunetti A., Weder B. (2003). A free press is bad news for corruption, *Journal of Public Economics*, 87 (7-8), pp. 1801-1824.

Buccirossi P., Spagnolo G. (2001). The Effects of leniency on Illegal Transactions : How (Not) to Fight Corruption, *Stockholm School of Economics WP*, 456.

Burguet R., Che Y.-K. (2004). Competitive Procurement with Corruption, *RAND Journal of Economics*, 35 (1), pp. 50-68.

Burtless G., Hausman J. A. (1978). The Effect of Taxation on Labor Supply : Evaluating the Gary Negative Income Tax Experiment, *Journal of Political Economy*, 86 (6), pp. 1103-1130.

Butz D. A. (1999). The Disconnection Between Principal-Agent Theory and Empirical Work : A Review of Bernard Salanie, The Economics of Contracts, *International Journal of the Economics of Business*, 6 (1), pp. 131-140.

Cabrales A., Charness G. (2003). Optimal Contracts, Adverse Selection, and Social Preferences : An Experiment, *University of California, Santa Barbara, Department of Economics WP*, 1103.

Cadot O. (1987). Corruption as a gamble, *Journal of Public Economics*, 33 (2), pp. 223-244.

Camerer C., Talley E. (2005). Experimental Study of the law, *in* A. M. Polinsky and S. Shavell (Eds.), *Handbook of law and economics*, Ch. 21. Amsterdam : North-Holland.

Carillo M. R., Pugno M. (2004). The underground economy and underdevelopment, *Economic Systems*, 28 (3), pp. 257-279.

Carlsen F., Grytten J. (1998). More physicians : improved availability or induced demand ?, *Health Economics*, 7 (6), pp. 495-508.

Carlsen F., Grytten J. (2000). Consumer satisfaction and supplier induced demand, *Journal of Health Economics*, 19 (5), pp. 731-753.

Carrillo J. D. (2000a). Corruption in Hierarchies, *Annales d'Economie et de Statistiques*, 59, pp. 37-61.

Carrillo J. D. (2000b). Graft, Bribes, and the Practice of Corruption, *Journal of Economics & Management Strategy*, 9 (3), pp. 257-286.

Celik G., Sayan S. (2005). To Give In or Not To Give In To Bribery ? Setting the Optimal Fines for Violations of Rules when the Enforcers are Likely to Ask for Bribes, *Mimeo.*

Chander P., Wilde L. (1992). Corruption in tax administration, *Journal of Public Economics*, 49 (3), pp. 333-349.

Chang J.-j., Lai C.-c., Yang C. C. (2000). Casual police corruption and the economics of crime : Further results, *International Review of Law and Economics*, 20 (1), pp. 35-51.

Chiappori P.-A., Salanié B. (2000). Testing for Asymmetric Information in Insurance Markets, *Journal of Political Economy*, 108 (1), pp. 56-78.

Chiappori P.-A., Salanié B. (2003). Testing Contract Theory : A Survey of Some Recent Work, *in* M. Dewatripont, L. Hansen and S. Turnovsky (Eds.), *Advances in Economics and Econometrics, Eight World Congress*, pp. 115-149. Cambridge (MA) : Cambridge University Press.

Choi J. P., Thum M. (2005). Corruption and the shadow economy, *International Economic Review*, 46 (3), pp. 817-836.

Christensen L. R., Jorgenson D. W., Lau L. J. (1975). Transcendental Logarithmic Utility Functions, *American Economic Review*, 65 (3), pp. 367-383.

Clark D. J., Riis C. (2000). Allocation efficiency in a competitive bribery game, *Journal of Economic Behavior & Organization*, 42 (1), pp. 109-124.

Clark K., Sefton M. (2001). The Sequential Prisoner's Dilemma : Evidence on Reciprocation, *Economic Journal*, 111 (468), pp. 51-68.

Coase R. H. (1937). The Nature of the Firm, *Economica*, 4 (16), pp. 386-405.

Collie D. R. (2004). Sustaining Collusion With Asymmetric Costs, *Royal Economic Society Annual Conference*, 155.

Colombino U. (1998). Evaluating the effects of new telephone tariffs on residential users' demand and welfare. A model for Italy, *Information Economics and Policy*, 10 (3), pp. 283-303.

Compte O., Lambert-Mogiliansky A., Verdier T. (2005). Corruption and competition in procurement auctions, *RAND Journal of Economics*, 36 (1), pp. 1-15.

Cooley T. F., Marimon R., Quadrini V. (2004). Aggregate Consequences of Limited Contract Enforceability, *Centre for Economic Policy Research DP*, 4173.

Cooper R., DeJong D. V., Forsythe R., Ross T. W. (1996). Cooperation without Reputation : Experimental Evidence from Prisoner's Dilemma Games, *Games and Economic Behavior*, 12 (2), pp. 187-218.

Cooter R. D., Garupa N. (2000). The Virtuous Circle of Distrust : A Mechanism to Deter Bribes and Other Cooperative Crimes, *Berkeley Olin Program in Law & Economics WP*, 32.

Cowell F. A. (1981). Taxation and Labour Supply with Risky Activities, *Economica*, 48 (192), pp. 365-379.

Cowell F. A. (1985). Tax evasion with labour income, *Journal of Public Economics*, 26 (1), pp. 19-34.

Cowell F. A. (1990). *Cheating the Government : The Economics of Evasion*. Cambridge (MA) : MIT Press.

Cribari-Neto F., Zarkos S. G. (2003). Econometric and Statistical Computing Using Ox, *Computational Economics*, 21 (3), pp. 277-295.

Crifo P., Rullière J.-L. (2004). Incentives and Anonymity Principle : Crowding Out Toward Users, *Center For Economis Studies Institute for Economic Research (CESifo) WP*, 1316.

Crozier M. (1963). *Le phénomène bureaucratique*. Paris : Editions du Seuil.

Cule M., Fulton M. (2005). Some implications of the unofficial economy-bureaucratic corruption relationship in transition countries, *Economics Letters*, 89 (2), pp. 207-211.

Culyer A. J., Evans R. G. (1996). Mark Pauly on welfare economics : Normative rabbits from positive hats, *Journal of Health Economics*, 15 (2), pp. 243-251.

Cummings R. G., Martinez-Vazquez J., McKee M., Torgler B. (2005). Effects of Tax Morale on Tax Compliance : Experimental and Survey Evidence, *Centre for Research in Economics, Management and the Arts WP*, 29.

Cunningham R. (2004). Professionalism Reconsidered : Physician Payment In A Small-Practice Environment, *Health Affairs*, 23 (6), pp. 36-48.

Damania R., Fredriksson P. G., List J. A. (2003). Trade liberalization, corruption, and environmental policy formation : theory and evidence, *Journal of Environmental Economics and Management*, 46 (3), pp. 490-512.

Danthine J.-P., Kurmann A. (2005). The Macroeconomic Consequences of Reciprocity in Labor Relations, *Université de Lausanne, Département d'Econométrie et Economie Politique WP*, 05-08.

D'Aspremont C., Dos Santos Ferreira R., Gérard-Varet L.-A. (2003). Competition for market share or for market size : Oligopolistic equilibria with varying competitive toughness, *Center for Operation Research and Econometrics DP*, 10.

De Jaegher K., Jegers M. (2000). A model of physician behaviour with demand inducement, *Journal of Health Economics*, 19 (2), pp. 231-258.

Deber R., Narine L., Baranek P., Sharpe N., Duvalko K. M., Zlotnik-Shaul R., Coyte P., Pink G., Williams P. (1998). The Public-Private Mix in Health Care, *in* National Forum on Health (Ed.), *Striking a balance : health care systems in Canada and elsewhere*, pp. 423-545. Sainte-Foy (Qc) : Editions MutliMondes.

Delattre E., Dormont B. (2003). Fixed fees and physician-induced demand : A panel data study on French physicians, *Health Economics*, 12 (9), pp. 741-754.

Demange G., Geoffard P.-Y. (2002). Reforming incentive schemes under political constraints : The physician agency, *DELTA WP*, 2002-14.

Demougin D., Helm C. (2005). Moral Hazard and Bargaining Power, *Mimeo*.

Di Tella R., Schargrodsky E. (2003a). Controlling corruption through high wages, *in* Transparency International (Ed.), *Global Corruption Report 2003*, pp. 377-379. London : Robin Hodess.

Di Tella R., Schargrodsky E. (2003b). The Role of Wages and Auditing during a Crackdown on Corruption in the City of Buenos Aires, *Journal of Law & Economics*, 46 (1), pp. 269-292.

Dickinson D., Villeval M.-C. (2004). Does Monitoring Decrease Work Effort ? The Complementarity Between Agency and Crowding-Out Theories, *IZA DP*, 1222.

Diewert W. E. (1993). Laspeyres, Ernst Louis Etienne, *in* W. E. Diewert and A. O. Nakamura (Eds.), *Essays in Index Number Theory*, Vol. I, pp. 69-70. Amsterdam : North-Holland.

Dollar D., Fisman R., Gatti R. (2001). Are women really the "fairer" sex ? Corruption and women in government, *Journal of Economic Behavior & Organization*, 46 (4), pp. 423-429.

Doornik J. A., Ooms M. (2001). *Introduction to Ox : An Object-Oriented Matrix Language*. London : Timberlake Consultants Press.

Doornik J. A., Shephard N., Hendry D. F. (2004). Parallel Computation in Econometrics : A Simplified Approach, *University of Oxford, Nuffield College WP*, W16.

Dranove D. (1988). Demand Inducement And The Physician-Patient Relationship, *Economic Inquiry*, 26 (2), pp. 281-298.

Dreher A., Herzfeld T. (2005). The Economic Costs of Corruption : A Survey and New Evidence, *Mimeo*.

Dudley R. A., Miller R. H., Korenbrot T. Y., Luft H. S. (1998). The Impact of Financial Incentives on Quality of Health Care, *Milbank Quarterly*, 76 (4), pp. 649-686.

Dufwenberg M., Gneezy U. (2000). Price competition and market concentration : an experimental study, *International Journal of Industrial Organization*, 18 (1), pp. 7-22.

Edlefsen L. E. (1981). The Comparative Statics of Hedonic Price Functions and Other Nonlinear Constraints, *Econometrica*, 49 (6), pp. 1501-1520.

Elberfeld W., Wolfstetter E. (1999). A dynamic model of Bertrand competition with entry, *International Journal of Industrial Organization*, 17 (4), pp. 513-525.

Emery J. C. H., Auld M. C., Lu M. (1999). Paying for physician services in Canada : The institutional, historical, and policy contexts, *Institute of Health Economics WP*, 99-06.

Encinosa W. E. I., Gaynor M., Rebitzer J. B. (1997). The Sociology of Groups and the Economics of Incentives : Theory and Evidence on Compensation Systems, *NBER WP*, 5953.

Euwals R., van Soest A. (1999). Desired and actual labour supply of unmarried men and women in the Netherlands, *Labour Economics*, 6 (1), pp. 95-118.

Evans R. G. (1974). Modeling the economic objectives of the physician, *in* R. Fraser (Ed.), *Health economics symposium, Proceedings of the First Canadian Conference 4-6 Sept.*, pp. 33-45.

Evans R. G. (1983). Health Care in Canada Patterns of Funding and Regulation, *Journal of Health Politics, Policy and Law*, 8 (1), pp. 1-43.

Evans R. G., Parish E. M. A., Sully F. (1973). Medical Productivity, Scale Effects, and Demand Generation, *Canadian Journal of Economics*, 6 (3), pp. 376-393.

Falk A., Fehr E., Fischbacher U. (2005). Driving Forces Behind Informal Sanctions, *Econometrica*, 73 (6), pp. 2017-2030.

Falk A., Fischbacher U. (2002). "Crime" in the lab-detecting social interaction, *European Economic Review*, 46 (4-5), pp. 859-869.

Falk A., Ichino A. (2005). Clean Evidence on Peer Effects, *Journal of Labor Economics*, Forthcoming.

Falk A., Kosfeld M. (2004). Distrust - The Hidden Cost of Control, *IZA Discussion Paper*, 1203.

Feess E., Heesen E. (2002). Self-Reporting and Ex Post Asymmetric Information, *Journal of Economics*, 77 (2), pp. 141-153.

Feess E., Walzl M. (2003). Corporate leniency programs in the EU and the USA, *German WP in Law and Economics*, 2003 (24).

Feess E., Walzl M. (2004). Self-reporting in Optimal Law Enforcement when there are Criminal Teams, *Economica*, 71 (283), pp. 333-348.

Fehr E., Falk A., Fischbacher U. (2000). Testing Theories of Fairness - Intentions Matter, *University of Zürich, Institute for Empirical Research in Economics (IEW) WP*, 63.

Fehr E., Gächter S. (2000a). Cooperation and Punishment in Public Goods Experiments, *American Economic Review*, 90, pp. 980-994.

Fehr E., Gächter S. (2000b). Fairness and Retaliation : The Economics of Reciprocity, *Journal of Economic Perspectives*, 14 (3), pp. 159-181.

Fehr E., Gächter S. (2002). Do Incentive Contracts Crowd Out Voluntary Cooperation ?, *University of Zürich, Institute for Empirical Research in Economics (IEW) WP*, 34.

Fehr E., Gächter S., Kirchsteiger G. (1997). Reciprocity as a Contract Enforcement Device : Experimental Evidence, *Econometrica*, 65 (4), pp. 833-860.

Fehr E., Kirchsteiger G., Riedl A. (1993). Does Fairness Prevent Market Clearing ? An Experimental Investigation, *Quarterly Journal of Economics*, 108 (2), pp. 437-459.

Fehr E., Schmidt K. (2002). Theories of Fairness and Reciprocity - Evidence and Economic Applications, *in* M. Dewatripont, L. Hansen and S. Turnovsky (Eds.), *Advances in Economics and Econometrics - Eighth World Congress*. Cambridge (MA) : Cambridge University Press.

Fehr E., Schmidt K. M. (2004). Fairness and Incentives in a Multi-task Principal-Agent Model, *Scandinavian Journal of Economics*, 106 (3), pp. 453-474.

Feinstein J. S. (1999). Approaches for Estimating Noncompliance : Examples from Federal Taxation in the United States, *Economic Journal*, 109 (456), pp. F360-F369.

Feldman R., Sloan F. (1988). Competition Among Physicians, Revisited, *Journal of Health Politics, Policy and Law*, 13 (2), pp. 239-262.

Feldstein M. (1995). The Effect of Marginal Tax Rates on Taxable Income : A Panel Study of the 1986 Tax Reform Act, *Journal of Political Economy*, 103 (3), pp. 551-572.

Feldstein M. S. (1970). The Rising Price of Physician's Services, *Review of Economic Statistics*, 52 (2), pp. 121-133.

Ferrall C., Gregory A. W., Tholl W. G. (1998). Endogenous Work Hours and Practice Patterns of Canadian Physicians, *Canadian Journal of Economics*, 31 (1), pp. 1-27.

Fouraker L., Siegel S. (1963). *Bargaining Behavior*. New York (NJ) : McGraw-Hill.

Franciosi R., Kujal P., Michelitsch R., Smith V., Deng G. (1995). Fairness : Effect on Temporary and Equilibrium Prices in Posted-Offer Markets, *Economic Journal*, 105 (431), pp. 938-950.

Frank B., Schulze G. G. (2000). Does economics make citizens corrupt ?, *Journal of Economic Behavior & Organization*, 43 (1), pp. 101-113.

Franzoni L. A. (2004). Discretion in Tax Enforcement, *Economica*, 71 (283), pp. 369-389.

Frey B. S. (1993). Does monitoring increase work effort ? The rivalry with trust and loyalty, *Economic Inquiry*, 31 (4), pp. 663-670.

Frey B. S., Oberholzer-Gee F. (1997). The Cost of Price Incentives : An Empirical Analysis of Motivation Crowding- Out, *American Economic Review*, 87 (4), pp. 746-755.

Friedman E., Johnson S., Kaufmann D., Zoido-Lobaton P. (2000). Dodging the grabbing hand : the determinants of unofficial activity in 69 countries, *Journal of Public Economics*, 76 (3), pp. 459-493.

Friedman J. W. (1971). A Non-cooperative Equilibrium for Supergames, *Review of Economic Studies*, 38 (1), pp. 1-12.

Fuchs V. R. (1986). Physician-induced demand : A parable, *Journal of Health Economics*, 5 (4), pp. 367.

Garicano L., Palacios I., Prendergast C. (2005). Favoritism Under Social Pressure, *Review of Economics and Statistics*, 87 (2), pp. 208-216.

Garoupa N. (1999). Optimal Law Enforcement and Criminal Organization, *Universitat Pompeu Fabra, Department of Economics and Business WP*, 366.

Garoupa N., Klerman D. (2004). Corruption and the optimal use of nonmonetary sanctions, *International Review of Law and Economics*, 24 (2), pp. 219-225.

Gaynor M., Gertler P. (1995). Moral Hazard and Risk Spreading in Partnerships, *Rand Journal of Economics*, 26 (4), pp. 591-613.

Gërxhani K. (2004). The Informal Sector in Developed and Less Developed Countries : A Literature Survey, *Public Choice*, 120 (3-4), pp. 267-300.

Glazer J., McGuire T. G. (1993). Should physicians be permitted to 'balance bill' patients ?, *Journal of Health Economics*, 12 (3), pp. 239-258.

Gneezy U., Rustichini A. (2000). Pay enough or don't pay at all, *Quarterly Journal of Economics*, 115 (3), pp. 791-201.

Goel R. K., Nelson M. A. (2005). Economic freedom versus political freedom : cross-country influences on corruption, *Australian Economic Papers*, 44 (2), pp. 121-133.

Gosden T., Forland F., Kristiansen I. S., Sutton M., Leese B., Giuffrida A., Sergison M., P (2001). Impact of payment method on behaviour of primary care physicians : a systematic review, *Journal of Health Services Research and Policy*, 6 (1), pp. 44-55.

Gosden T., Pedersen L., Torgerson D. (1999). How should we pay doctors ? A systematic review of salary payments and their effect on doctor behaviour, *Quarterly Journal of Medicine*, 92 (1), pp. 47-55.

Gosden T., Sibbald B., Williams J., Petchey R., Leese B. (2003). Paying doctors by salary : a controlled study of general practitioner behaviour in England, *Health Policy*, 64 (3), pp. 415-423.

Gottfredson M. R., Hirshi T. (1990). *A General Theory of Crime*. Stanford (CA) : Stanford University Press.

Goudie A. W., Stasavage D. (1998). A framework for the analysis of corruption, *Crime, Law and Social Change*, 29 (2-3), pp. 113-159.

Gourieroux C., Monfort A. (1993). Simulation-based inference : A survey with special

reference to panel data models, *Journal of Econometrics*, 59 (1-2), pp. 5-33.

Greene W. (2004). The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects, *Econometrics Journal*, 7 (1), pp. 98-119.

Grossman S. J., Hart O. D. (1983). An Analysis of the Principal-Agent Problem, *Econometrica*, 51 (1), pp. 7-46.

Gruber J., Owings M. (1996). Physician Financial Incentives and Cesarean Section Delivery, *Rand Journal of Economics*, 27 (1), pp. 99-123.

Grytten J., Sørensen R. (2001). Type of contract and supplier-induced demand for primary physicians in Norway, *Journal of Health Economics*, 20 (3), pp. 379-393.

Hackner J. (1996). Optimal symmetric punishments in a Bertrand differentiated products duopoly, *International Journal of Industrial Organization*, 14 (5), pp. 611-630.

Hamaguchi Y., Kawagoe T. (2005). An Experimental Study of Leniency Programs, *Research Institute of Economy, Trade and Industry DP*, 05-E-003.

Hanoch G. (1980). Hours and Weeks in a Theory of Labor Supply, *in* J. P. Smith (Ed.), *Female Labor Supply : Theory and Estimation*, pp. 119-165. Princeton (NJ) : Princeton University Press.

Haque N., Sahay R. (1996). Do Government Wage Cuts Close Budget Deficits ? Costs of Corruption, *IMF Staff Papers*, 43, pp. 754-778.

Harrison G. W., List J. A. (2005). What Constitutes a Field Experiment in Economics ?, *Mimeo*.

Harsanyi J. C. (1967). Games with Incomplete Information Played by "Bayesian" Players, I-III. Part I. The Basic Model, *Management Science*, 14 (3), pp. 159-182.

Harsanyi J. C. (1968a). Games with Incomplete Information Played by "Bayesian" Players, I-III. Part II. Bayesian Equilibrium Points, *Management Science*, 14 (5), pp. 320-334.

Harsanyi J. C. (1968b). Games with Incomplete Information Played by "Bayesian" Players, I-III. Part III. The Basic Probability Distribution of the Game, *Management Science*, 14 (7), pp. 486-502.

Hauk E., Saez-Marti M. (2002). On the Cultural Transmission of Corruption, *Journal of Economic Theory*, 107 (2), pp. 311-335.

Hausman J. A. (1979). The econometrics of labor supply on convex budget sets, *Economics Letters*, 3 (2), pp. 171-174.

Hausman J. A. (1980). The effect of wages, taxes, and fixed costs on women's labor force participation, *Journal of Public Economics*, 14 (2), pp. 161-194.

Hausman J. A. (1985). The Econometrics of Nonlinear Budget Sets, *Econometrica*, 53 (6), pp. 1255-1282.

Hausman J. A., Taylor W. E. (1981). Panel Data and Unobservable Individual Effects, *Econometrica*, 49 (6), pp. 1377-1398.

Heckman J. J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System, *Econometrica*, 46 (4), pp. 931-959.

Heckman J. J. (1981). Statistical models for discrete panel data, *in* C. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, Ch. 3, pp. 114-178. Cambridge (MA) : MIT Press.

Heckman J. J. (1997). Instrumental Variables : A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations, *Journal of Human Resources*, 32 (3), pp. 441-462.

Heckman J. J., Smith J. A. (1995). Assessing the Case for Social Experiments, *Journal of Economic Perspectives*, 9 (2), pp. 85-110.

Helpman E., Laffont J.-J. (1975). On moral hazard in general equilibrium theory, *Journal of Economic Theory*, 10 (1), pp. 8-23.

Hemenway D., Killen A., Cashman S. B., Parks C. L., Bicknell W. J. (1990). Physicians' responses to financial incentives. Evidence from a for-profit ambulatory care center, *New England Journal of Medicine*, 322 (15), pp. 1059-1063.

Hennig-Schmidt H., Rockenbach B., Sadrieh A. (2005). In Search of Workers' Real Effort Reciprocity - A Field and a Laboratory Experiment, *Governance and the Efficiency of economic SYstems DP*, 55.

Herzfeld T., Weiss C. (2003). Corruption and legal (in)effectiveness : an empirical investigation, *European Journal of Political Economy*, 19 (3), pp. 621-632.

Hines J. R. J. (1995). Forbidden Payment : Foreign Bribery and American Business After 1977, *NBER WP*, 5266.

Holmstrom B. (1982). Moral Hazard in Teams, *Bell Journal of Economics*, 13 (2), pp. 324-340.

Holmstrom B., Milgrom P. (1991). Multitask Principal-Agent Analyses : Incentive Contracts, Asset Ownership, and Job Design, *Journal of Law, Economics, & Organization*, 7 (3), pp. 24-52.

Holt C. A. (1995). Industrial Organization : A survey of Laboratory Research, *in* J. H. Kagel and A. E. Roth (Eds.), *The Handbook of Experimental Economics*, pp. 349-444. Princeton (NJ) : Princeton Uninversity Press.

Honoré B., Vella F., Verbeek M. (2005). Attrition, Selection Bias and Censored Regressions in Panel Data, *in* L. S. Mátyás, Patrick (Ed.), *The Econometrics of Panel Data. A Handbook of the Theory with Applications, 3rd ed., Forthcoming*, Ch. 22.

Hoynes H. W. (1996). Welfare Transfers in Two-Parent Families : Labor Supply and Welfare Participation Under AFDC-UP, *Econometrica*, 64 (2), pp. 295-332.

Huck S., Normann H.-T., Oechssler J. (2004). Two are few and four are many : number effects in experimental oligopolies, *Journal of Economic Behavior & Organization*, 53 (4), pp. 435-446.

Hunt J., Laszlo S. (2005). Bribery : Who Pays, Who Refuses, What Are The Payoffs ?, *Mimeo*.

Hutchison B., Birch S., Hurley J., Lomas J., Stratford-Devai F. (1996). Do physician-payment mechanisms affect hospital utilization ? A study of Health Service Organizations in Ontario, *Canadian Medical Association Journal*, 154 (5), pp. 653-661.

Innes R. (1999a). Self-policing and optimal law enforcement when violator remediation is valuable, *Journal of Political Economy*, 107 (6), pp. 1305-1325.

Innes R. (1999b). Remediation and self-reporting in optimal law enforcement, *Journal of Public Economics*, 72 (3), pp. 379-393.

Jain A. K. (2001). Corruption : A Review, *Journal of Economic Surveys*, 15 (1), pp. 71-121.

Jencks S. F., Cuerdon T., Burwen D. R., Fleming B., Houck P. M., Kussmaul A. E., Nilas (2000). Quality of Medical Care Delivered to Medicare Beneficiaries : A Profile at State and National Levels, *Journal of the American Medical Association*, 284 (13), pp. 1670-1676.

Jensen M. C. (1983). Organization Theory and Methodology, *Accounting Review*, 58 (2), pp. 319-339.

Johnson S., Kaufmann D., McMillan J., Woodruff C. (2000). Why do firms hide ? Bribes and unofficial activity after communism, *Journal of Public Economics*, 76 (3), pp. 495-520.

Jung Y. H., Snow A., Trandel G. A. (1994). Tax evasion and the size of the underground economy, *Journal of Public Economics*, 54 (3), pp. 391-402.

Kalb G., Williams J. (2003). Delinquency and gender, *Applied Economics Letters*, 10 (7), pp. 425-429.

Kaplow L., Shavell S. (1994). Optimal Law Enforcement with Self-Reporting of Behavior, *Journal of Political Economy*, 102 (3), pp. 583-606.

Kaufmann D. (1997). Corruption : The facts, *Foreign Policy*, 107, pp. 114-110.

Keane M., Moffitt R. (1998). A Structural Model of Multiple Welfare Program Participation and Labor Supply, *International Economic Review*, 39 (3), pp. 553-589.

Keser C., Willinger M. (2000). Principals' principles when agents' actions are hidden, *International Journal of Industrial Organization*, 18 (1), pp. 163-185.

Kiefe C. I., Allison J. J., Williams O. D., Person S. D., Weaver M. T., Weissman N. W. (2001). Improving Quality Improvement Using Achievable Benchmarks For Physician Feedback : A Randomized Controlled Trial, *Journal of the American Medical Association*, 285 (22), pp. 2871-2879.

Kirchsteiger G., Prat A. (1999). Common Agency and Computational Complexity : Theory and Experimental Evidence, *Tilburg University, Center for Economic Research (CentER) WP*, 36.

Klitgaard R. (1988). *Controlling Corruption*. Berkeley (CA) : University of California Press.

Klochko M. A., Ordeshook P. C. (2003). Corruption, Cooperation and Endogenous Time Discount Rates, *Public Choice*, 115 (3-4), pp. 259-283.

Konrad K. A., Skaperdas S. (1997). Credible threats in extortion, *Journal of Economic Behavior & Organization*, 33 (1), pp. 23-39.

Konrad K. A., Skaperdas S. (1998). Extortion, *Economica*, 65 (260), pp. 461-477.

Kralj B., Kantarevic J., Weinkauf D. (2005). 'Taxing' Doctors : The Impact of Income Caps on the Provision of Medical Services, *IZA DP*, 1784.

Kreps D. M. (1997). Intrinsic Motivation and Extrinsic Incentives, *American Economic Review*, 87 (2), pp. 359-364.

Kreps D. M., Milgrom P., Roberts J., Wilson R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma, *Journal of Economic Theory*, 27 (2), pp. 245-252.

Kugler M., Verdier T., Zenou Y. (2005). Organized crime, corruption and punishment, *Journal of Public Economics*, Forthcoming.

Labelle R., Stoddart G., Rice T. (1994a). A re-examination of the meaning and importance of supplier-induced demand, *Journal of Health Economics*, 13 (3), pp. 347-368.

Labelle R., Stoddart G., Rice T. (1994b). Editorial : Response to Pauly on a re-examination of the meaning and importance of supplier-induced demand, *Journal of Health Economics*, 13 (4), pp. 491-494.

Laffont J. J. (2000). *Incentives and Political Economy*. New-York (NJ) : Oxford University Press.

Laffont J.-J., Martimort D. (2002). *The Theory of Incentives : The Principal-Agent Model*. Princeton (NJ) : Princeton University Press.

Lambertini L., Sasaki D. (2002). Non-Negative Quantity Constraints and the Duration of Punishment, *Japanese Economic Review*, 53 (1), pp. 77-93.

Lambert-Mogiliansky A. (2002). Why firms pay occasional bribes : the connection economy, *European Journal of Political Economy*, 18 (1), pp. 47-60.

Lambsdorff J. G. (1999). Corruption in Empirical Research - A Review, *Transparency International WP*.

Lambson V. E. (1987). Optimal Penal Codes in Price-setting Supergames with Capacity Constraints, *Review of Economic Studies*, 54 (3), pp. 385-397.

Laussel D., Le Breton M. (2001). Conflict and Cooperation : The Structure of Equilibrium Payoffs in Common Agency, *Journal of Economic Theory*, 100 (1), pp. 93-128.

Lazear E. P. (1995). *Personnel Economics*. Cambridge (MA) : MIT Press.

Lazear E. P. (1999). Personnel Economics : Past Lessons and Future Directions, *Journal of Labor Economics*, 17 (2), pp. 199-236.

Lazear E. P. (2000a). Performance Pay and Productivity, *American Economic Review*, 90 (5), pp. 1346-1361.

Lazear E. P. (2000b). The Future of Personnel Economics, *Economic Journal*, 110 (467), pp. F611-F639.

Lemieux T., Fortin B., Frechette P. (1994). The Effect of Taxes on Labor Supply in the Underground Economy, *American Economic Review*, 84 (1), pp. 231-254.

Levaggi R., Rochaix L. (2003). Optimal payment schemes for physicians, *Portuguese Economic Journal*, 2 (2), pp. 87-107.

Levin M., Satarov G. (2000). Corruption and institutions in Russia, *European Journal of Political Economy*, 16 (1), pp. 113-132.

Levitt S., Miles T. (2005). Empirical Study of Public Enforcement of Law, *in* A. M. Polinsky and S. Shavell (Eds.), *Handbook of law and economics*, Ch. 7. Amsterdam : North-Holland.

Loayza N. V. (1996). The economics of the informal sector : a simple model and some empirical evidence from Latin America, *World Bank Policy Research WP*, 1727, pp. 129-162.

Lui F. T. (1985). An Equilibrium Queuing Model of Bribery, *Journal of Political Economy*, 93 (4), pp. 760-781.

Lui F. T. (1986). A dynamic model of corruption deterrence, *Journal of Public Economics*, 31 (2), pp. 215-236.

Ma C.-T. A. (1994). Health care payments systems : cost and quality incentives, *Journal of Economics & Management Strategy*, 3 (1), pp. 93-112.

Ma C.-T. A., McGuire T. G. (1997). Optimal Health Insurance and Provider Payment, *American Economic Review*, 87 (4), pp. 685-704.

Macho-Stadler I., Pérez-Castrillo J. D. (2001). *An Introduction to the Economics of Information*. New-York (NJ) : Oxford University Press.

MaCurdy T., Green D., Paarsch H. (1990). Assessing Empirical Approaches for Analyzing Taxes and Labor Supply, *Journal of Human Resources*, 25 (3), pp. 415-490.

Magill M., Quinzii M. (2005). An Equilibrium Model of Managerial Compensation, *Institute of Economic Policy Research WP*, 05 (22).

Malin J. L., Keating N. L. (2005). The Cost-Quality Trade-Off : Need for Data Quality Standards for Studies That Impact Clinical Practice and Health Policy, *Journal of Clinical Oncology*, 23 (21), pp. 4581-4584.

Manion M. (1996). Corruption by design : bribery in Chinese enterprise licensing, *Journal of Law, Economics & Organization*, 12 (1), pp. 167-195.

Manion M. (1998). Correction to 'corruption by design', *Journal of Law, Economics & Organization*, 14 (1), pp. 180-182.

Marjit S., Mukherjee V., Mukherjee A. (2000). Harassment, corruption and tax policy, *European Journal of Political Economy*, 16 (1), pp. 75-94.

Marjit S., Mukherjee V., Mukherjee A. (2003). Harassment, corruption and tax policy : reply, *European Journal of Political Economy*, 19 (4), pp. 899-900.

Marjit S., Shi H. (1998). On controlling crime with corrupt officials, *Journal of Economic Behavior & Organization*, 34 (1), pp. 163-172.

Mason C. F., Phillips O. R. (2002). In Support of Trigger Strategies : Experimental Evidence from Two-Person Noncooperative Games, *Journal of Economics & Management Strategy*, 11 (4), pp. 685-716.

Mauro P. (1995). Corruption and Growth, *Quarterly Journal of Economics*, 110 (3), pp. 681-712.

McFadden D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior, *in* P. Zarembka (Ed.), *Frontiers in Econometrics*, Ch. 4, pp. 105-142. New York (NJ) : New York Academic Press.

McFadden D. L. (1978). Modelling the Choice of Residential Location, *in* A. Karlkvist (Ed.), *Spatial Interaction Theory and Residential Location*, pp. 75-96. Amsterdam : North Holland.

McGlynn E. A., Asch S. M., Adams J., Keesey J., Hicks J., DeCristofaro A., Kerr E. A. (2003). The Quality of Health Care Delivered to Adults in the United States, *New England Journal of Medicine*, 348 (26), pp. 2635-2645.

McGuire T. G. (2000). Physician Agency, *in* A. J. Culyer and J. P. Newhouse (Eds.), *Handbook of Health Economics*, Vol. 1A, Ch. 9, pp. 461-536. Amsterdam : North-Holland.

McGuire T. G., Pauly M. V. (1991). Physician response to fee changes with multiple payers, *Journal of Health Economics*, 10 (4), pp. 385-410.

McKenna C. (2002). GPs vote "yes" for new contract framework, *BMJ*, 325 (7356), pp. 119.

McNeil B. J. (2001). Hidden Barriers to Improvement in the Quality of Care, *New England Journal of Medicine*, 345 (22), pp. 1612-1620.

Meidinger C., Rullière J.-L., Villeval M.-C. (2003). Does Team-Based Compensation Give Rise to Problems When Agents Vary in Their Ability ?, *Experimental Economics*, 6 (3), pp. 253-272.

Meyer B. D., Heim B. T. (2003). Structural Labor Supply Models when Budget Constraints are Nonlinear, *Mimeo*.

Mirrlees J. A. (1975). The Theory of Moral Hazard and Unobservable Behaviour : Part I, *Review of Economic Studies*, 66 (1), pp. 3-21 (Publié en 1999).

Mirrlees J. A. (1997). Information and Incentives : The Economics of Carrots and Sticks, *Economic Journal*, 107 (444), pp. 1311-1329.

Mishra A. (2002). Hierarchies, incentives and collusion in a model of enforcement, *Journal of Economic Behavior & Organization*, 47 (2), pp. 165-178.

Mocan H. N., Rees D. I. (2005). Economic Conditions, Deterrence and Juvenile Crime : Evidence from Micro Data, *American Law and Economics Review*, 7 (2), pp. 319-349.

Mocan N. H. (2004). What Determines Corruption ? International Evidence from Micro Data, *NBER WP*, W10460.

Møllgaard P. (2002). Must Trust Bust ?, *Copenhagen Business School, Department of Economics WP*, 02-2002.

Mookherjee D., Png I. P. L. (1992). Monitoring vis-a-vis Investigation in Enforcement of Law, *American Economic Review*, 82 (3), pp. 556-565.

Mookherjee D., Png I. P. L. (1995). Corruptible Law Enforcers : How Should They Be Compensated ?, *Economic Journal*, 105 (428), pp. 145-159.

Mookherjee D., Ray D. (2002). Contractual Structure and Wealth Accumulation, *American Economic Review*, 92 (4), pp. 818-849.

Motchenkova E. (2004a). Effects of Leniency Programs on Cartel Stability, *Tilburg University, Center for Economic Research (CentER) DP*, 98.

Motchenkova E. (2004b). Determination of Optimal Penalties for Antitrust Violations in a Dynamic Setting, *Tilburg University, Tilburg Law and Economics Center DP*, 2004-19.

Motchenkova E., Kort P. M. (2004). Analysis of the Properties of Current Penalty Schemes for Violations of Antitrust Law, *Journal of Optimization Theory and Applications*, Forthcoming.

Motta M. (2004). *Competition Policy. Theory and Practice*. Cambridge (MA) : Cambridge University Press.

Motta M., Polo M. (2003). Leniency programs and cartel prosecution, *International Journal of Industrial Organization*, 21 (3), pp. 347-379.

Mueller D. C. (2003). *Public Choice III*. Cambridge (UK) : Cambridge University Press.

Mullainathan S., Thaler R. H. (2001). Behavioral Economics, *in* N. J. Smelser and P. B. Baltes (Eds.), *International Encyclopedia of the Social &Behavioral Sciences*, pp. 1094-1100. Oxford (UK) : Elsevier.

Myerson R. B. (2004). Comments on "Games with Incomplete Information Played by 'Bayesian' Players, I-III", *Management Science*, 50 (12), pp. 1818-1824.

Nassiri A., Rochaix-Ranson L. (2000). L'offre de services médicaux analyse sur données de panel d'une expérience naturelle au Québec, *Revue d'Economie Politique*, (4), pp. 541-570.

Normann H., Wallace B. (2005). The Impact of the Termination Rule in Cooperation Experiments, *Royal Holloway, University of London DP*, 04/11.

Nyffeler R. (2005). Different Modeling Strategies for Discrete Choice Models of Female Labour Supply : Estimates for Switzerland, *Universität Bern, Department of Economics DP*, 05-08.

Olson M. (1978). *Logique de l'action collective*. Paris : Presses Universitaires de France.

Osborne M. J., Rubinstein A. (1994). *A Course in Game Theory*. Cambridge (MA) : MIT Press.

Paarsch H. J., Shearer B. S. (1999). The Response of Worker Effort to Piece Rates : Evidence from the British Columbia Tree-Planting Industry, *Journal of Human Resources*, 34 (4), pp. 643-667.

Paldam M. (2001). Corruption and Religion. Adding to the economic model, *Kyklos*, 54 (2-3), pp. 383-414.

Paldam M. (2002). The cross-country pattern of corruption : economics, culture and the seesaw dynamics, *European Journal of Political Economy*, 18 (2), pp. 215-240.

Pauly M. V. (1994a). Editorial : A re-examination of the meaning and importance of supplier-induced demand, *Journal of Health Economics*, 13 (3), pp. 369-372.

Pauly M. V. (1994b). Reply to Roberta Labelle, Greg Stoddart and Thomas Rice, *Journal of Health Economics*, 13 (4), pp. 495-496.

Pencavel J. (1986). Labor supply of men : A survey, *in* O. C. Ashenfelter and R. Layard (Eds.), *Handbook of Labor Economics*, Vol. 1, Ch. 1, pp. 3-102. Amsterdam : North-Holland.

Pitchford R. (1998). Moral hazard and limited liability : The real effects of contract bargaining, *Economics Letters*, 61 (2), pp. 251-259.

Plott C. R. (1982). Industrial Organization Theory and Experimental Economics, *Journal of Economic Literature*, 20 (4), pp. 1485-1527.

Polinsky A. M., Shavell S. (2001). Corruption and optimal law enforcement, *Journal of Public Economics*, 81 (1), pp. 1-24.

Prat A., Rustichini A. (1998). Sequential common agency, *Tilburg University, Center for Economic Research (CentER) DP*, 9895.

Prat A., Rustichini A. (2003). Games played through agents, *Econometrica*, 71 (4), pp. 989-1026.

Prendergast C. (1996). What Happens Within Firms ? A Survey of Empirical Evidence on Compensation Policies, *NBER WP*, 5802.

Prendergast C. (1999). The Provision of Incentives in Firms, *Journal of Economic Literature*, 37 (1), pp. 7-63.

Prendergast C. (2000a). What Trade-Off of Risk and Incentives ?, *American Economic Review*, 90 (2), pp. 421-25.

Prendergast C. (2000b). Investigating Corruption, *World Bank Policy Research WP*, 2500.

Prendergast C. (2002a). Uncertainty and incentives, *Journal of Labor Economics*, 20 (2), pp. S115-S137.

Prendergast C. (2002b). The Tenuous Trade-Off between Risk and Incentives, *Journal of Political Economy*, 110 (5), pp. 1071-1102.

Quiggin J., Chambers R. G. (2003). Bargaining power and efficiency in principal-agent relationships, *University of Queensland, Risk and Sustainable Management Group WP*, R03-1.

Rabin M. (1998). Psychology and Economics, *Journal of Economic Literature*, 36 (1), pp. 11-46.

Rauch J. E. (1991). Modelling the informal sector formally, *Journal of Development Economics*, 35 (1), pp. 33-47.

Rauch J. E., Evans P. B. (2000). Bureaucratic structure and bureaucratic performance in less developed countries, *Journal of Public Economics*, 75 (1), pp. 49-71.

Reinikka R., Svensson J. (2005). Using Micro-Surveys to Measure and Explain Corruption, *World Development*, Forthcoming.

Rey P. (2003). Towards a Theory of Competition Policy, *in* M. Dewatripont, L. P. Hansen and S. J. Turnovsky (Eds.), *Advances in Economics and Econometrics : Theory and Applications, Eight World Congress*, pp. 82-132. Cambridge (MA) : Cambridge University Press.

Rice T. (1983). The impact of changing Medicare reimbursement rates on physician-induced demand, *Medical Care*, 21, pp. 803-815.

Rice T. H., Labelle R. J. (1989). Do Physicians Induce Demand for Medical Services ?,

Journal of Health Politics, Policy and Law, 14 (3), pp. 587-601.

Rigolini J., Gatti R., Paternostro S. (2003). Individual Attitudes Towards Corruption : Do Social Effects Matter ?, *World Bank WP*, 3122.

Rizzo J. A., Blumenthal D. (1994). Physician labor supply : Do income effects matter ?, *Journal of Health Economics*, 13 (4), pp. 433-453.

Rochaix L. (1989). Information asymmetry and search in the market for physicians' services, *Journal of Health Economics*, 8 (1), pp. 53-84.

Rochaix L. (1993). Financial incentives for physicians : the Quebec experience, *Health Economics*, 2 (2), pp. 163-176.

Rogerson W. P. (1994). Choice of treatment intensity by a no-profit hospital under prospective pricing, *Journal of Economics & Management Strategy*, 3 (1), pp. 7-51.

Rose-Ackerman S. (1975). The economics of corruption, *Journal of Public Economics*, 4 (2), pp. 187-203.

Rose-Ackerman S. (1978). *Corruption : a study in political economy*. New-York (NJ) : New-York Academic Press.

Rose-Ackerman S. (2001). Trust, Honesty, and Corruption : Theories and Survey Evidence from Post-Socialist Societies., *Research Project on Honesty and Trust at Collegium Budapest, WP*.

Rosenthal M. B., Fernandopulle R., Song H. R., Landon B. (2004). Paying For Quality : Providers' Incentives For Quality Improvement, *Health Affairs*, 23 (2), pp. 127-141.

Rosenthal R. W. (1981). Games of perfect information, predatory pricing and the chain-store paradox, *Journal of Economic Theory*, 25 (1), pp. 92-100.

Rothschild R. (1999). Cartel stability when costs are heterogeneous, *International Journal of Industrial Organization*, 17 (5), pp. 717-734.

Sæther E. M. (2003). A Discrete Choice Analysis of Norwegian Physicians' Labor Supply and Sector Choice, *University of Oslo, Health Economics Research Program WP*, 19.

Sæther E. M. (2005). Wage Policies for Health Personnel, *Ph.D. Dissertation, University of Oslo, Health Economics Research Program*.

Saha B. (2003). Harassment, corruption and tax policy : a comment on Marjit, Mukherjee and Mukherjee, *European Journal of Political Economy*, 19 (4), pp. 893-897.

Salanié B. (1997). *The Economics of Contracts – A primer*. Cambridge (MA) : MIT Press.

Salanié B. (2003). Testing Contract Theory, *CESifo Economic Studies*, 49 (3), pp. 461-477.

Sanyal A. (2004). Bribes in a Supply Line, *Economica*, 71 (281), pp. 155-168.

Sappington D. E. M. (1991). Incentives in Principal-Agent Relationships, *Journal of Economic Perspectives*, 5 (2), pp. 45-66.

Schaafsma J. (1994). A new test for supplier-inducement and application to the Canadian market for dental care, *Journal of Health Economics*, 13 (4), pp. 407-431.

Schargrodsky E. (2003). Corruption perception index 2002, *in* Transparency International (Ed.), *Global Corruption Report 2003*, pp. 345-350. London : Robin Hodess.

Schneider F., Enste D. H. (2000). Shadow Economies : Size, Causes, and Consequences, *Journal of Economic Literature*, 38 (1), pp. 77-114.

Schulze G. G., Frank B. (2003). Deterrence versus intrinsic motivation : Experimental evidence on the determinants of corruptibility, *Economics of Governance*, 4 (2), pp. 143-160.

Selden T. M. (1990). A model of capitation, *Journal of Health Economics*, 9 (4), pp. 397-409.

Selten R. (1965). Spieltheoretische Behandlung eines Oligopolmodells mit Nachfragetragheit, *Zeitschrift fur disgesamte Staatswissenschaft*, 121, pp. 301-324, 667-689.

Selten R. (1973). A simple model of imperfect competition, where four are few and six are many, *International Journal of Game Theory*, 2 (1), pp. 141-201.

Selten R. (1975). Reexamination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory*, 4 (1), pp. 25-55.

Selten R., Mitzkewitz M., Uhlich G. R. (1997). Duopoly Strategies Programmed by Experienced Players, *Econometrica*, 65 (3), pp. 517-555.

Serra D. (2004). Empirical Determinants of Corruption : A Sensitivity Analysis, *Global Poverty Research Group WP*, 12.

Shapiro C., Stiglitz J. E. (1984). Equilibrium Unemployment as a Worker Discipline Device, *American Economic Review*, 74 (3), pp. 433-444.

Shapiro S. P. (2005). Agency Theory, *Annual Review of Sociology*, 31 (1), pp. 263-284.

Sharkey W. W., Sibley D. S. (1993). A Bertrand model of pricing and entry, *Economics Letters*, 41 (2), pp. 199-206.

Shearer B., Paarsch H. (2000). Piece Rates, Fixed Wages and Incentive Effects : Statistical Evidence from Payroll Records, *International Economic Review*, 41 (1), pp. 59-92.

Shekelle P. (2003). New contract for general practitioners, *BMJ*, 326 (7387), pp. 457-458.

Shleifer A., Vishny R. W. (1993). Corruption, *Quarterly Journal of Economics*, 108 (3), pp. 599-617.

Showalter M. H., Thurston N. K. (1997). Taxes and labor supply of high-income physicians, *Journal of Public Economics*, 66 (1), pp. 73-97.

Simon H. A. (1978). Rationality as Process and as Product of Thought, *American Economic Review*, 68 (2), pp. 1-16.

Skidmore M. J. (1996). Promise and Peril in Combating Corruption : Hong Kong's ICAC, *Annals of the American Academy of Political and Social Science*, 547, pp. 118-130.

Sloan F. A. (1975). Physician Supply Behavior in the Short Run, *Industrial and Labor Relations Review*, 28 (4), pp. 549-569.

Smith P. C., York N. (2004). Quality Incentives : The Case Of U.K. General Practitioners, *Health Affairs*, 23 (3), pp. 112-118.

Smith R. (2003). The failures of two contracts, *BMJ*, 326 (7399), pp. 1097-1098.

Smith V. L. (1962). An Experimental Study of Competitive Market Behavior, *Journal of Political Economy*, 70 (2), pp. 111-137.

Smith V. L. (1982). Microeconomic Systems as an Experimental Science, *American Economic Review*, 72 (5), pp. 923-955.

Sørensen R. J., Jostein G. (1999). Competition and supplier-induced demand in a health care system with fixed fees, *Health Economics*, 8 (6), pp. 497-508.

Spagnolo G. (2002). Self-Defeating Antitrust Laws : How Leniency Programs Solve Bertrand's Paradox and Enforce Collusion in Auctions, *Stockholm School of Economics, Mimeo.*

Spagnolo G. (2003). Divide et Impera. Optimal Deterrence Mechanisms Against Cartels and Organized Crime, *Fondazione Eni Enrico Mattei Note di Lavoro WP*, 42-2000.

Spulber D. F. (1995). Bertrand Competition when Rivals' Costs are Unknown, *Journal of Industrial Economics*, 43 (1), pp. 1-11.

Stevens D. E., Thevaranjan A. (2005). Is there Room Within Principal-Agent Theory for Ethics ?, *Mimeo*.

Steves F., Rousso A. (2003). Anti-corruption programmes in post-communist transition countries and changes in the business environment, 1999-2002, *European Bank for Reconstruction and Development WP*, 85.

Stigler G. J. (1970). The Optimum Enforcement of Laws, *Journal of Political Economy*, 78 (3), pp. 526-536.

Stiglitz J. E. (1975). Incentives, Risk, and Information : Notes Towards a Theory of Herarchy, *Bell Journal of Economics*, 6 (2), pp. 552-579.

Stiglitz J. E. (2000). The contributions of the economics of information to twentieth century economics, *Quarterly Journal of Economics*, 115 (4), pp. 1441-1478.

Sung H.-E. (2003). Fairer Sex or Fairer System ? Gender and Corruption Revisited, *Social Forces*, 82 (2), pp. 703-723.

Suphachalasai S. (2005). Bureaucratic Corruption and Mass Media, *University of Cambridge, Department of Land Economics, Environmental Economy and Policy Research WP*, 052005.

Swamy A., Knack S., Lee Y., Azfar O. (2001). Gender and corruption, *Journal of Development Economics*, 64 (1), pp. 25-55.

Swann C. A. (2002). Maximum Likelihood Estimation Using Parallel Computing : An Introduction to MPI, *Computational Economics*, 19 (2), pp. 145-178.

Tanzi V. (1995). Corruption : arm's-length relationships and markets, *in* G. Fiorentini and S. Peltzman (Eds.), *The Economics of Organised Crime*, pp. 161-181. Cambridge (MA) : Cambridge University Press.

Tanzi V. (1998). Corruption Around the World : Causes, Consequences, Scope, and Cures, *IMF Staff Papers*, 45 (4), pp. 559-594.

Thal J. (2004). Optimal Collusion under Cost Asymmetry, 13^{th} *WZB Conference on Markets and Political Economy.*

Tirole J. (1986). Hierarchies and Bureaucracies : On the Role of Collusion in Organizations, *Journal of Law, Economics, & Organization*, 2 (2), pp. 181-214.

Tirole J. (1988). The Multicontract Organization, *Canadian Journal of Economics*, 21 (3), pp. 459-466.

Tirole J. (1992). Collusion and the Theory of Organizations, *in* J. J. Laffont (Ed.), *Advances in Economic Theory : Sixth World Congress*, Vol. II, pp. 151-206. Cambridge (MA) : Cambridge University Press.

Tirole J. (1994a). *Théorie de l'organisation industrielle*. Paris : Economica.

Tirole J. (1994b). The Internal Organization of Government, *Oxford Economic Papers*, 46 (1), pp. 1-29.

Tirole J. (1996). A Theory of Collective Reputations (with Applications to the Persistence of Corruption and to Firm Quality), *Review of Economic Studies*, 63 (1), pp. 1-22.

Tirole J. (1999). Incomplete Contracts : Where Do We Stand ?, *Econometrica*, 67 (4), pp. 741-781.

Tirole J. (2002). Rational irrationality : Some economics of self-management, *European Economic Review*, 46 (4-5), pp. 633-655.

Toye J., Moore M. (1998). Taxation, Corruption and Reform, *European Journal of Development Research*, 10 (1), pp. 60-84.

Train K. E. (1999). Halton Sequences for Mixed Logit, *University of Berkeley, Departement of Economics Mimeo*.

Train K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge (UK) : Cambridge University Press.

Trandel G., Snow A. (1999). Progressive income taxation and the underground economy, *Economics Letters*, 62 (2), pp. 217-222.

Treisman D. (2000). The causes of corruption : a cross-national study, *Journal of Public Economics*, 76 (3), pp. 399-457.

Van Rijckeghem C., Weder B. (2001). Bureaucratic corruption and the rate of temptation : do wages in the civil service affect corruption, and by how much ?, *Journal of Development Economics*, 65 (2), pp. 307-331.

van Soest A. (1995). Structural Models of Family Labor Supply : A Discrete Choice Approach, *Journal of Human Resources*, 30 (1), pp. 63-88.

van Soest A., Das M., Gong X. (2002). A structural labour supply model with flexible preferences, *Journal of Econometrics*, 107 (1-2), pp. 345-374.

Wang X. H., Yang B. Z. (2001). Fixed and sunk costs revisited, *Journal of Economic Education*, 32 (2), pp. 178-185.

White H. (1985). Agency as control, *in* J. Pratt and R. Zeckhauser (Eds.), *Principals and agents : the structure of business*, pp. 187-212. Boston (MA) : Harvard Business School Press.

Yava C. (1998). A comment on 'corruption by design', *Journal of Law, Economics & Organization*, 14 (1), pp. 174-179.

Yellen J. L. (1984). Efficiency Wage Models of Unemployment, *American Economic Review*, 74 (2), pp. 200-205.

Zabalza A., Pissarides C., Barton M. (1980). Social security and the choice between full-time work, part-time work and retirement, *Journal of Public Economics*, 14 (2), pp. 245-276.

Zeiliger R. (2000). A presentation of Regate, Internet based Software for Experimental Economics, *http ://www.gate.cnrs.fr/ zeiliger/regate/RegateIntro.ppt, GATE.*

Liste des Tableaux

1 Structure de l'analyse 9

1.1 Effet des instruments de lutte contre la corruption 45
1.2 Fonction de production 63
1.3 Taux d'acceptation et de corruption 67
1.4 Transfert moyen en fonction du comportement de l'agent 69
1.5 Corrélation avec le transfert proposé 70
1.6 Comparaisons inter-expériences 73
1.7 Comparaisons intra-expériences 75
1.A Décision de corruption 81

2.1 Statistiques descriptives de la rémunération mixte 110
2.2 Rémunération des médecins du Québec considérés dans l'analyse 113
2.3 Statistiques descriptives de l'effet de la réforme 114
2.4 Estimateurs de Différence en Différence 117
2.5 Déterminants théoriques de l'allocation du loisir 136
2.6 Distribution de l'échantillon entre les niveaux de discrétisation 140
2.7 Prédiction de la consommation effective 158
2.8 Profil de pratique des chirurgiens 160
2.9 Paramètres estimés de la fonction d'utilité Translog 161
2.10 Qualité du modèle estimé 162
2.11 Utilités marginales 163
2.12 Variations induites par l'introduction de la rémunération mixte 164
2.13 Variations induites par une rémunération mixte obligatoire 166

3.1 Statique comparative de l'équilibre de silence collusif 221
3.2 Distribution de l'intensité de la concurrence 235
3.3 Possibilités de collusion dans chaque traitement 237
3.4 Taux d'évasion sous le traitement Contrôle 239
3.5 Distribution des prix de marché sous le traitement Contrôle 240
3.6 Taux d'évasion 241
3.7 Prix choisi moyen 242
3.8 Prix d'équilibre moyen 243
3.9 Comportement de dénonciation 244
3.10 Evasion collusive observée 246
3.11 Evasion collusive 250
3.12 Evasion collusive et coordination 253
3.A Tableau remis aux participants 264
3.B Comportements experimentaux et variables d'environnement 267
3.C Distribution de l'intensité de la concurrence au sein des traitements . . 268
3.D Possibilités de collusion dans chaque traitement 269
3.E Taux d'évasion sous le traitement Contrôle 270
3.F Distribution des prix de marché sous le traitement Contrôle 270
3.G Taux d'évasion 271
3.H Prix choisi moyen 271
3.I Prix d'équilibre moyen 272
3.J Evasion collusive observée 272
3.K Evasion collusive 273
3.L Evasion collusive et coordination 275

Liste des Graphiques

1.1 Forme séquentielle du jeu de corruption 56

1.2 Evolution du niveau de corruption au sein des traitements 75

1.A Ecran de contrôle de l'agent 86

2.1 Passage à la rémunération mixte 125

3.1 Evolution des mesures d'intensité 234

3.2 Prix d'équilibre par traitement 244

3.A Ecran de contrôle d'une firme 265

3.B Evolution des variables de décisions moyennes 266

Table des matières

Remerciements **v**

Résumé **ix**

Abstract **xi**

Sommaire **xv**

Introduction Générale **1**

1 Analyses du comportement de corruption **15**

(i) Contrat de délégation : la relation principal – agent 18

a) Détection 21

b) Salaire d'efficience 23

c) Coût moral 27

(ii) La relation principal – corrupteur ? 32

a) Corruption et relations multi-principaux 33

b) De la corruption des lois... 36

(iii) Pacte de corruption : la relation agent – corrupteur 38

a) Comportement du corrupteur 39

b) Pot-de-vin d'équilibre 42

c) Propriétés du pacte de corruption 44

(iv) Principal – agent – corrupteur : l'effet de délégation 52

1.1 Jeu de corruption à trois joueurs 54
1.1.1 Description du jeu 55
1.1.2 Hypothèses de travail 57
1.2 Protocole de l'expérience 59
1.2.1 Plan d'expériences 60
a) Expérience de Corruption (EC) 60
b) Expérience de Délégation Explicite (EDE) 61
c) Traitements 62
1.2.2 Déroulement des sessions 63
a) Paramètres 63
b) Conditions pratiques 64
1.3 Déterminants du comportement de corruption 66
1.3.1 Mise en œuvre des pactes de corruption 67
1.3.2 Effet de délégation 71
1.4 Conclusion 77
Annexes 81
1.A Regression 81
1.B Instructions de l'expérience 82
1.C Questionnaire pré-expérimental 87
1.C.1 Expérience de Corruption 87
1.C.2 Expérience de Délégation Explicite 88

2 Coût et qualité de l'offre de soins, quelle(s) rémunération(s) ? 89
2.1 Institutions : la contrainte budgétaire des médecins du Québec 97
2.1.1 Modes de rémunération et comportements de pratique 98
2.1.2 Le règne de la rémunération à l'Acte 104
2.1.3 Le mode de Rémunération Mixte 107
a) Objectifs et dispositions 107
b) Consommation potentielle sous la rémunération mixte 109
c) Réalisations : un premier aperçu 113

2.2 Analyse théorique du passage à la Rémunération Mixte 119
2.2.1 Modélisation du comportement des médecins 120
2.2.2 Quantités optimales : analyse du modèle à effort exogène 123
2.2.3 Arbitrage qualité/quantités 128
a) Effets revenu 129
b) Effets prix 131
2.3 Modèle économétrique 138
2.3.1 Modèle de choix discrétisé : éléments de base 139
2.3.2 Aspects spécifiques 144
2.4 Résultats : les vertus de la flexibilité 148
2.4.1 Présentation des données 149
2.4.2 Construction des variables 150
a) Heures de travail 150
b) Actes 152
c) Prix des actes 153
d) Plafonds et taux de rémunération 156
2.4.3 Résultats d'estimation 159
a) Préférences estimées 159
b) Simulations 164
2.5 Conclusion 167
Annexes 171
2.A Programme Ox 171
2.A.1 Lexique des variables et matrices utilisées 171
2.A.2 Programme 175

3 Demande de travail au noir en environnement concurrentiel 195
3.1 Demande de travail au noir et concurrence à la Bertrand 203
3.1.1 Cadre d'analyse 204
3.1.2 La malédiction de Bertrand 209

3.2 Dénonciation : l'équilibre de silence collusif . . . 213
3.2.1 Dénonciation et concurrence . . . 214
a) Silence collusif . . . 215
b) Silence collusif et collusion tacite . . . 218
3.2.2 Statique comparative du modèle . . . 220
a) Absence de clauses de clémence . . . 221
b) Programme actif de clémence . . . 223
3.3 Présentation des marchés expérimentaux . . . 225
3.3.1 Cadre de l'expérience . . . 225
3.3.2 Protocole expérimental . . . 227
a) Description de l'expérience . . . 227
b) Traitements . . . 228
c) Déroulement des sessions . . . 230
3.4 Malédiction de Bertrand et évasion collusive : résultats empiriques . . . 232
3.4.1 Contrepartie empirique du cadre théorique . . . 233
3.4.2 Comportements observés . . . 239
a) Traitement de contrôle : malédiction de Bertrand . . . 239
b) Statistiques descriptives . . . 241
3.4.3 Conditions de mise en œuvre de l'évasion collusive . . . 246
a) Emergence de l'évasion collusive . . . 247
b) Evasion collusive et coordination . . . 251
3.5 Conclusion . . . 256
Annexes . . . 259
3.A Instructions de l'expérience . . . 259
3.B Paramètres de l'expérience . . . 265
3.C Robustesse à une mesure alternative de la taille du marché . . . 266
3.C.1 Comportement des firmes et mesures d'intensité . . . 267
3.C.2 Description des marchés expérimentaux . . . 268
3.C.3 Mise en œuvre de l'évasion collusive . . . 272

Conclusion Générale **277**

Bibliographie **281**

Liste des Tableaux **309**

Liste des Graphiques **311**

Table des matières **313**

Résumé

La théorie de l'agence a offert une analyse approfondie des conditions sous lesquelles les incitations parviennent à réconcilier les intérêts divergents du principal et de l'agent. Les essais présentés dans cette thèse évaluent la pertinence empirique de ces résultats face à l'intervention d'une tierce partie dans trois situations-types : le comportement de corruption, les choix de pratique des médecins spécialistes et la demande de travail au noir.

D'abord, les situations de corruption correspondent à l'imbrication de deux contrats : un contrat de délégation, qui lie un Principal et un Agent ; et un pacte de corruption conclu entre cet Agent et une tierce partie, appelée Corrupteur. Nous proposons, dans un premier temps, une revue de la littérature récente consacrée à cette question mettant en parallèle les résultats existants quant aux déterminants du comportement de corruption et les propriétés de chacun de ces deux contrats. Nous montrons dans un second temps que l'existence simultanée de ces deux contrats met l'agent en face d'un conflit de réciprocités. Les résultats du jeu de corruption expérimental à trois joueurs que nous réalisons confirment l'importance de ce mécanisme. Ce conflit de réciprocités est à l'origine d'un *effet de délégation* qui constitue une explication supplémentaire à l'influence du salaire sur le comportement de corruption.

Ensuite, la gestion de l'offre de soins des médecins doit répondre à des objectifs contradictoires : le contrôle du coût du système de santé est le premier souci des autorités qui l'administrent, tandis que les patients se préoccupent principalement de la qualité des soins (en termes de santé). Pour mieux comprendre la capacité des incitations à résoudre cette contradiction, nous proposons une analyse théorique et économétrique des effets de l'introduction d'une rémunération mixte au Québec en 1999. Le comportement des médecins est décrit par leurs choix en termes de marges extensives (nombre d'heures et d'actes) et de marge intensive (temps consacré aux patients). Les résultats d'estimation mettent en évidence l'importance de la flexibilité dans les choix de rémunération.

Enfin, la demande de travail au noir est une activité illégale dont le bénéfice dépend des stratégies de marché adoptées par les firmes concurrentes. Nous proposons une analyse théorique et expérimentale de l'effet de cette particularité sur les déterminants de la demande de travail au noir. L'accent est mis, en particulier, sur l'efficacité potentielle d'un nouvel instrument de répression : la dénonciation. Nous montrons d'abord que l'intensité de la concurrence conduit à la *malédiction de Bertrand* : les firmes choisissent l'évasion mais la concurrence en élimine tout bénéfice. Nous étudions ensuite les conditions sous lesquelles un marché peut mettre en œuvre une *évasion collusive*, qui permet d'obtenir des bénéfices positifs de l'évasion. La dénonciation est une condition favorable à l'émergence de cette stratégie, et tend donc à encourager l'embauche au noir. Ces résultats, confirmés par les comportements observés, militent donc contre l'introduction de ce type d'instrument.

Dans leur ensemble, ces applications mettent en évidence le rôle central de la structure d'intérêts qu'entretiennent les joueurs en présence : radicalement divergents, convergents mais contradictoires ou divergents mais disposant d'un mécanisme de réconciliation.

Mots-clés : Incitations, Relations Principal-Agent, Activités illégales, Offre de soins de santé, Modes de rémunération, Comportement des firmes, Collusion, Econométrie appliquée, Economie expérimentale (*Codes JEL* : K42, K12, I11, I18, J33, D21, L41, C25, C91).

Abstract and Keywords : See p. xi.

Laboratoires de rattachement

Groupe d'Analyse et de Théorie Economique (GATE). UMR 5824 CNRS - Université Lumière Lyon 2 – ENS LSH – 93 chemin des Mouilles – BP 167 – 69131 Ecully Cedex – France.

Centre Interuniversitaire sur le Risque, les Politiques Economiques et l'Emploi (CIRPEE). Université Laval – Département d'économique, Pavillon J.A. de Sève – Québec (Qc) G1K 7P4 – Canada.