Essays on competition policy and applied econometrics

Louis Pape

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX025

Thèse de doctorat

Essays on Competition Policy and Applied Econometrics

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (EDIPP)
Spécialité de doctorat: Sciences économiques

Thèse présentée et soutenue à Palaiseau, le 7 avril 2022, par

Louis-Daniel Pape

Composition du Jury :

Philippe Gagnepain Professor, Paris School of Economics (CES, UMR 8174)	Président
Michael Bognanno Professor, Temple University (Economics Department)	Rapporteur
Ghazala Azmat Professor, Sciences Po (Département d'économie)	Examinateur
Ioana Marinescu Associate Professor, University of Pennsylvania (School of Social Policy)	Examinateur
Christian Belzil Professor, Ecole Polytechnique (Département d'économie)	Directeur de thèse
Michael Visser Directeur de Recherche, CNRS (CREST, UMR 9194)	Co-directeur de thèse

Acknowlegments

I would like to acknowledge and thank the plethora of individuals who participated, in one way or another, to this thesis.

First, I am very grateful to Ghazala Azmat, Michael Bognanno, Philippe Gagnepain, and Ioana Marinescu for taking the time to read, review, and comment my work, as well as serving on my committee.

I am very lucky and appreciative for my advisors, Christian Belzil and Michael Visser, who managed the difficult task of both guiding me through this thesis whilst simultaneously allowing me to develop my autonomy. Our discussions on research matters, and more generally, on the correct way to envision the role of the academic economist, have helped me to, not only become a better researcher, but also maintain my passion for the field over the years.

This thesis is comprised of articles written with several co-authors. I am forever indebted to them and sincerely grateful for having thought me the ways of economic research, for their trust in me, and for their patience. I have learned tremendously from Ioana Marinescu's energy and steady thinking, as well the time she took to show me how to write, communicate, and publish an article. It was a pleasure to work with Ivan Ouss who introduced me to the intricacies of the DADS. I would not have written this thesis without the encouragements and teachings of Alessandro Iaria which both pushed me to pursue research in industrial organization and to think critically about the essence of its most successful applications. Christophe Bellégo has taught me the virtue of perseverance and has always encouraged me to think outside the box. Finally, David Benatia has been very generous with me. Working with him has not only helped me improve my understanding of econometrics but also made me develop an authentic admiration for the immense creativity displayed in the work of econometricians.

I was fortunate to have benefited from the excellent research environment provided by CREST and Ecole Polytechnique. The quality of this environment results from the vision and dedication of its current and past directors, Arnak Dalayan, Guillaume Hollard, and Francis Kramarz, whom I would like to thank. I am particularly appreciative to the latter for having encouraged me to work with the administrative data of the CASD, as well as helping me see early on when some of my research ideas were off; saving me precious time. I have also benefited from Philippe Choné's counsel from the time of my master's thesis, along with the advice given by the other members of my Comité de Suivi, Elia Lapenta and Franck Malherbet.

I am very thankful for the support, expertise and availability of many researchers, including Marie-Laure Allain, Pierre Boyer, Pierre Cahuc, Julien Combe, Xavier d'Haultfoeuille, Laurent Davezies, Roxana Fernandez, Bertrand Garbinti, Olivier Gossner, Thierry Kamionka, Yukio Koriyama, Yves Le Yaounq, Laurent Linnemer, Jean-Baptiste Michau, Isabelle Mejean, Guy Meunier, Mathias Nunez, Roland Rathelot, Alessandro Riboni, Benoît Schmutz, Anthony Strittmatter, Alain Trognon and Arne Uhlendorff.

I would also like to thank Murielle Jules, Weronica Leduc, Audrey Lemaréchal, Eliane Nitiga-Madelaine, Lyza Racon, Arnaud Richet, Sri Srikandan, Fanda Traore, and Edith Verger for their kindness and help with logistic and administrative affairs during this thesis.

I am very fortunate to have benefited from the guidance of Thibault Vergé who whet my appetite for competition policy and encouraged me to pursue a gap year at the Autorité de la Concurrence. His support was also essential in helping me find a postdoctoral fellowship at Telecom Paris. To this end, I would like to thank Étienne Pfister for welcoming me at the Autorité de la Concurrence during this gap year. My work there, and in

particular with Nicolas Lluch and Cédric Nouël de Buzonnière, was eye opening and has deeply influenced the way I think about research in competition policy. This experience motivated me to pursue research within digital economics, and I am very lucky and grateful to be able to do so with Ulrich Laitenberger at Telecom Paris, as a postdoctoral research fellow. I look forward to our work together !

This thesis builds upon many discussions and heated arguments held with fellow Ph.D students which I would like to thank. Hoping that I have not forgotten anyone, I am thinking of Remi Avignon, Alicia Bassiere, Antoine Bertheau, Guillaume Bied, Guidogiorgio Bodrato, Léa Bou Sleiman, Clémence Bourcet, Emmanuel Bovary, Pauline Carry, Arthur Cazaubiel, Badr-Eddine Chérief-Abdellatif, Gwen-Jiro Clochard, Héloïse Cloléry, Pierre-Edouard Collignon, Morgane Cure, Andréa Epivent, Antoine Ferey, Germain Gauthier, Lucas Girard, Morgane Guignard, Etienne Guigue, Yannick Guyonvarch, Jérémy Hervelin, Tomas Jagelka, Gustave Kenedi, Raphael Lafrogne-Joussier, Alice Lapeyre, Clémence Lenoir, Claire Leroy, Pauline Leveneur, Esther Mbih, Denys Medee, Federica Meluzzi, Hugo Molina, Julien Monardo, Martin Mugnier, Sandra Nevoux, Elio Nimier-David, Félix Pasquier, Berangère Patault, Elia Pérennès, Fabien Perez, Bertille Picard, Inès Picard, Anasuya Raj, Emanuel Rapsch, Emilie Sartre, Felix Schleef, Clemence Tricaud, Jérôme Trinh, Giulia Vattuone, Benjamin Walter, Ao Wang, Tang Yuanzhe.

I salute all of the (honorary) citizens of bureau 4100, Reda Aboutajdine, Thomas Delemotte, Antoine Valtat and especially Alexis Larousse for all of our laughs and discussions over the years. I am glad to see Arnaud Châtelain become the new custodian of this amazing realm.

La décision d'entamer cette thèse doit beaucoup à des personnes exceptionnelles dont j'ai eu la chance de suivre l'enseignement au lycée de Sèvres, Anne-Marie Durand, Czelaw Michalewski, Thibault Richard et Jean-Marie Soubirou.

Je souhaite témoigner ma reconnaissance envers tous les amis qui ont patiemment écouté quelques théories balbutiantes autour de cafés et autres nectars. Je pense particulièrement à mes amis de Sèvres, de Londres et de Cambridge, ainsi que ceux rencontrés à mon retour en France, à Alésia, à Malakoff, à Montrouge et à l'association pongiste du 17e.

Les mots ne peuvent dire tout ce que je dois à ma famille, Mom, Dad, Larry, et Sam. Vos soutiens indéfectibles durant cette thèse, ainsi que vos encouragements à poursuivre la voie de l'économie, m'ont été essentiels. Enfin, je n'aurais pu accomplir cette thèse sans avoir eu à mes côtés l'extraordinaire sens de l'humour et de la vie de Laure, à qui je dédie cette thèse. Merci.

Contents

Chapter 1

Thesis Introduction and Summary

1.0.1 English Summary

This thesis includes three articles on the subject of competition policy along with an econometric contribution. Regarding the former, the two first articles focus on the relationship between antitrust enforcement and the labor market. The first measures levels of labor market concentration in France (2011-2015) through the Herfindahl–Hirschman Index (HHI) defined over local labor markets. We find a negative relationship between labor market concentration and both wages and employment. This suggests the existence of oligopsonistic and monopsonistic labor market power on behalf of employers ; potentially resulting from and influenced by competition policy. The second article quantifies the importance of antitrust laws in protecting workers' wages. We look at a mechanism allocating American baseball players, in a quasi-random way, to a competitive labor market from a highly monopsonistic market ; the difference resulting from the application of antitrust laws. We find such laws allow wages to increase by at least 30%. A third article looks at the need for regulatory oversight in the mobile phone video game industry. Exploiting both quasi-natural variation stemming from the game's structure and artificially induced variation from a Randomized Controlled Trial (RCT), we estimate demand for content on behalf of consumers using a discrete choice model. We use this model to simulate alternative pricing schemes, including individual level pricing, in order to study the potential for additional profits. We find limited evidence that higher profits can be obtained through price discrimination. Rather, revenues could be improved through a lower fixed price ; suggesting the absence of a need for regulatory oversight. Finally, this thesis includes a methodological contribution. This fourth article discusses the problem of zeros in log-linear and log-log regressions when the analyst is interested in measuring a semi-elasticity or an elasticity ; as is commonly done in empirical industrial organization, labor economics, international economics, development economics, and health economics. We show this issue to be highly relevant by measuring its relevence in the works of recent publications in the American Economic Review. We not only explain this issue, we also develop a new approach called « iterated Ordinary Least Squares ». The latter is more flexible in terms of moment conditions and easier to estimate, in particular in the context of endogenous regressors, than the more classical solution consisting in Poisson regression. Using a new statistical test based on comparing the empirical pattern of zeros in the data with the ones implied by moment-based models, we show that our model can sometimes be favored in comparison to other classical solutions, as shown by revisiting recent publications from the fields of international trade and development.

1.0.2 French Summary

Cette thèse propose trois contributions aux questions de politique de concurrence ainsi qu'une nouvelle approche économétrique. Sur la politique de concurrence, elle inclut deux études portant sur le rôle des politiques antitrust vis-à-vis du marché du travail. La première évalue les niveaux de concentration industrielle, au sens du Herfindahl–Hirschman Index (HHI), dans les marchés locaux du travail en France sur la période récente (2011-2015). A l'aide de données administratives permettant de suivre l'ensemble des flux des travailleurs entre les entreprises, cet article confirme une relation négative entre la concentration et la masse salariale. Ce travail révèle donc l'existence de rentes oligopsonistiques et monopsonistiques résultant, en partie, de la politique de concurrence. Pour étudier ce phénomène, la dernière partie de l'article simule des fusions horizontales entre les deux plus grandes entreprises de chaque secteur. On identifie ainsi des pertes d'emplois significatives dans le secteur même, ainsi que des pertes collatérales par le biais des professions subissant une réduction de la concurrence entre employeurs. Le deuxième article confirme cette relation entre politique de concurrence et salaires à travers l'étude des mécanismes de promotion salariale au sein des équipes américaines de baseball. Exploitant la soudaine et aléatoire promotion de certains joueurs pouvant accéder à un marché du travail pseudo-concurrentiel (à travers l'accès à l'arbitration salariale) à partir d'un marché initialement monopsonistique, nous montrons que ce changement dans la structure du marché permet d'expliquer une modification à la hausse des salaires d'au moins 30%. Le reste de l'article argumente que cette augmentation peut bien s'interpréter comme une perte de salaire par rapport à un marché concurrentiel. D'abord, les salaires ne changent pas de manière forte lorsque les joueurs rejoignent un marché réellement concurrentiel (free agents). Ensuite, un modèle théorique clarifie des conditions pour que l'arbitration salariale ne génère pas de distorsions par rapport à un marché du travail concurrentiel. Ainsi, nous pouvons observer que l'absence de concurrence peut aussi se régler à travers la création d'institutions arbitrales tierces, suggérant de possibles nouveaux leviers pour la politique de l'emploi. Une troisième étude pose la question de l'éventuel encadrement des pratiques de discriminations tarifaires dans le milieu du jeu vidéo sur téléphone mobile. A travers des variations quasi-naturelles au sein du jeu étudié ainsi que des exclusions résultant d'un essai randomisé contrôlé (RCT), nous pouvons affirmer que les joueurs sont myopes : ils n'anticipent pas les barrières de payements futures. A travers cette observation, nous identifions la demande de contenus ainsi que son hétérogénéité pour les joueurs, par un modèle de demande à choix discrets basé sur la théorie de l'utilité aléatoire de McFadden. Nous simulons alors différents menus de prix pour étudier leurs capacités à augmenter les profits de l'entreprise. Nous trouvons que ces pratiques n'augmenteraient que marginalement ses profits mais que celle-ci pourrait augmenter ses revenus simplement en établissant un prix fixe plus modeste. Cela suggère l'absence de besoin de régulation et ainsi que de nouvelles façons d'exploiter le contenu des jeux afin de mieux quantifier les besoins et sensibilités des consommateurs. Enfin, une contribution d'ordre méthodologique complète cette thèse. Elle soulève la question des bonnes pratiques économétriques lorsque l'analyste est confronté à des zéros dans le contexte d'une estimation d'élasticité ou semi-élasticité – cette dernière étant un objectif récurrent en économie industrielle, en économie du travail, en économie internationale, en économie du développement et en économie de la santé. Nous montrons la pertinence de la question à travers une revue bibliométrique des publications dans l'American Economic Review où nous trouvons que des chercheurs à la pointe du domaine peuvent faire appel à des simplifications méthodologiques difficiles à justifier théoriquement. Nous expliquons alors les diverses solutions existantes et compatibles avec ces données de comptage et

proposons une nouvelle approche nommée « iterated Ordinary Least Squares ». Cette dernière est plus flexible en termes de moments et plus facile à estimer (surtout avec des variables endogènes) que la méthode plus classique de la régression Poisson. De plus, la méthode est utilisable même dans le contexte de variables de contrôle nombreuses et de haute dimension, grâce à une transformation dite within. Avec l'aide d'un test statistique basé sur l'équivalence théorique entre la prévalence des zéros dans les données et celle prédite par ces modèles, nous montrons que notre modèle peut-être préféré par rapport aux méthodes classiques, à travers l'étude de données issues de publications récentes portant sur le commerce international, par exemple. Dans ce cas, la méthode Poisson suggère que les accords commerciaux n'ont aucun impact sur les flux internationaux. Notre méthode montre, au contraire, que ces derniers peuvent largement augmenter les échanges de biens et services à travers les pays.

Chapter 2

Wages, Hires, and Labor Market Concentration

This first chapter is co-written with Ioana Marinescu and Ivan Ouss. It is published in the Journal of Economic Behavior and Organization, Volume 184, April 2021, p. 506-605.

Abstract: How does employer market power affect workers? We compute the concentration of new hires by occupation and commuting zone in France using linked employer-employee data. Using instrumental variables, we find that a 10% increase in labor market concentration decreases hires by 3.2% and their hourly wage by nearly 0.5%, as hypothesized by monopsony theory. Based on a simple merger simulation, we find that a merger between the top two employers in the retail industry would be most damaging, with about 30 million euros in annual loss to the wage bill of new hires, and a 3,000 decrease in annual hires.

2.1 Introduction

How does employer market power affect workers? A burgeoning literature has shown that labor market concentration has a negative impact on wages (Azar et al., 2017a; Benmelech et al., 2018; Hershbein et al., 2018; Rinz, 2018; Lipsius, 2018; Abel et al., 2018; Martins, 2018; Qiu and Sojourner, 2019; Bassanini et al., 2020; Schubert et al., 2020; Dodini et al., 2020). From a policy perspective, this suggests that antitrust and competition authorities should scrutinize prospective mergers between two companies for their anticompetitive effects in the labor market (Marinescu and Hovenkamp, 2018; Naidu et al., 2018a; Marinescu and Posner, 2019). However, doing so requires the assessment of both wage and employment effects of consolidations. While prior literature has examined the wage effects of labor market concentration, it did not examine employment effects. Furthermore, the data used was often incomplete in terms of industries and occupations covered. Therefore, it was not possible to assess the size of the expected economy-wide wage *and* employment losses resulting from employer consolidation via mergers.

In contrast, we leverage rich administrative data on firms and workers in France to measure how increases in labor and product market concentration affect both wages and employment. More specifically, our administrative data from France includes the date, occupation, and location of all new hires. We link this data to workers' employment histories and to firm-level data. We define labor market concentration as the Herfindahl-Hirschman

Index for new hires in an occupation (4-digits), commuting zone, and quarter. We find that the mean labor market concentration in France is 0.151. Then, we run wage regressions controlling for worker and firm fixed effects, and for firm size and value added per worker. Using our estimates of the impacts of labor market concentration, we then simulate the economy-wide effects of a horizontal merger between the two largest (by employment) firms in each industry.

Our first finding concerns labor market concentration and its wage and employment impacts. In our preferred wage specification, we control for market (occupation by commuting zone), worker and firm fixed effects, and instrument labor market concentration with the inverse number of employers in other geographic markets for the same quarter and occupation, following a similar strategy to Azar et al. (2017a); Rinz (2018); Qiu and Sojourner (2019). We find that a 10% increase in labor market concentration decreases the wages for new hires by nearly 0.5%. This negative effect was found to be robust across specifications. Furthermore, we find some evidence that the effects of labor market concentration are less negative in more unionized industries and more severe when the worker is employed on a part-time basis. In our preferred specification to measure the employment effects of concentration, we control for occupation by commuting zone fixed effects and instrument labor market concentration in the same way as before. We find that a 10% increase in labor market concentration lowers the number of new hires by about 3.2%. That labor market concentration decreases wages and hires is exactly what economic theory would predict in an oligopsonistic labor market (Manning, 2011; Azar et al., 2019).

Our second finding concerns the impact of *product* market concentration on wages (Qiu and Sojourner, 2019) and hires. Product market concentration is calculated at the industry by commuting zone level. Since labor and product market concentrations are positively correlated, we add this variable in all our main regressions to limit omitted variable bias. For our preferred specification for new hires described above, we find that a 10% increase in product market concentration increases hourly wages by 0.65%, with a larger effect in more unionized industries. This result is robust across specifications and consistent with rent sharing in unionized industries. Furthermore, we find that product market concentration decreases hires as predicted by oligopsony theory such that, in our preferred specification, a 10% increase in product market concentration lowers the number of new hires by 3.3%.

Our third finding sheds new light on the expected impact of mergers and how antitrust authorities could anticipate their effects. We simulate the impact of horizontal mergers between the two largest employers in each industry, thus focusing on the mergers that would increase labor market concentration the most in each industry. We calculate the changes in labor market concentration that such major mergers would entail. We then apply our preferred estimate for the impact of labor market concentration to estimate the loss to the number of new hires and the associated wage bill. We find that the economy-wide impact of the merger varies with initial labor market concentration: mergers yield the highest number of lost hires in labor markets with low levels of labor market concentration prior to the merger, which tend to be the markets with the largest number of hires prior to the merger. Overall, compared to other industries, a merger in retail would be the most damaging: a merger between the top two employers in the retail industry would yield 30 million euros of yearly lost wages for new hires, and about 3 000 hires lost annually. When we also take into account the impacts on workers in other industries that share an occupational labor market with workers in the retail industry (e.g. stock clerks in the temporary work industry), the damage extends to 40 million euros loss to the annual wage bill along with 3 900 jobs. Effects on workers outside the retail industry are not negligible since they amount to about 30% of the total effect. After the retail industry, a merger between the top two employers in the building maintenance

industry would be almost as damaging with annual wage losses of about 16 million euros for new hires, and a 2 200 decrease in yearly hires.

We make three key contributions to the literature. First, we use administrative data to obtain the most comprehensive dataset to date on the labor market concentration of new hires by occupation. Relying on hires is more accurate than relying on job postings (Azar et al., 2017a, 2018, 2019; Hershbein et al., 2018) because not all companies post their jobs online. Data on hires is more accurate for measuring current competition in the labor market than data on the stock of employment, especially when such a stock is based on industries (Benmelech et al., 2018; Rinz, 2018; Lipsius, 2018; Abel et al., 2018) rather than occupations. Another advantage of focusing on new hires is that the wages of new hires are more likely to be impacted by market conditions than the wages of stayers (Montornès and Sauner Leroy, 2009; Pissarides, 2009). Our extensive data further allows us to control not only for value added and firm fixed effects (Benmelech et al., 2018) but also for worker fixed effects[1], thereby reducing the scope of omitted variable bias arising from worker composition effects.

Our data allows us to explore the effect of labor market concentration in the European context of France: we show that the impact of labor market concentration on wages and employment is negative even when unions are powerful and labor market regulations are stringent. This adds to the evidence from Portugal (Martins, 2018) and Sweden (Dodini et al., 2020), showing a negative impact of labor market concentration on wages. Another study using French administrative data (Bassanini et al., 2020) focuses on stayers instead of new hires, and shows that labor market concentration reduces the wages of stayers.

Our second key contribution is to go beyond the wage effects of labor market concentration that prior literature has estimated to examine the effects of labor and product market concentration on worker flows. We find that both labor and product market concentration negatively affect hires, but the effect is more precisely estimated for labor market concentration. Exits also decrease with labor and product market concentration, leading to a more sclerotic labor market. Overall, the net effect of labor and product market concentration on employment is negative.

Our third key contribution is to shed light on how consolidation may affect both wages and employment by simulating horizontal mergers between the two largest players (by employment) in each industry, adding to the literature on the effects of mergers on workers (Brown and Medoff, 1987; Shleifer and Summers, 1988; Gokhale et al., 1995; Conyon et al., 2001; Gugler and Yurtoglu, 2004; Margolis, 2006; Lehto and Böckerman, 2008; Siegel and Simons, 2010; Prager and Schmitt, 2018; Arnold, 2019). Comprehensive data is critical to measure the full impact of mergers: in particular, we find that 30% of the impact of mergers affects workers in industries *other* than the industry where the merger took place. Through this exercise, we provide a simple method that can be used in practice by competition authorities to assess the likely impact of a merger. In particular, we find that in France, mergers in retail and in building maintenance would be the most damaging in terms of lost wages and jobs.

The paper proceeds as follows. First, Section 2.2 defines our measure of labor and product market concentration, introduces the French matched employer-employee dataset, and describes the statistical relationship between our main variables of interest. Second, Section 3.3 presents our main econometric evidence with regards to impact of labor and product market concentration on wages and the number of new hires. Finally,

[1] Worker fixed effects are identified off of workers who are hired multiple times during our sample frame 2011-2015.

Section 2.4 presents the counter-factual exercise consisting in simulating the impact of mergers of the top two employers in each industry on the labor market.

2.2 Measuring Labor and Product Market Concentration

2.2.1 Data

2.2.1.1 Primary Sources

Two main data sources are used in this paper. They form what is commonly referred to as linked employer-employee data. First, the *Déclaration Annuelle de Données Sociales* (DADS) provides us with individual level data on workplace location, wages, hours worked, occupation, industry, gender, and age. For multi-establishment firms, the data provides establishment identifier and location. This allows us to distinguish between workers employed in different establishments of the same firm. Maintained by the French National Institute for Statistical and Economic Studies (INSEE), this administrative dataset covers all French private and public sector workers. Further description of this data can be found, for example, in Abowd et al. (1999). Whilst this dataset is not freely accessible, any researcher can request access to it through the Secure Data Access Centre (CASD).

The subfile *DADS Salariés* is an exhaustive repeated cross-section of workers that allows us to identify individual workers and their primary source of income (i.e, the job providing them with the most income during a given year). When there are no main sources of income, one is created by aggregating the different income sources. We use this subfile to construct our measures of concentration (see below) and employment flows. The subfile *DADS Panel* provides a worker identifier allowing us to control for individual fixed effects in hourly wage regressions. However, it only records workers born during the month of October. For this reason, it is not exhaustive (making it unreliable to construct concentration indices) and less relevant to study employment flows.[2]

Second, the firm identifier (*code SIREN*) allows us to link workers to the characteristics of their respective firms. These characteristics are those provided in standard financial disclosures at the yearly level. These financial disclosures stem from the database *Système unifié de statistique d'entreprises* (SUSE; unified system of firm statistics) collected by INSEE and the French Treasury (*Direction Générale des Impôts* (DGI)). Its main dataset is called the *Fichier complet unifié de SUSE* (FICUS; complete unified file of SUSE). In 2007, it was replaced with *Élaboration des Statistiques Annuelles d'Entreprise* (ESANE). From this database, we get our measure of firm revenues along with our controls for firm size and value added per employee.

2.2.1.2 Secondary Sources

Two secondary datasets are used for the purpose of exploring alternative specifications. First, the survey on financial links between companies, *Enquête sur les liaisons financières entre sociétés* (LIFI), is used to identify the business group to which firms belong and allows us to construct a measure of labor market concentration at the business group level. This survey describes the ownership and subsidiaries of companies, identified by

[2] Nonetheless, we provide in Appendix 2.C.7 the baseline regressions on new hires based on the DADS Panel. They are broadly in line with those estimated on the repeated cross-section.

the *Siren* number. The survey reviews all firms satisfying one of the following criteria : (a) owns over 1.2 million euros of another company's shares, (b) employs over 500 employees, (c) has a turnover over 60 million euros per year, (d) was a business group headquarter in the previous year, or (e) was foreign owned in the previous year (i.e, at least 50% of its shares are owned from a foreign firm). Respondents must identify a subsidiary if they own over 30,000 euros of its shares. This allows us to identify business groups in a comprehensive way. We match a firm to its business group using, as above, the *code SIREN*. When a firm had no business group, we assigned its firm identification number (*code SIREN*) as its business group identifier. This dataset has already been used by Cestone et al. (2017), which can be consulted for further information.

Second, we measure unionization rate at the industry level using the *Enquête Réponse* (2011). We use this measure to look at the relationship between market power and unionization. As explained below, we take this as a proxy for the unions' bargaining strength. This survey was administered by the French Ministry of Labor to 18 536 individual workers (with at least a 15 months tenure) in firms with at least 11 employees in the private and semi-private sector (excluding agriculture and the public administrations). It asked workers if they were part of a union. Respondents could answer (i) "yes", (ii) "No, and I have never been a member", and (iii) "No, but I have been in the past". We recode answers as a binary variable (yes or no). To recover a unionization rate by industry, we aggregate this new variable at the 2-digit industry level because there were too few respondents to measure the unionization rate at the 4-digit level accurately.[3] We provide these raw unionization rates in Table 2.M.1 in the Appendix 2.M.1, along with the number of individuals used to calculate these rates.

2.2.2 Sample Selection

Our sample selection procedure is the following. First, we only keep new hires i.e. those who have employment contract start dates during the quarter of observation.[4] Due to wage rigidity and employment protection, there is evidence (Montornès and Sauner Leroy, 2009; Pissarides, 2009) that wage cuts and workforce adjustments tend to disproportionately affect new hires.[5] Second, we only keep private sector employees. We also exclude state-sponsored workers, apprentices, and interns. We drop workers below 18 and above 67. Beyond the public sector, we also drop non-governmental organizations, the art industry, museums, sports clubs, unions, and home production. Our data covers new employees from 2011 to 2015 included. During these years, the data collection system did not change, it is both the most recent and complete (in terms of response rate) version of the data.[6] Finally, for the purposes of studying hourly wages, we discard for each year the 5% lowest hourly wages and the 1% highest wages. This leaves us with valid observations (above the minimum hourly wage) and excludes outliers.

To measure concentration, we always use the cross-sectional data on new hires (DADS Salariés). For regressions, we use two different samples depending on the outcome of interest. When we analyze the impact of concentration on hourly wages, we use the smaller worker panel (DADS Panel), so that we can control for worker fixed effects. When we turn to the impact of concentration on new hires, we use the cross-sectional

[3] We drop workers in the Temporary Employment Industry because we cannot distinguish between the Temp. workers who rarely belong to a union and the permanent employees of the Temp. agencies who have a standard pattern of unionization.

[4] In the data, workers who start on January 1st may be starting on January 1st or continuing a job from the previous year. Therefore, we exclude observations whose job spells start on January 1st for each year of observation.

[5] Our data period, 2011-2015, is a period of growth in France. During such a period, higher labor market concentration could slow wage growth, even if it does not lead to wage cuts.

[6] Starting in 2009, the whole population in employment is covered by the dataset, commuting zones were redefined in 2010, and non-response in 4-digit occupation is low and stable starting in 2010.

data (DADS Salariés). Consulting Table 2.A.2 in Appendix 2.A, one can observe that our two samples cover broadly the same number of commuting zones (at least 304), occupations (at least 403), and industries (at least 601). The panel data covers one million individuals. As expected, the more exhaustive repeated cross-section records nearly three times more firms (around 1 million) than the panel.

2.2.3 Definition of the Herfindahl-Hirschman Index (HHI)

2.2.3.1 Labor Market Herfindahl-Hirschman Index (HHI)

We now define our measures of labor (L) and product (P) market concentration. Labor Market Concentration is measured through the Labor Market Herfindahl-Hirschman Index (HHI), as in Azar et al. (2017a). This index measures concentration through market shares. Let $J_{o,m,t}$ be the set of firms hiring in occupation $o \in O$ (measured at the 4-digit level) in geographical area $m \in M$ (measured at the commuting zone) at time $t \in T = \{Q12011, ..., Q42015\}$ (measured at the quarterly level from 2011 to 2015). The number of workers of this occupation, time, and commuting zone hired by firm $j \in J_{o,m,t}$ is denoted $N_{j,o,m,t}$. The firm's labor market share $s^L_{j,o,m,t}$ is then:

$$s^L_{j,o,m,t} = \frac{N_{j,o,m,t}}{\sum_{k \in J_{o,m,t}} N_{k,o,m,t}} \tag{2.2.1}$$

For example, if at a given time and commuting zone there is a total of 100 cleaners being hired, a firm hiring 10 of these cleaners would have a 10% market share. The labor market Herfindahl-Hirschman Index, $\text{HHI}^L_{o,m,t}$, sums the squares of these market shares:

$$\text{HHI}^L_{o,m,t} = \sum_{k \in J_{o,m,t}} \left\{ s^L_{k,o,m,t} \right\}^2 \tag{2.2.2}$$

This index is always between zero (excluded) and one. When it is equal to one, a single firm employs all new hires. One way to interpret the HHI is through the 2010 horizontal merger guidelines of the American Department of Justice and Federal Trade Commission. An HHI between 0,15 and 0,25 is indicative of a moderately concentrated market and above 0.25 of a highly concentrated market.

2.2.3.2 Product Market Herfindahl-Hirschman Index (HHI)

We also construct a Product Market Herfindahl-Hirschman Index (HHI). To do so, we locate firms according to the commuting zone in which their employees are located. We then use the national sales of these firms to measure a commuting-zone specific product market share. If the firms at the local level had the same share of sales (relative to competitors present in the commuting zone) as at the national level and the product market were local, then this way of calculating would mimic a localized measure of labor market concentration.

More formally, we consider firms with at least one employee in commuting zone $m \in M$, at time $t \in T$ (national sales are recorded at the yearly level), in industry $i \in I$ (measured at the 4-digit level). These firms are collected in a set $V_{i,m,t}$. For any firm j in this set, we observe the national sales (measured in nominal euros) during that year, denoted by $\mathrm{R}_{j,t}$. We can then define the product market share as:

$$s^P_{j,m,t} = \frac{R_{j,t}}{\sum_{k \in V_{i,m,t}} R_{k,t}} \tag{2.2.3}$$

The product market Herfindahl-Hirschman Index, $HHI^P_{i,m,t}$, sums the squares of these market shares:

$$HHI^P_{i,m,t} = \sum_{k \in V_{i,m,t}} \left\{ s^P_{k,m,t} \right\}^2 \tag{2.2.4}$$

Despite using national sales of a company instead of the more ideal local sales, we believe this measure to be adequate for large markets. For example, by virtue of the size of the Parisian market, the sales share of each firm in Paris is likely to be similar to the national share of sales. In a small local market, one firm – e.g. a locally-owned supermarket selling regional foods – could be dominant, even though its share of *national* sales is very small relative to national supermarket chains that operate in the same commuting zone. Although our product market HHI under-estimates the degree of competition when local firms serve a large share of the market, we believe the number of such markets to be few. Moreover, to the degree that this error is systematic, our control variables will rely on variation across time rather than in levels for identification. Finally, we explore alternatives to this imperfect measure in Section 2.3.4 below. We show that our baseline estimates, for the impact of labor market concentration, are not significantly affected by the exclusion of firm level controls (including the product market HHI), the use of product market concentration defined using national sales (which we call the *global* product market HHI) in the industries exposed to international trade, the reliance on commuting-zone by industry by time dummy variables, or the use of employment-weighted product market concentration.

2.2.4 Descriptive Statistics

Table 2.1 provides descriptive statistics for the estimation sample used to study the relationship between our measures of concentration and hourly wages.[7] The average labor market HHI at the firm level is 0.15 whilst its median nears 0.06. This difference between the mean and the median reflects the existence of a few markets with high levels of labor market concentration.

Table 2.1: Summary Statistics : Individual Level Data

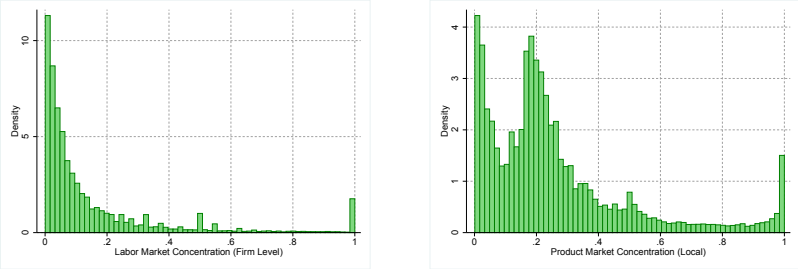
	count	min	max	p50	mean	sd
Gross Hourly Wage	2225026	9	44.0625	11.88235	13.34	4.567
Labor Market Concentration (Firm)	2225026	.0005867	1	.0644531	0.151	0.217
Labor Market Concentration (Business Group)	2225026	.0008135	1	.0752775	0.163	0.220
Product Market Concentration	2225026	.0011498	1	.201532	0.260	0.234
Product Market Concentration (Global)	2225026	.00007	1	.0021053	0.00608	0.0161
Age (in years)	2225026	18	67	28	31.80	11.46
Gender (1 if Male)	2225026	0	1	1	0.564	0.496
Unionization Rate	1476658	0	45.71	10.32	9.862	3.882
Nb. Full-Time Equivalent Employees	2225026	.001	250825	139	16527.5	42083.9
Value Added per Emp. (in nominal euros)	2225026	.0001111	49235.57	33.22288	76.81	273.0

Note: Each observation used to construct this table is a job spell at the individual level (DADS Panel). Their associated level of concentration is calculated based on the repeated cross-section (DADS Salariés).
Source: DADS, FICUS, and authors' calculations.

This can be seen more clearly by considering Figure 2.1(a) which depicts the density of the HHI in the labor market across workers. There appears to be a significant portion of workers who face a single employer (monopsony). The same can be said based on Figure 2.1(b) for the product market having a single seller

[7] Table 2.A.3 in Appendix 2.A provides an equivalent table based on the estimation sample used to study new hires. It is based on data aggregated at the 4-digit occupation by 4-digit industry by quarter level, from the repeated cross-section data provided by the DADS Salariés subfile.

(monopoly). Looking at Table 2.A.1 in Appendix 2.A, which provides these same summary statistics for the five most common occupations, this feature of the distribution in labor and product market concentration appears to extend to the most common occupations.



(a) Labor Market Concentration (b) Product Market Concentration

Figure 2.1: Histogram of Labor and Product Market HHI

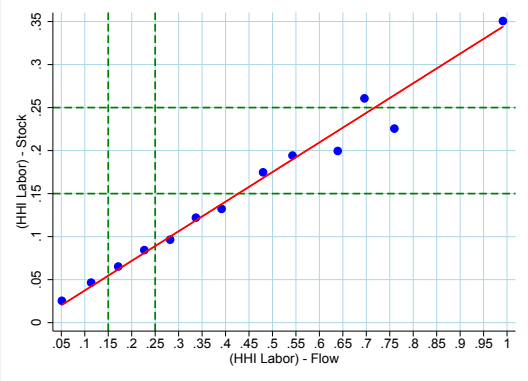
Note: These figures were constructed using individual level data (DADS Panel). Given that each worker is assigned a level of concentration (based on the cross-section DADS Salariés), these histograms reflect the distribution of concentration across the new hires.
Source: DADS, FICUS, and authors' calculations.

Before turning to the econometric evidence, it is important to discuss our measure of labor market concentration and the way it relates to other measures used across the literature. This paper measures labor market concentration through employment flows, because this is the most relevant way of capturing job opportunities for workers looking for a job. Indeed, if a worker was hired, it manifests that a job was available. By contrast, the total number of workers is not as direct an indication of the number of available jobs. Prior literature has used employment stocks to measure labor market concentration, albeit by industry rather than occupation (Benmelech et al., 2018; Rinz, 2018; Lipsius, 2018; Abel et al., 2018). Therefore, it is interesting to examine the differences between stock and flow measures of labor market concentration by occupation. We present in figure 2.2(a) a binscatter allowing one to convert flow levels of labor market concentration to stock levels. To construct it, we calculated for each market (occupation by commuting zone) the average HHI and provided its best fit line. Clearly, there is a near linear relationship between labor market concentration based on flows and on stocks: the R-squared of the superimposed regression line is equal to 43%.[8] The main regression tables for the wage regressions found in the following section are also provided using the stock level labor market concentration in Appendix 2.K.[9] When the HHI is measured at the industry level (Figure 2.2(b)), there is a similar relationship between our preferred flow-based measure of labor market concentration and the stock based measure. This latter measure has been used in prior literature, as it is often more easily available. Although the relationship is weaker, with an R^2 of 20%[10], there is nonetheless evidence of a strong correlation between the two measures.

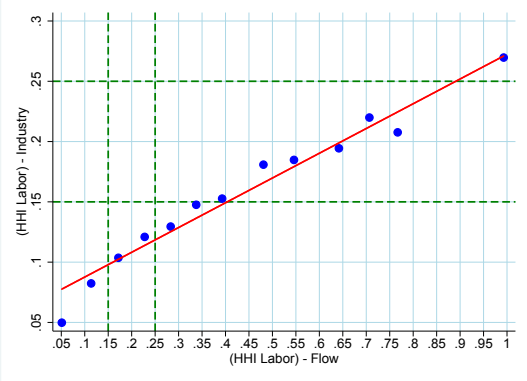
[8] i.e: log(Stock HHI) $= -1.45 + 0.9101$log(Flow HHI)

[9] The reader is referred to Bassanini et al. (2020) who consider employees who do not change firms (i.e, stayers) for an alternative population of interest and approach.

[10] i.e: log(Industrial Stock HHI) $= -1.69 + 0.5322$log(Flow HHI).



(a) HHI - Stock v. HHI - Flow



(b) HHI - Industry-Stock v. HHI - Flow

Figure 2.2: Binscatter : comparison of different measures of labor market concentration

Note: These figures are binscatters performed at the occupation by commuting zone level. We use the stock of employees in 2014 and 2015 to calculate an HHI by 4-digit occupation and occupation (denoted *HHI - Stock*). We also calculate an HHI defined over industries (at the 4-digit industry by a commuting zone level) which we denote by *HHI - Industry-Stock*. These new measures of concentration are then averaged across our usual market definition (an occupation by a commuting zone) and matched with the flow level of labor market concentration. In each case, we rely on the repeated cross-section DADS Salariés. The sample was limited to 2014 and 2015 for computational convenience.
Source: DADS and authors' calculations.

While stock and flow based labor market concentration measures are highly correlated across markets, their levels are quite different. The figures use vertical and horizontal lines to indicate thresholds used by the US federal antitrust authorities to gauge levels of concentration. By the standard of the Department of Justice / Federal Trade Commission 2010 horizontal merger guidelines, 0.15 is the threshold between low and medium concentration while 0.25 is the threshold between medium and high concentration. In Figure 2.2(a) and Figure 2.2(b), we see that stock-based measures of concentration show systematically lower levels of concentration than flow-based measures, which makes sense as not all firms hire in every given quarter. As a result, the 0.25 threshold for high concentration in the stock measure of labor market concentration corresponds to a concentration as high as 0.7 in the flow-based measure of labor market concentration! Even the threshold of 0.15 for medium concentration in the stock-based measure of labor market concentration corresponds to a flow-based HHI of about 0.4, which is way above the high concentration threshold. This shows that measuring labor market concentration by stocks severely underestimates the level of concentration among new hires.[11] If only a stock-based measure of concentration is available, thresholds of about 0.05 and 0.15 correspond to the relevant medium and high concentration thresholds in the flow-based measure. To the extent that new hires adequately measure available job opportunities for workers, competition authorities should use the flow based measure, or, if only the stock-based measure is available, realize that it corresponds to much higher levels of flow-based labor market concentration.

The existence of business groups may lead to under-estimating labor market concentration to the extent that firms within a group do not compete for workers. However, this turns out not to be a big problem empirically: as suggested by Figure 2.3 below, the two measures are almost perfectly correlated and estimation results are not sensitive to measuring labor market concentration at the business group versus the individual firm level. Table

[11] The concentration among new hires is also relevant for the wages of job stayers because it reflects their potential outside options at a given point in time (Bassanini et al., 2020).

2.1 also shows that the *levels* of concentration measured at the firm level or the group level are very similar, even if concentration is as expected slightly higher at the group level with a mean of 0.163 instead of 0.151 at the firm level. We nonetheless provide in Appendix 2.B.4 estimates of our baseline specifications using the business group measure of labor market concentration.

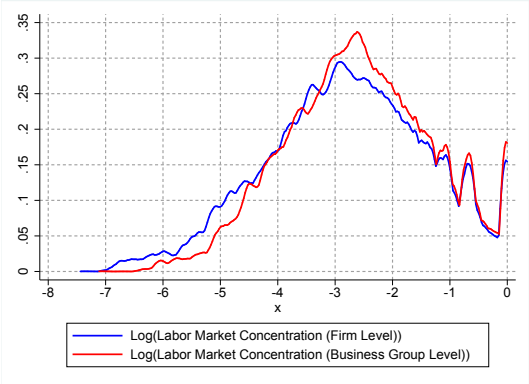


Figure 2.3: Distribution of the Log(Labor HHI) at the Firm and Business Group Level

Note: This figure was constructed using individual level data (DADS Panel) in 2011-2015. Each worker has both a measure of labor market concentration at the business group and at the firm level. The concentration levels were calculated using the repeated section (DADS Salariés). Source: DADS, LIFI, and authors' calculations.

2.3 Econometric Assessment of Monopsony

Monopsony Theory (Boal and Ransom, 1997; Manning, 2011; Robinson, 1969) predicts that both employment and hourly wages should fall as a result of a rise in labor market concentration (Azar et al., 2019). Indeed, the key intuition for monopsony power is by analogy with monopoly power: profit-maximizing employers with monopsony power keep both wages and employment below the competitive equilibrium. The presence of concentration in the product market (monopoly power) reduces output, which should result in fewer workers employed. On the other hand, the impact of product market concentration on wages is unclear (Qiu and Sojourner, 2019). In the presence of rent sharing, one would expect greater product market concentration to increase wages to the extent that profits increase. Table 2.2 summarizes the predicted effects.

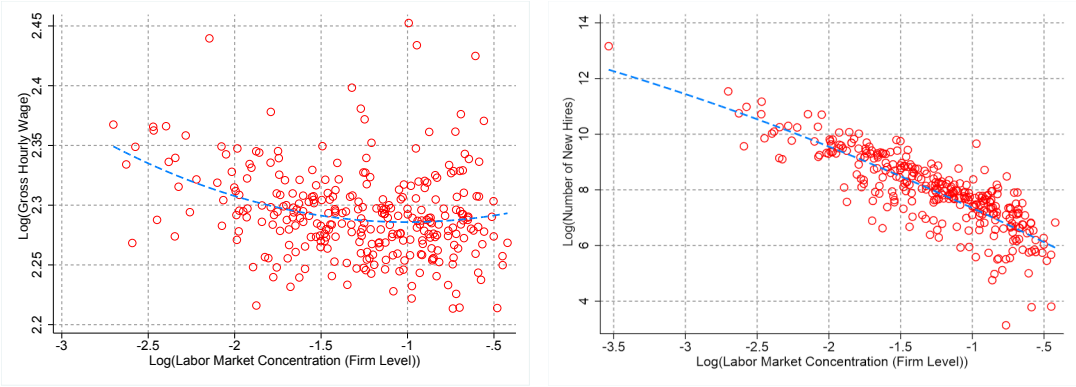
	Employment	**Hourly Wage**
Product Market HHI	-	+ ?
Labor Market HHI	-	-

Table 2.2: Expected Effects of Labor and Product Market Concentration

2.3.1 Research Context

2.3.1.1 Descriptive Evidence

Our goal is to assess these predicted effects of labor and product market concentration in the French labor market. To this end, we consider the correlation across commuting zones between, on the one hand, labor market concentration, and, on the other hand, hourly wages and the number of new hires. Figure 2.4(a) depicts the log of the average gross hourly wage against the log labor market HHI by commuting zone. There is a clear negative relationship between hourly wages and labor market concentration. Figure 2.4(b) shows a strong negative relationship between market size (in terms of recruitment flows) and labor market concentration. Both of these observations are consistent with the core predictions of the monopsony model. Of course, the negative relationship between concentration and hires is somewhat mechanical since fewer hires also typically entails fewer firms hiring. Our regression analysis will address this issue by using the instruments described below.

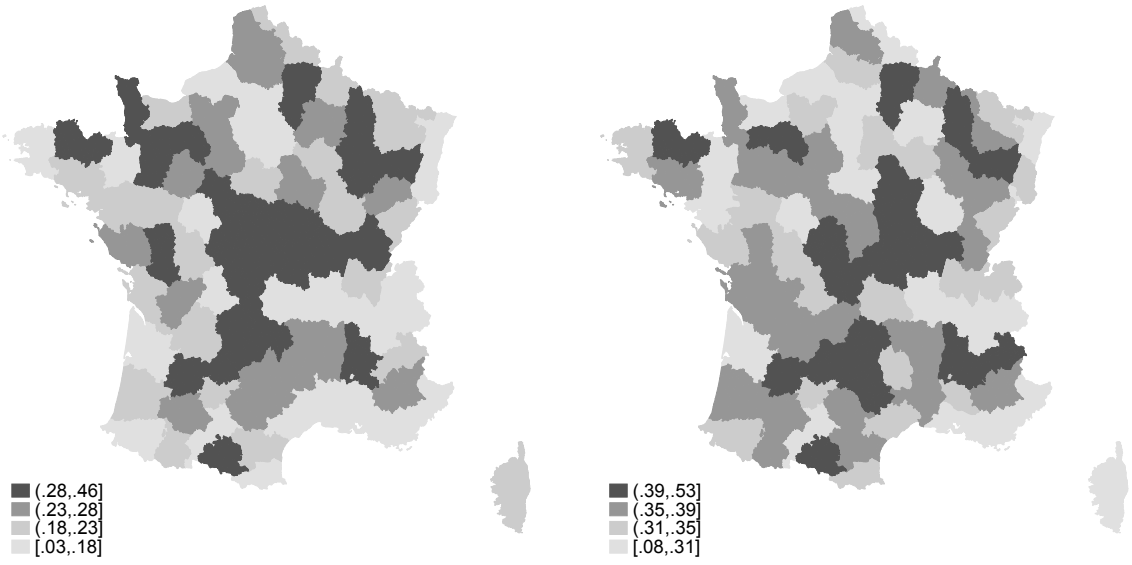


(a) (b)

Figure 2.4: Hourly Wage and Number of New Hires Against Labor Market Concentration

Note: Each point represents a commuting zone through its average level of labor market concentration, number of new hires, and mean hourly wage (2011-2015). These averages were calculated using the worker panel (DADS Panel) but relying on concentration measured calculated on the repeated cross-section (DADS Salariés).
Source: DADS and authors' calculations.

Moreover, our task is complicated by the existence of both observed and unobserved confounders. These confounders motivate the use of regression analysis with fixed effects and control variables, along with the use of instrumental variables. Indeed, we can observe that concentration varies systematically across the French territory. Maps 2.5(a) and 2.5(b) display the mean labor market and product market HHI per *département* (administrative unit similar to a US county). The product market HHI is calculated on the basis of the identity of firms that have at least one employee in a worker's industry in the same geographic market. The labor market HHI is calculated on the basis of the identity of the firms that hire in the same occupation as the worker and same commuting zone. Even though sales *shares* come from national sales, this way of calculating the product and labor market HHIs will yield relatively high levels of concentration in less populated areas where fewer firms hire, whether that is within an occupation or within an industry. We see (i) that areas with high product and labor market concentration overlap, (ii) low population density areas have high concentration market structures, and (iii) given that low population density areas have low wages, one could be led to believe that the (presumably) negative impact of labor market concentration on wages dominates the (presumably) positive impact of product market concentration.



(a) (b)

Figure 2.5: Map of Labor and Product Market HHI

Note: This figure was constructed using individual level data (DADS Panel). Each worker has both a measure of labor market concentration and product market concentration calculated based on our repeated cross section (DADS Salariés). These measures are aggregated at the *département* level.
Source: DADS, FARE, and authors' calculations.

2.3.1.2 Identification

To test the predictions of Monopsony Theory, we must be able to disentangle the effects of market concentration from both observable and unobservable confounders. To this end, we estimate two sets of regressions which include fixed-effects, control variables, and which we identify through instrumental variables. The first set focuses on the log(Gross Hourly Wage) observed at the individual worker level (and estimated on the longitudinal DADS Panel). The baseline results are provided in Section 2.3.2 below. The second set is concerned with employment flows, measured through the log(Number of New Hires) in a given combination of occupation, industry, and commuting zone, calculated for each time quarter. These latter regressions are performed on the cross-sectional DADS Salariés and rely on market-level aggregates such as the mean age and gender among new hires. The baseline results are provided in Section 2.3.3 below.

In each case, we attempt to disentangle observable confounders from evolving market structure effects. Our linked employer-employee dataset allows us to account for several fixed and time-varying covariates which could threaten identification. To this end, in our wage regressions, we control for time, occupation, commuting zone, occupation by commuting zone, firm[12], and worker fixed effects. We also include gender (only identified without individual fixed effects) and age, along with the logarithm of the number of full-time equivalent employees in the hiring firm and the logarithm of the value-added per full-time equivalent in the hiring firm. The latter two

[12]Controlling for plant fixed effects does not substantially change the estimates.

variables are included to control for the potential correlation which may exist between labor and product market concentration on the one hand, and firm productivity on the other. In our employment regressions, we control for time, occupation, commuting zone, occupation by commuting zone, and occupation by commuting zone by industry fixed effects. We construct analogue control variables to those used for hourly wage regressions, including the share of men among the hires, their mean age, along with the mean log firm size (defined above) and mean log value-added per employee in the firms to which new hires belong.

Moreover, there are potentially unobserved time-varying market-specific variables that we did not control for, that are correlated with our measures of concentration, and that affect wages. For example, according to Search and Matching Theory, wages are determined by labor market tightness (i.e, the ratio of job openings to job seekers), productivity, and the workers' unemployment benefits (Rogerson et al., 2005). We control for proxies of productivity and given that unemployment benefits are determined nationally, we are able to control for workers' out-of-work benefits by controlling for time fixed effects. However, we are unable to control for time-varying changes in labor market tightness at the market level because we do not observe job openings and job seekers.

To address this issue, we follow the strategy deployed in Azar et al. (2017a); Martins (2018); Qiu and Sojourner (2019) and instrument our concentration indices with a Hausman instruments. These instruments measure national shocks, assume them to be uncorrelated to local shocks, and use their variation to identify exogenous changes in the endogenous variable of interest. This type of instrumental variables strategy is commonly used in the field of Industrial Organization to address price endogeneity within the product market. For example, Nevo (2001a) relies on the prices in other geographic markets to instrument for city-level prices of various products in the ready-to-eat cereal industry.

For the labor market HHI, we consider as instrument the average number of firms in other markets recruiting a given occupation. Formally, we have for each quarter $t \in T$, commuting zone $m \in M$ with cardinality $|M|$, and 4-digit occupation $o \in O$:

$$\text{Instrument Worker}_{o,m,t} = \frac{1}{|M|-1} \sum_{v \in M-\{m\}} -\log\Big(\sum_{k \in J_{o,v,t}} 1(N_{k,o,v,t} > 0)\Big) \tag{2.3.1}$$

where $N_{k,o,v,t}$ are the number of new hires for firm k.

This provides us with variation in market concentration that is driven by national-level changes in the occupation, and not by changes in the occupation in that particular local market. For example, if the labor market tightness for cleaners (the most common occupation) falls in the Paris area, this could both decrease wages and increase concentration, since fewer firms would likely be recruiting. By instrumenting with the number of firms hiring cleaners in other areas, we rule out an effect of labor market tightness in Paris on Labor HHI.

We rely on a similar strategy to instrument the product market HHI. Our instrument is the average number of firms in other commuting zones hiring within the same industry. This instrument will fluctuate when there are national shocks to the industry but not when there are local shocks to the industry. More formally, we have for each quarter $t \in T$, commuting zone $m \in M$ and 4-digit industry $i \in I$:

$$\text{Instrument Firm}_{i,m,t} = \frac{1}{|M|-1} \sum_{v \in M-\{m\}} -\log\Big(\sum_{k \in V_{i,v,t}} 1(N_{k,i,v,t} > 0)\Big) \tag{2.3.2}$$

where $N_{k,i,v,t}$ are the number of new hires for firm k in industry i.

First-Stage estimates for the hourly-wage are provided in Appendix 2.B.3 whilst those for new hires are available in Appendix 2.C.4. The instruments appear to be relevant based on their statistical significance and high F-statistic which is always above 10. This is particularly true at the aggregate level used to study the impact of concentration on the number of new hires. In this case, the F-statistic is above 100 in all cases but the most demanding specification (with industry by occupation by commuting zone fixed effects). This observation holds when we weight (analytically) our regressions by the number of hires and mean number of hires across time (for a combination of occupation, industry, and commuting zone).

2.3.2 Impact on Hourly Wages

2.3.2.1 Baseline

We first estimate the impact of concentration on hourly wages. To do so, we rely on longitudinal individual level data provided by the DADS Panel, where each observation is an employment spell. We provide the estimates using ordinary least squares (OLS) along with those relying on the instrumental variables (IV) described in Equations 2.3.1 and 2.3.2 above. In our most demanding specification, we estimate for worker $e \in E$, firms by $j \in J$ in industry $i \in I$, occupation $o \in O$, and commuting zone $m \in M$ at time quarter $t \in T$:

$$\begin{aligned}\log(w_{e,j,o,m,t}) = \alpha_L \log(\text{Labor HHI}_{o,m,t}) + \alpha_P \log(\text{Product HHI}_{i,m,t}) \\ + X'_{e,j,t}\lambda + \Psi_j + \Omega_e + \zeta_{o,m} + \Xi_t + \varepsilon_{e,j,o,m,t}\end{aligned}$$

where $w_{e,j,o,m,t}$ is the gross hourly wage, α_L is the elasticity of the hourly wage with respect to labor market concentration, α_P is the elasticity of the hourly wage with respect to product market concentration, Ψ_j are firm fixed effects, Ω_e are individual fixed effects. The vector $X'_{e,j,t}$ collects control variables (with associated parameter vector λ) such log(Nb. Employees) measured in terms of full-time equivalent employees per year in the hiring firm, and log(Value Added per Employee) measured in terms of annual revenue per full-time equivalent employee in the hiring firm. In specifications without individual fixed effects, we can also identify the effect of a male gender dummy and of age (measured in years, as a continuous variable). $\zeta_{o,m}$ are commuting zone by occupation fixed effects and Ξ_t are time fixed effects. $\varepsilon_{e,j,o,m,t}$ is the error term. We provide standard errors clustered at the commuting zone level to account for common shocks and their persistence across time within a labor market.

In practice, we estimate several specifications. This allows us to examine the trade-off between having a parsimonious model and a more demanding model with fewer potential unmeasured confounders but less variation to identify parameters of interest. Six specifications are presented with increasingly demanding fixed effects. The first provides only time and occupation fixed effects. The second adds commuting zone fixed effects. The third combines the two previous ones by also including occupation by commuting zone fixed effects. The fourth and fifth append, respectively, firm and worker fixed effects[13] to the third specification. The final column provides both firm and worker fixed effects, as described in Equation 2.3.2.1. Results for the ordinary least squares estimation are reported in Table 2.3.

[13]Worker fixed effects are identified through individuals who are new hires several times within the time frame covered by our data. These workers' job mobility may be different from that of the general population.

We find results consistent with monopsony: labor market concentration is negatively associated with the wage. This is true across specifications and, although the magnitudes are small, all coefficients are statistically significant at the 1% level. The most negative coefficient is in column (1), which controls for time and occupation fixed effects: a rise of 10% in labor market concentration lowers hourly wages by 0.13%. This suggests that the partial correlation of concentration and wages is fairly strong across geographic labor markets: at a given point in time and for a given occupation, geographic labor markets with higher concentration have lower wages for new hires. The effect is quantitatively weaker when we rely on across time variation by controlling for occupation by commuting zone fixed effects (column (3)). The effect that is closest to zero is in column (6), which includes worker and firm fixed effects along with the occupation by commuting zone fixed effects: an increase by 10% in labor market concentration lowers hourly wages by 0.02%. The size of the coefficients falls as more rigorous fixed effects are added. On the product market side, estimated effects are also small. They range from an elasticity of -0.005% in column (1) to 0.002% in column (4), once firm fixed effects are added. Estimates for age and gender appear to be in the usual range, providing credence to our analysis. We can also report positive coefficients associated with the value added per worker and the firm size. Overall, the adjusted R^2 stays constant, rising slightly when firm and worker fixed effects are introduced.

Table 2.3: Hourly Wage (OLS) : Baseline

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.0132***	-0.00782***	-0.00304***	-0.00244***	-0.00216*	-0.00206***
	(0.00278)	(0.00138)	(0.000921)	(0.000620)	(0.00123)	(0.000763)
Log(Product HHI)	-0.00460***	-0.000793	-0.000604	0.00166	-0.00246	-0.00189***
	(0.00144)	(0.00125)	(0.00144)	(0.00166)	(0.00166)	(0.000647)
Age (in years)	0.00339***	0.00336***	0.00327***	0.00274***		
	(0.000414)	(0.000409)	(0.000470)	(0.000484)		
Gender	0.0304***	0.0295***	0.0287***	0.0242***		
	(0.00106)	(0.00104)	(0.00167)	(0.00248)		
Log(Value Added per Employee)	0.0230***	0.0223***	0.0202***	-0.000885	0.0112***	0.000702
	(0.00166)	(0.00174)	(0.00186)	(0.000725)	(0.000577)	(0.000745)
Log(Nb. Employees)	0.00815***	0.00791***	0.00781***	0.0000949	0.00722***	0.00162**
	(0.000257)	(0.000237)	(0.000203)	(0.00138)	(0.000121)	(0.000710)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.523	0.527	0.556	0.664	0.741	0.793
Adjusted R^2	0.523	0.527	0.548	0.629	0.633	0.677
N. Clusters	304	304	304	304	304	304
F	1133.4	966.3	1193.2	230.7	1111.9	11.69
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00216 \times 0.1 \times 100 = -0.0216\%$.

Next, we consider the two stage least squares estimates provided in Table 2.4. The signs of the coefficients

are in accord with the basic predictions of the monopsony model across specifications and the magnitudes have increased compared to the OLS estimates. In terms of labor market concentration, at one extreme, one finds that a 10% increase in the HHI leads to a 0.97% fall in hourly wages (column (1)). At the other extreme, in column (4), parameter estimates suggest that a 10% increase in labor market concentration lowers hourly wages by 0.48%. All coefficients are statistically significant at the 1% or 10% levels. In terms of product market concentration, we find positive and statistically significant coefficients (at the 1% level) in specifications that do not include firm fixed effects. At most, a 10% increase in the Product HHI would lead to a 0.82% increase in hourly wages (column 1).[14] We consider column (5) as our preferred specification because controlling individual fixed effects explain more of the wage heterogeneity (see OLS results) and controlling for firm fixed effects seems to reduce too drastically the amount of variation in the data. We thus find that a 10% increase in labor market concentration decreases hourly wages by 0.52%, while an equivalent increase in product market concentration increases hourly wages by 0.65%.

Table 2.4: Hourly Wage (IV) : Baseline

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.0970***	-0.0850***	-0.0674***	-0.0478***	-0.0518***	-0.0199**
	(0.0128)	(0.0102)	(0.00829)	(0.00782)	(0.00999)	(0.00993)
Log(Product HHI)	0.0823***	0.0770***	0.0755***	0.0262	0.0654**	-0.0270
	(0.0243)	(0.0261)	(0.0248)	(0.0261)	(0.0300)	(0.0291)
Age (in years)	0.00336***	0.00335***	0.00328***	0.00274***		
	(0.000355)	(0.000391)	(0.000457)	(0.000486)		
Gender	0.0257***	0.0282***	0.0268***	0.0241***		
	(0.00144)	(0.00185)	(0.00194)	(0.00255)		
Log(Value Added per Employee)	0.0196***	0.0194***	0.0179***	-0.00121*	0.00977***	0.000636
	(0.00210)	(0.00314)	(0.00317)	(0.000703)	(0.00162)	(0.000711)
Log(Nb. Employees)	0.00584***	0.00557***	0.00465***	-0.000337	0.00511***	0.00112
	(0.000258)	(0.000253)	(0.000264)	(0.00156)	(0.000199)	(0.000761)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	1199.2	1492.9	2015.9	213.8	691.6	9.515
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0518 \times 0.1 \times 100 = -0.518\%$.

2.3.2.2 Heterogeneity

Overall, these last two tables provide robust evidence that labor market concentration has a negative impact on hourly wages. We now run two additional sets of regressions which will allow us to explore some of the underlying heterogeneity and thereby learn more about potential mechanisms through which concentration

[14] Table 2.B.9 and Table 2.B.10 in Appendix 2.B.5.1 present wage regressions on the more exhaustive repeated cross-section (which includes more variation across firms and individuals). We recover a positive and statistically significant parameters associated with the product market HHI, both with and without the use of firm fixed effects.

affects wages.

First, we document a relationship between unionization and the impact of labor market concentration.[15] We interact our measures for labor and product market concentration with the 2-digit industry unionization rate observed in the *Enquête Réponse* (2011).[16] We provide in Appendix 2.B.1 results for the ordinary least squares specification in Table 2.B.1 and those using instrumental variable in Table 2.B.2. Focusing on the latter, there appears to be a positive impact of unionization on hourly wages, as made clear by the coefficient denoted *Unionization Rate* which reports positive and statistically significant coefficients across specifications. This is in line with expectations and the literature on wages and unionization patterns (e.g, Barth et al. (2017)). The interaction coefficient between labor market concentration and unionization rate is positive across all specifications. It is statistically significant at the 1% level when market fixed effects are included. Based on our preferred specification (IV specification, column (5)), the impact of labor market concentration on wages is positive with a unionization rate above 37.2%. Similarly, we find a positive interaction between product market concentration and unionization, across all specifications. These results on unionization are consistent with those of Benmelech et al. (2018) and Qiu and Sojourner (2019). All in all, this suggests that institutional factors moderate the impact of labor market concentration on wages.

Second, we find that labor market concentration can have very negative effects for workers operating outside standard full-time contracts. In particular, Table 2.B.3 in Appendix 2.B.2 provides the ordinary least squares estimates from interacting each variable and control in our baseline regression with a dummy variable equal to one if the worker is on a part-time contract. This allows us to identify a stronger negative relationship between labor market concentration and hourly wages for the subpopulation of workers in part-time, temporary, or on-call work arrangements. We find that the coefficient relating to the interaction between labor market concentration and part-time employment is negative and statistically significant at the 1% level, once occupation by commuting zone fixed effects are accounted for. Indeed, the least square estimators suggest that the impact of labor market concentration on part-time employment is nearly two times larger than in the overall population.[17] Table 2.B.4, which provides the two stage least squares estimates, finds even larger effects. Again, once market fixed effects are accounted for, there is a negative point estimate associated with the interaction between labor market concentration and part-time work. This coefficient is statistically significant at the 5% level when both firm and worker fixed effects are added, in column (6). In this case, we find that the impact of labor market concentration is $\frac{-0.0446-0.0136}{-0.0136} = 4.27$ times larger for workers with a part-time contract. Overall, these results suggest that workers who are less protected by institutions (such as unions) are more likely to be negatively affected by labor market concentration.

[15] Unionization rates may be endogenous, by virtue of simultaneity, if the level of hourly wages affects the unionization rates of workers and, in turn, the unionization rate affects the hourly wage. This consideration is beyond the scope of this paper and is left for future research.

[16] This measure of unionization may suffer from a potential measurement error by virtue of the relatively small sample size on which the survey is based.

[17] For example, taking the estimates from our preferred specification in column (5), we find that the effect for the non-permanent subset of workers is $\frac{-0.00174-0.00241}{-0.00174} = 2.38$ times larger than for the general population.

2.3.3 Impact on New Hires

2.3.3.1 Baseline

We now consider the impact of market structure on the number of new hires. We do so because the existence of centralized wage bargaining regimes and high minimum wages limit the scope to adapt wages in response to changes in bargaining power related to labor market concentration. This suggests that firms may express changes to their bargaining position by changing their number of hires, as suggested by Monopsony Theory. To this end, we measure employment as a flow : the number of labor contracts signed during a quarter. This measure is calculated based on our repeated cross-section (DADS Salariés) by aggregating our observations. We aggregate at the 4-digit occupation by 4-digit industry code by commuting zone and quarter because this preserves both labor market concentration (at the occupation by commuting zone level) and product market concentration (industry by commuting zone level) along with their instruments. Summary statistics for the subsample used for estimation are provided in Table 2.A.3 in Appendix 2.A. We provide the estimates using ordinary least squares (OLS) along with those relying on the instrumental variables (IV) described in Equations 2.3.1 and 2.3.2.

We denote by $E_{o,i,m,t}$ the number of new hires in 4-digit occupation $o \in O$, 4-digit industry $i \in I$ in commuting zone $m \in M$ in quarter $t \in T$. In our most demanding specification, we estimate an equation of the form:

$$\log(E_{o,i,m,t}) = \beta_L \log(\text{Labor HHI}_{o,m,t}) + \beta_P \log(\text{Product HHI}_{i,m,t}) + X'_{o,i,m,t}\lambda + \zeta_{o,i,m} + \Xi_t + \varepsilon_{o,i,m,t}$$

where β_L is the elasticity of the number of new hires with respect to labor market concentration, β_P is the elasticity of the number of new hires with respect to product market concentration, $\zeta_{o,i,m}$ are occupation by industry by commuting zone fixed effects, and λ is a vector of parameters associated with the control variables measured in $X'_{o,i,m,t}$, which include the mean age of the new hires, the share of these new hires which are men, the mean log(value-added per employee) across new hires, and the average log(firm size) in the firms recruiting the new hires. Ξ_t are time fixed effects whilst $\varepsilon_{o,i,m,t}$ is the error term. We provide standard errors clustered at the commuting zone level to account for common shocks and their persistence across time within a labor market.

Four specifications are presented with increasingly demanding fixed effects. The first provides only time and occupation fixed effects. The second adds commuting zone fixed effects. The third combines the two previous ones by also including occupation by commuting zone fixed effects. The final column goes further by including an occupation by industry by commuting zone fixed effect. Results for the ordinary least squares estimation are reported in Table 2.5.

Table 2.5: New Hires (OLS) : Baseline

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.120***	-0.0593***	-0.00478	-0.0937***
	(0.00378)	(0.00641)	(0.00350)	(0.00675)
Log(Product HHI)	-0.116***	-0.105***	-0.110***	-0.0429**
	(0.0124)	(0.0114)	(0.0117)	(0.0208)
Mean Age (in years)	-0.00336***	-0.00335***	-0.00328***	-0.000557***
	(0.0000701)	(0.0000707)	(0.0000712)	(0.0000586)
Share of Men	-0.0499***	-0.0527***	-0.0506***	0.00218
	(0.00615)	(0.00640)	(0.00577)	(0.00146)
Mean Log(Value Added per Employee)	0.0727***	0.0745***	0.0767***	-0.00502***
	(0.00378)	(0.00374)	(0.00359)	(0.00130)
Mean Log(Nb. Employees)	0.0511***	0.0501***	0.0491***	0.0132***
	(0.00200)	(0.00198)	(0.00180)	(0.00154)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.171	0.174	0.225	0.742
Adjusted R^2	0.171	0.174	0.205	0.683
N. Clusters	308	307	305	305
F	2521.9	1289.1	1007.8	225.3
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.00478 \times 0.1 \times 100 = -0.0478\%$.

The results presented are in line with the basic predictions of theory. We find that both labor market and product market concentration are negatively related with the number of recruited workers. This can be seen across the four specifications where nearly all relevant coefficients are statistically significant at the 1% level. However, magnitudes vary greatly across specifications. Column (1) suggests that a 10% increase in labor market concentration would lead to a 1.2% fall in employment. This effect falls to 0.0478% in column (3) where occupation by commuting zone fixed effects are included. For the product market, our results suggest that a 10% increase in the product market can lower employment by up to 1.1% (column (1)), or, more conservatively, by 0.4% (column (4)). We can also report positive coefficients associated with the value added per worker and the firm size. At the same time, the coefficients related to mean age and the share of men tend to be negative. Overall, the adjusted R^2 stays constant, rising significantly in the final column when occupation by industry by commuting zone fixed effects are introduced.

Table 2.6: New Hires (IV) : Baseline

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.432**	-0.312***	-0.325***	-0.585***
	(0.207)	(0.0820)	(0.0824)	(0.187)
Log(Product HHI)	-0.289***	-0.305***	-0.329***	-3.096***
	(0.0153)	(0.0167)	(0.0159)	(0.798)
Mean Age (in years)	-0.00239***	-0.00303***	-0.00300***	-0.000240
	(0.000728)	(0.000141)	(0.000124)	(0.000261)
Share of Men	-0.0589***	-0.0419***	-0.0424***	0.00523
	(0.00418)	(0.00981)	(0.00726)	(0.00360)
Mean Log(Value Added per Employee)	0.0817***	0.0599***	0.0620***	-0.00117
	(0.0182)	(0.00468)	(0.00461)	(0.00384)
Mean Log(Nb. Employees)	0.0653***	0.0668***	0.0691***	0.0290***
	(0.00296)	(0.00316)	(0.00381)	(0.00481)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	492.7	881.7	699.2	47.89
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new Employees and the mean age (in years) of the new Employees. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new Employees) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new Employees) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new Employees by approximately $-0.325 \times 0.1 \times 100 = -3.25\%$.

The impact of labor market concentration on the number of new hires could be biased by the mechanical effect that when fewer firms recruit, the labor HHI tends to be higher. This is why it is especially important to use an instrument to check the validity of these results. Using instrumental variables confirms the negative effect of labor and product market concentration on the number of new hires, as shown in Table 2.6. With instruments, the impact of labor market concentration appears to be even greater. Indeed, a 10% increase in labor market concentration leads to a 3.25% fall in new hires (column (3)). Similarly, we find that the impact of product market concentration has increased. The same column reports a negative and statistically significant coefficient according to which a 10% increase in product market concentration lowers the number of new hires by 3.29%. The final column reports a very large estimate for the impact of the product market concentration, perhaps due to the highly demanding number of included fixed effects. For this reason, we consider column (3) as our preferred specification because it is a conservative estimate within the class of instrumental variables estimates and can be said to be robust to occupation by commuting zone fixed effects.

In addition, we provide weighted regressions in Appendix 2.C. We rely on analytic weights based on (i)

the number of new hires and (ii) the mean number of new hires in that combination of occupation by industry by commuting zone. We take the latter to be the more accurate measure for it is constant across time. The instrumental variable estimates are provided in Table 2.C.2 and 2.C.4 respectively. In both cases, the main coefficients of interest are negative and statistically significant. More surprisingly, the magnitude of the effects are larger. For example, column (3) of Table 2.C.4 reports that a 10% increase in labor market concentration lowers the number of new hires by 14.4%. This suggests that the impact on employment flows are particularly important in large labor markets.

Finally, we rely on Poisson Regression to gauge the impact of excluding small labor markets. That is, for each combination of occupation by industry by commuting zone, we complete missing observations by setting the the number of new hires to zero. Unable to take the log of zero, we estimate Equation 2.3.3.1 by taking its exponential form as in Poisson or pseudo-Poisson regression. Results are provided in Table 2.C.5 in Appendix 2.C.[18] In comparison to the ordinary least squares results in the unweighted case presented in Table 2.5, the estimates are of the same sign but of much greater magnitude. Indeed, according to column (3), we find that a 10% increase in labor market concentration would lead a 4.3% fall in the number of new hires. This suggests that the impact of concentration is underestimated by virtue of a survival bias.

2.3.3.2 Heterogeneity

This last set of results provide compelling evidence that labor and product market concentration have a negative impact on the number of new hires. To better understand the implications, we now present two additional sets of regression which clarify the mechanisms at play in terms of exits and net employment.

First, we estimate our baseline specification using as dependent variable the number of new exits instead of the number of new hires. We define an exit as a job spell whose termination date falls before the last day of the quarter. The expected effect of concentration on new exits is unclear. On the one hand, an increase in labor market concentration could lead to a smaller workforce and therefore more exits. On the other hand, higher labor market concentration could reduce the separation rate because employees fear having to face a concentrated labor market. Results are in Appendix 2.C.5. Looking at the unweighted regressions, we observe that both the ordinary and two stage least squares estimates, presented respectively in Tables 2.C.12 and 2.C.13, provide negative and statistically significant estimates with regards to both labor and product market concentration. In the latter table, we find for our preferred specification (column (3)) that a 10% rise in labor market concentration lowers the number of exits by 1.9%. Similar results, though of greater magnitudes, are found when the observations are weighted by the number of exits or the mean number of exits (for a given combination of occupation, industry, and commuting zone) across time. This suggests that labor market concentration lowers employment flows, both in terms of entries and exits.

Second, we determine the net impact of labor market concentration on employment by using as dependent variable the net number of employees instead of the number of new hires. In this case, Monopsony Theory predicts more clearly that labor and product market concentration should be negatively related to net employment. We consider a job spell to be part of net employment if the starting date is within the measured quarter and the termination date falls after the end of the quarter. Appendix 2.C.6 displays the results. The ordinary least squares estimates are provided in Table 2.C.19 along with the two stage least squares in Table 2.C.20.

[18] Unfortunately, it was not computationally feasible to estimate this model using instrumental variables.

We find a negative and often statistically significant effect of both the labor and product market concentration on net employment. In our preferred specification relying on instrumental variables, we observe that a 10% increase in labor market concentration lowers net employment by 0.81%. The weighted regressions provide more pronounced results with magnitudes more than doubled. For example, the same column in Table 2.C.24, which relies on weighting each observation by the mean number of employees (for a combination of occupation, industry, and commuting zone) across time, shows that a 10% increase in labor market concentration lowers net employment by 2.22%. Therefore, we find results in line with monopsony theory in terms of net employment.

Overall, our analysis shows that wages are lower when labor market concentration increases, while wages are higher when product market concentration increases. The number of hires decrease with labor market concentration, while product market concentration often has a negative relationship to new hires.

2.3.4 Robustness, Sensitivity, and Alternative Specifications

2.3.4.1 Robustness

We now present various alternative specifications, which allows us to gauge the robustness of our results to different modelling choices. We first consider the impact of dropping the log(number of full-time equivalent employees in the hiring firm) from our specification. Given that the firm size is affected by market structures and that the latter is likely to be correlated with unobserables (such as labor market tightness), it is probable that the firm size is also endogenous. We provide our baseline estimates for hourly wages and new hires in Appendix 2.J. We focus on estimates identified using our instrumental variables. Generally speaking, the results are very similar. We nonetheless observe that the absence of firm size has lowered the magnitude of the coefficients associated with labor market concentration for the specification which includes both worker and firm fixed effects (column (6), Table 2.J.2). In terms of new hires, Table 2.J.4 shows very similar results to our baseline estimates. Similar observations can be made when we weight our observations, as in Table 2.J.8. Altogether, this evidence suggests that controlling for firm size is not significantly biasing our estimates.

Second, we examine the robustness of our results with regards to the changes in underlying sample. Indeed, observations are dropped when there are too few observations in the data to estimate their respective fixed effects (i.e, they are singletons). This means that, as we increase the number of fixed effects across our different specifications, the underlying sample can change. To gauge the effect of keeping the underlying sample fixed, we present in Appendix 2.D the estimates resulting from using the same data as required for the estimation of the most demanding specification. In the case of hourly wages, Table 2.D.2 provides the instrumental variable results using solely the data required to estimate column (6) with both worker and firm fixed effects. The results are broadly similar to those of our baseline. The same can be said for the estimates provided for new hires in Table 2.D.4 which relies on the sample required to estimate column (4) where we include both time and occupation by industry by commuting zone fixed effects. The use of weights in Table 2.D.8 does not lead to significant departures from our baseline estimates. We conclude that changes in underlying samples are not significantly affecting our results.

2.3.4.2 Sensitivity

We now consider the sensitivity of our estimates of the impact of labor market concentration in relation to potential measurement error in the product market concentration. As explained above, our measure relies on national sales instead of the more ideal local sales. Beyond our goal of assessing the presence of monopsony and monopoly power on the French labor market, this measurement error is problematic because it has the potential to bias our estimate of the impact of labor market concentration, by virtue of the correlation which exists between labor and product market concentration. For this reason, we propose two ways to gauge the importance of this measurement error.

First, we consider the impact of removing product market concentration from our covariates. This replaces a potential measurement error with an omitted variable. Although not a particularly decisive solution, it allows us to verify that the presence of product market concentration is not driving our results. The estimation results are presented in Appendix 2.E. Using instrumental variables to study the hourly wage, the estimates reported in Table 2.E.2 show negative and statistically significant (at the 1% level) estimates in relation to labor market concentration. However, the magnitude has now slightly fallen. Similarly, both the unweighted and weighted regressions on new hires, presented respectively in Tables 2.E.4 and 2.E.8, show comparable estimates to those of our baseline. For this reason, we conclude that the measurement error in the product market HHI is not driving the sign of the parameters associated with labor market concentration.

Second, we supply estimates of the impact of labor market concentration which are independent of any measurement error in the product market concentration index. To do so, we replace the latter with 4-digit industry by commuting zone by time fixed effects. This absorbs any of the variation measured by the product market HHI at the cost of a loss of precision induced by the need to estimate many more parameters. Moreover, the impact of product market concentration can no longer be ascertained. The results from this exercise are provided in Appendix 2.H. Table 2.H.2 provides the estimates based on instrumental variables when the dependent variable is the hourly wage. This table reports negative and statistically significant point estimates for labor market concentration in all specifications but those including worker fixed effects. The magnitudes appear to have fallen. In terms of the instrumental variable regressions on the number of new hires, the unweighted case reported in Table 2.H.4 provides estimates which are negative and statistically significant, of similar magnitude to those in our baseline. In contrast, in the weighted case shown in Table 2.H.8, the statistical significance of the parameters associated with labor market concentration falls to 5% or 10% levels. This is despite the coefficients being of similar magnitude to those in the base case. We conclude from this exercise that our measure of product market concentration has not significantly impacted our parameters measuring the the impact of labor market concentration on hourly wages and new hires.

2.3.4.3 Alternative Specifications

We conclude this section of the paper by considering two alternative strategies to assess the role of product market concentration in relation to hourly wages and new hires. First, we focus on tradeable industries. We assume the product market of those industries to be of a *global* (G), rather than *local*, nature. That is, the level of competition in these product markets can be assumed to be constant across the country. We can then construct an appropriate measure of concentration for these industries by relying on national sales. More formally, we consider firms in industry $i \in I$ (measured at the 4-digit level) at time $t \in T$ (national sales are

recorded at the yearly level). These firms are collected in a set $F_{i,t}$. For any firm j in this set, we observe the national sales during that year, denoted by $\mathrm{R}_{j,t}$. We can then define the *global* product market share as:

$$s^G_{j,t} = \frac{R_{j,t}}{\sum_{k \in F_{i,t}} R_{k,t}} \tag{2.3.3}$$

The *global* product market Herfindahl-Hirschman Index, $\mathrm{HHI}^G_{i,t}$, sums the squares of these market shares:

$$\mathrm{HHI}^G_{i,t} = \sum_{k \in F_{i,t}} \left\{ s^G_{k,t} \right\}^2 \tag{2.3.4}$$

Given the absence of variation at the commuting zone level, it is not possible to instrument this variable using the instrument described in Equation 2.3.2. We nonetheless provide in Appendix 2.F the results from using this measure on our full sample but only instrumenting labor market concentration.

This measure of concentration is relevant for industries with goods that can be traded across the country and internationally. We consider an industry to belong to this *tradeable sector* if over 5% of its sales are earned by export. Indeed, the latter provides evidence that the goods and services can effectively be moved and traded by the French firms. Choosing a relatively low threshold selects a subsample which is not so small such that our regressions would be necessarily underpowered. Moreover, in comparison to a method using both imports and exports, this method assures us that the French firms are actively trading their goods across geographic markets.

Appendix 2.G provides the estimates from running our regressions on the subsample of workers in industries included in our tradeable sector whilst relying on our *global* measure of product market concentration. Table 2.G.2 provides the results when using an instrumental variable to identify the effects of labor market concentration on hourly wages. Although the coefficient associated with our global measure of product market concentration is positive in nearly all specifications, it is only statistically significant in two cases. In particular, when we have time and occupation by commuting zone fixed effects in column (4), we observe that a 10% increase in the global product market HHI raises hourly wages by 0.1%. However, although the estimates for labor market concentration are almost all negative, none are statistically significant. This suggests that our estimation strategy provides noisy estimates, as reflected in the 76% drop in sample size in comparison to the sample size used for our baseline estimates in section 2.3.2.[19] Therefore, this exercise does not find conclusive evidence in favor an effect of the product market concentration on hourly wage.

However, this same exercise applied to the number of new hires provides more conclusive results. Table 2.G.4 displays the unweighted estimates from instrumenting the labor market concentration index in a regression where the dependent variable is the number of new hires. We find negative and statistically significant point estimates (at the 1% level) for both the labor and product market concentration. Whilst the magnitude for the labor market HHI is similar to the one observed in our baseline estimates, the size of the effects for the global product market HHI is much smaller. In our preferred specification, we find that a 10% increase in the latter lowers the number of new hires by 0.7%. This effect is more important when the regressions are weighted by the mean number of new hires for a given combination of industry by occupation by commuting zone. Indeed, as shown in column (3) of Table 2.G.8, a 10% increase in the global product market HHI lowers the number of new hires by 2%. Even so, one should note that we could not instrument this alternative measure of concentration

[19] This number is calculated on the basis of column (1).

and, given the pattern observed across our set of results (i.e, the increase in magnitudes observed when using instrumental variables for identification), these estimates can best be construed as forming lower bounds.

Second, we construct an alternative measure of product market concentration weighted by local employment for all industries.[20] This approach uses fully the available data by allowing us to provide greater importance to the sales of firms with many employees within a given labor market. To do so, we assume that the local turnover of a firm can be approximated as the national turnover taken in proportion to the local employment share of the firm. Although this approach also relies on a strong assumption, it does not make the same assumptions and therefore provides a natural robustness check. More formally, we consider firms with employees in commuting zone $m \in M$, in year $t \in T$ (national sales are recorded at the yearly level), in industry $i \in I$ (measured at the 4-digit level). These firms are collected in a set $V_{i,m,t}$. For any firm j in this set, we observe the national sales during that year, denoted by $\mathrm{R}_{j,t}$ and the total number of employees $N_{j,m,t}$ hired in that local market. We can then define the local turnover as as:

$$S_{j,m,t} = \frac{N_{j,m,t}}{\sum_{b \in M} N_{j,b,t}} \times R_{j,t} \tag{2.3.5}$$

and in turn the market share in terms of local employee (E) adjusted turnover:

$$s^{E}_{j,m,t} = \frac{S_{j,m,t}}{\sum_{k \in V_{i,m,t}} S_{k,m,t}} \tag{2.3.6}$$

The employment weighted product market Herfindahl-Hirschman Index, $\mathrm{HHI}^{E}_{i,m,t}$, sums the squares of these market shares:

$$\mathrm{HHI}^{E}_{i,m,t} = \sum_{k \in V_{i,m,t}} \left\{ s^{E}_{k,m,t} \right\}^{2} \tag{2.3.7}$$

Estimation results are available in Appendix 2.I. Table 2.I.2 provides the estimates using instrumental variables to look at the impact of concentration on hourly wages. The effect of labor market concentration is reported as negative and statistically significant at the 5% or 1% level although the magnitude has slightly fallen in comparison to our baseline estimates. The estimates for our employee weighted product market concentration index are positive and statistically significant in all specifications which do not include firm fixed effects. In particular, in our preferred specification, we find that a 10% increase in the latter would increase hourly wages by 0.4%. Looking at new hires, Table 2.I.4 provides the results from the unweighted regressions on the number of new hires identified through the use of instrumental variables. We find similar coefficients to those in our baseline results in terms of labor market concentration. For the employment weighted product market HHI, we observe that a 10% increase would imply a 1.3% fall in the number of new hires. Similar observations can be made when observations are weighted by the mean number of new hires for a given combination of occupation, industry, and commuting zone, and presented in Table 2.H.8. All in all, this exercise has shown that our results are robust to alternative ways to calculate and identify the impact of product market concentration.

2.4 Merger Simulation

In this section, we simulate counter-factual horizontal mergers. This exercise allows us to express our point estimates in a way relevant to a policy maker and, in particular, to a Competition Authority. Indeed, we take our results to have direct implications for losses in employment and in the wage bill which may result from a

[20] In unreported results, we also weighted market shares by employment shares before constructing a product market concentration index. This exercise provided similar results.

horizontal merger among the largest firms. As a consequence, we attempt to identify the industries in which workers are most vulnerable to corporate consolidation and provide a rough approximation of the effects of such mergers on labor markets. These predictions can be tested in future research.

To run these counter-factual horizontal mergers, we assume that the two largest firms in each 2-digit industry (in terms of full-time equivalent headcount) merge.[21] Such mergers would likely raise antitrust concerns regarding product market competition, and here we shed light on the additional concern that these mergers should raise by reducing labor market competition. We calculate the post-merger labor market HHI and predict the effect on the number of new hires and their wage bill. To calculate the former, we use our prior estimate of the impact of labor market concentration on new hires found in Table 2.6 (column 3). That is, we calculate the loss in new hires by assuming an elasticity with respect to labor market concentration of -0.325. To calculate the impact on the new hires' wage bill, we estimate the same regressions as for new hires, but rely on the wage bill as a dependent variable. Estimates are provided in Appendix 2.B.5.2 and in Table 2.B.12 for the unweighted two stage least squares case. In our preferred specification (column (3)), we find that a 10% increase in labor market concentration lowers the wage bill by 7.31%. We use this estimate for our merger simulations.

To do so, we make several modelling choices. First, we keep firm characteristics fixed. Second, we neglect the effects for workers who are already in employment. Indeed, the latter are legally protected by European Law (Transfers of Undertakings, 2001) from being fired as a result of the merger. Also, the effects on wages are smaller (Bassanini et al., 2020), given that French wages downward rigid. Third, we also keep levels of product market concentration constant and thus assume no impact on wages from changes in product market concentration.

We only use the coefficient on labor market HHI for four reasons. First, based on our point estimates, including the effects of the merger on product market concentration would only amplify the magnitude of our results in terms of the loss to the number of new hires and to their wage bill. This means that our simulations can be interpreted as lower bound estimates. Second, the product market HHI is not well measured. Third, the regressions with just the labor market HHI yield essentially the same coefficient as when we also include the product market HHI.[22] Fourth, from a policy perspective, the potentially positive effects of the product market HHI on *hourly* wages occur by an anticompetitive mechanism, so they should not be taken into account as offsets.

For each combination of 4-digit industry $i \in I$, 4-digit occupation $o \in O$, commuting zone $m \in M$ in quarter $t \in T$, we calculate the loss to the wage bill B_{iot} as:[23]

$$B_{iot} = \text{Observed Wage Bill}_{iot} \times \Big[\exp(-0.731 \log(\text{New HHI}/\text{Observed HHI})) - 1 \Big] \tag{2.4.1}$$

For each simulated industry merger, we recalculate the new level of labor market concentration per labor market (at the quarter by commuting zone by occupation level). This requires us to make use of our more exhaustive repeated cross section (DADS Salariés). We then calculate the loss to the number of new hires H_{iot} for each market according to the change in labor market concentration. We use the following formula :

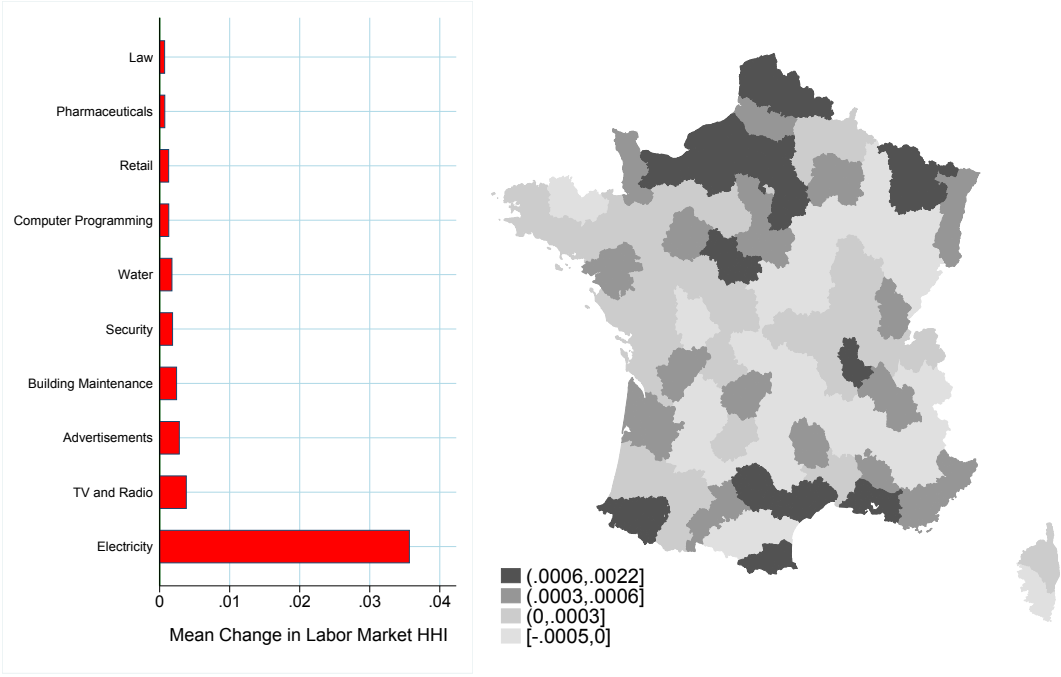
[21] This is in contrast to Jarosch et al. (2019) who simulate mergers by selecting the two largest employers within each of their areas. We take our approach as more indicative of the situation faced by Antitrust Authorities.

[22] See Appendix for alternative regression specifications which exclude the product market concentration.

[23] The formula can be read as writing the new wage bill as the old wage bill times a growth rate. The loss we want is the difference between the new wage bill and the old wage bill. So the loss is the old wage bill times the growth rate minus one.

$$H_{iot} = \text{Observed Nb. New Hires}_{iot} \times \left[\exp(-0.325 \log(\text{New HHI}/\text{Observed HHI})) - 1 \right] \quad (2.4.2)$$

Table 2.L.1 in Appendix 2.L reports descriptive statistics for the simulation. It shows that, after the merger of the two largest employers in each industry, labor market concentration would increase on average (weighted by industry employment) by 0.001 percentage points, that is, if a worker is in industry x, a merger between the top two employers in industry x would modify the average HHI of workers in industry x by 0.001 percentage points (this includes markets where the merger did not affect HHI because only one or none of the merging employers was present).



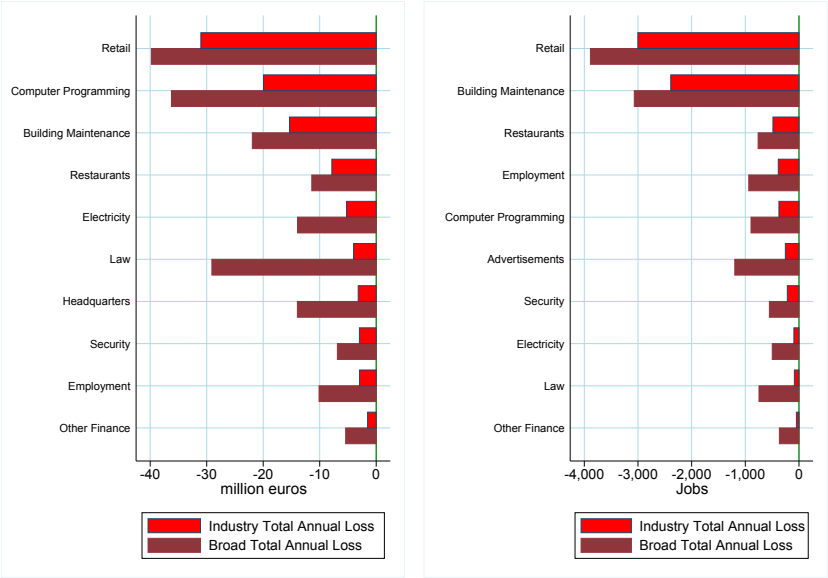
(a) (b)

Figure 2.6: Distribution of labor market concentration variation

Note: Figure 2.6(a) presents the top 10 greatest average percentage point change in level of labor market concentration per employee in a given industry as a result of the simulated merger in this same industry, over 2015. Figure 2.6(b): the average change in levels of labor market concentration described above are now averaged across French *départements*
Source: DADS Salariés and authors' calculations.

Nonetheless, the distribution of these merger effects on concentration is highly skewed across industries and geographies. In Figure 2.6(a), we graph the average change in labor market concentration following the merger, by industry and location. The electricity industry has the highest mean change in labor market concentration for new hires following a merger of the top two players. TV and radio has the second highest increase in HHI. Figure 2.6(b) displays the geographical location of the most affected workers. Workers most vulnerable to concentration increases from mergers appear to be in the rather disadvantaged areas of France, in the North and South of the country.

We now turn to the wage bill and employment impact of the simulated mergers. Figure 2.7(a) panel (a) shows the loss to the wage bill of newly hired workers in the industry that merged (in light red) and across all industries (in dark red). Mergers in Retail, Building Maintenance, and Computer Programming appear to be most harmful. In the retail industry, a merger by the top two players would lead to a yearly loss of over 30 million euros for workers in the industry, and 40 million euros when workers in all industries are taken into account.



(a) (b)

Figure 2.7: Total Wage Bill and Employment Effects

Figure 2.7(a) : Each line represents the sum of annual expected wage bill loss for new hires across France induced by a merger. It is calculated based on equation 2.4.1. Industry Total Annual Loss is calculated so as to include the loss to workers in the industry that merged (i.e, the impact on car repairers of a merger in the car repair industry). So, the merger in the retail industry would lead to a 30 million euro annual income loss to workers in the retail industry. Broad Total Annual Loss is calculated so as to include the loss to all workers in the economy, including those in the industry that merged. So, the merger in the retail industry would lead to 40 million euros in annual income loss across the economy.
Figure 2.7(b): Each line in light red represents the annual expected new jobs lost for new hires in *that* industry (i.e, a merger in the Building Maintenance industry would reduce annual recruitment by 2300 jobs in the Building Maintenance industry). It is calculated based on equation 2.4.2. Each line in dark red represents the annual expected new jobs lost for workers across France induced from a merger in *that* industry (i.e, a merger in the Building Maintenance industry would reduce annual recruitment by 3050 jobs across France). It is calculated based on equation 2.4.2.
Source: DADS Salariés and authors' calculations.

There are also significant employment losses due to increases in labor market concentration from the horizontal merger of the two top firms in each industry. Figure 2.7(b) panel (b) displays the loss of new hires in a given industry when there is a merger in that industry (e.g. the loss of jobs in the car repair industry induced by a merger by the car repair industry leaders) in light red. It also displays the overall loss from a merger in a given industry on all jobs in the economy (e.g. the loss of jobs induced by a merger in the car repair industry on all jobs in the French economy) in dark red. Retail and Building Maintenance appear at the top of the list, with the largest job losses, at almost 4,000 and over 3,000 jobs respectively.

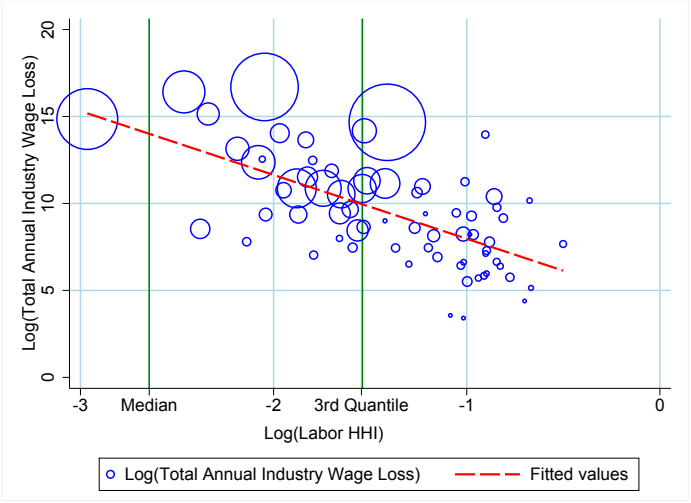
How are workers with different occupations affected by employment losses from mergers? Appendix 2.L Figures 2.1(a) and 2.1(b) display the job loss across the economy induced by industry mergers according to the

workers' occupation. Blue collar jobs (i.e, manual and non-manual workers) are most threatened by a merger in the Retail and Building Maintenance industry. White collar (i.e, managers) job loss is, as expected, smaller and mainly associated to a merger in the Computer Programming and Legal industry.

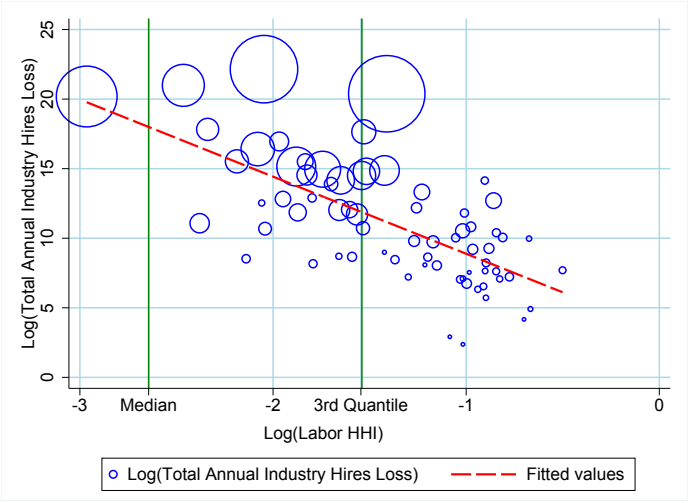
Do these effects correlate with workers in labor markets with high levels of labor market concentration? To answer this question, we plotted in Figure 2.8(a) the total wage bill loss to workers in the merging industry against the mean level of labor market concentration of the workers in that industry prior to the merger. Figure 2.8(b) displays the counterpart for industry new hires loss. We let the size of the indicator be proportional to the number of workers hired in that industry to distinguish size from intensity. These plots reveal that there can be significant losses for workers operating in areas of low labor market concentration. Industries with workers in highly concentrated labor markets have few hires to start with and, so, an increase in concentration does not scale up to large aggregate losses. Indeed, the best fit line is downward sloping, suggesting that losses are more pronounced in industries with workers from less concentrated labor markets. This can be explained in light of the log-log regression specification which emphasizes variations in the markets with initially low levels of labor market concentration.

While our simulation depends on a number of assumptions that may not always hold, our results offer a cautionary tale for antitrust enforcers. Once we add up effects in all markets, mergers in the industries with the highest levels of concentration are not necessarily the most damaging for workers' wages and for employment.[24]

[24] Figure 2.L.2 in Appendix 2.L shows these predicted losses relative to the total size of the industry. Despite these employment losses being small in relative terms, the number of lost hires in absolute terms remains important.



(a)



(b)

Figure 2.8: Total Wage and Employment Effects against Labor Market HHI

Figure 2.8(a) is constructed in the following way. Using equation 2.4.1, we calculate the total expected annual wage bill loss to new hires induced by the horizontal merger between the two largest employers of that industry. We associate to this value (on the x-axis) the log of the mean level of labor market concentration for new hires in the industry. Finally, the size of the marker for this observation is proportional to the number of new hires in the industry in 2015. The green lines represent the median and 75th percentile of the log labor market concentration levels.
Figure 2.8(b) is constructed in the same way. Using equation 2.4.2, we calculate the total expected annual hires loss in the industry by simulating the horizontal merger between the two largest employers of that industry. We associate to this value (on the x-axis) the log of the mean level of labor market concentration for new hires in the industry. Finally, we size the marker for this observation such that it is proportional to the number of new hires in the industry across 2015. The green lines represent the median and 75th percentile of the log labor market concentration levels.
Source: DADS Salariés and authors' calculations.

2.5 Conclusion

What are the labor market effects of labor and product market concentration? We leverage detailed French administrative data to show that labor market concentration decreases both the number of hires and the wages of new hires, as hypothesized by Monopsony Theory. Based on our instrumental variable estimates, a 10% increase in labor market concentration decreases hires by about 3.2% and the wages of new hires by nearly 0.5%, with less negative effects in more unionized industries. A 10% increase in product market concentration increases wages by 0.6%, with more positive effects in more unionized industries. The impact of product market concentration on wages is consistent with rent sharing. Product market concentration has a negative impact on the number of new hires.

Based on our estimate of the impact of labor market concentration on wages and the number of new hires, we can simulate the labor market impact of horizontal mergers between the two largest employers in each industry. We find that a horizontal merger has an impact not only on workers in the affected industry, but also on workers in other industries that share the same occupation: for a merger between the top two employers in the retail industry, about 30% of the impacts affect workers outside the retail industry. Compared to mergers in other industries, a merger between the top two employers in the retail industry would be the most damaging overall, with about 30 million euros annual decrease in the wages of new hires, and about a 3 000 decrease in annual hires.

Our comprehensive data allows us to show that employer market power has a substantial effect on labor market outcomes even in countries like France where union coverage is high and labor market institutions are protective of workers. Our findings suggest that antitrust and competition authorities should further scrutinize the effects of competition policy on workers.

2.A Descriptive Statistics

Table 2.A.1: Summary Statistics : Top 5 Highest Frequency Occupations

	count	min	max	p50	mean	sd
Transportation and Storage Worker (Unqualified)						
Gross Hourly Wage	83659	9	42.26316	11.70889	12.05	1.805
Labor Market Concentration (Business Group)	83659	.014444	1	.1028107	0.153	0.140
Labor Market Concentration (Firm)	83659	.0127772	1	.0894432	0.136	0.133
Product Market Concentration	83659	.0011498	1	.2144284	0.261	0.166
Product Market Concentration (Global)	83659	.00007	.3292432	.0010011	0.00320	0.00843
Age (in years)	83659	18	67	25	28.81	9.879
Gender	83659	0	1	1	0.603	0.489
Unionization Rate	17970	0	29.82	7.93	8.866	4.234
Nb. Full-Time Equivalent Employees	83659	.086	247296	2184	35333.1	51053.8
Value Added per Emp. (in nominal euros)	83659	.0001607	6171.898	27.90058	50.31	142.2
Cook (Beginner, Unqualified)						
Gross Hourly Wage	84119	9	41.6	11.15	11.46	1.851
Labor Market Concentration (Business Group)	84119	.0048275	1	.0398234	0.0678	0.0794
Labor Market Concentration (Firm)	84119	.0023827	1	.0214965	0.0464	0.0741
Product Market Concentration	84119	.0011498	1	.1853811	0.249	0.226
Product Market Concentration (Global)	84119	.0001886	.1382048	.0025168	0.00327	0.00422
Age (in years)	84119	18	67	23	27.73	10.25
Gender	84119	0	1	1	0.524	0.499
Unionization Rate	78669	0	23	10.32	10.43	0.978
Nb. Full-Time Equivalent Employees	84119	.125	165257.5	31	5327.0	24714.7
Value Added per Emp. (in nominal euros)	84119	.0080951	4649.973	31.77678	47.42	95.59
Wharehouse Person (Unqualified)						
Gross Hourly Wage	99540	9	42.70306	11.70662	12.08	1.720
Labor Market Concentration (Business Group)	99540	.0124863	1	.0953355	0.128	0.112
Labor Market Concentration (Firm)	99540	.0107163	1	.081216	0.112	0.105
Product Market Concentration	99540	.0011498	1	.2141763	0.249	0.144
Product Market Concentration (Global)	99540	.00007	.3913645	.0009887	0.00265	0.00769
Age (in years)	99540	18	67	24	28.15	9.716
Gender	99540	0	1	1	0.764	0.425
Unionization Rate	13800	0	23	5.65	7.613	4.094
Nb. Full-Time Equivalent Employees	99540	.035	247296	344.25	25479.4	48407.3
Value Added per Emp. (in nominal euros)	99540	.0033551	6171.898	29.03798	65.98	182.9
Storekeeper						
Gross Hourly Wage	105933	9	43.14286	11.39545	11.67	1.986
Labor Market Concentration (Business Group)	105933	.0191664	1	.0763395	0.108	0.0906
Labor Market Concentration (Firm)	105933	.0143594	1	.0610217	0.0916	0.0891
Product Market Concentration	105933	.0011572	1	.3428953	0.393	0.238
Product Market Concentration (Global)	105933	.00007	.3913645	.001912	0.00248	0.00388
Age (in years)	105933	18	67	23	26.66	9.524
Gender	105933	0	1	0	0.473	0.499
Unionization Rate	77882	0	18.27	7.93	7.986	1.398
Nb. Full-Time Equivalent Employees	105933	.152	165257.5	356	24393.6	44540.4
Value Added per Emp. (in nominal euros)	105933	.0012027	11243.25	36.61514	65.33	216.4
Cleaner						
Gross Hourly Wage	144134	9	42.2417	10.95527	11.34	2.226
Labor Market Concentration (Business Group)	144134	.0084093	1	.0496454	0.0777	0.0906
Labor Market Concentration (Firm)	144134	.004739	1	.045307	0.0707	0.0877
Product Market Concentration	144134	.0011498	1	.2358732	0.306	0.235
Product Market Concentration (Global)	144134	.00007	.3374615	.0033092	0.00416	0.00508
Age (in years)	144134	18	67	35	35.91	12.28
Gender	144134	0	1	0	0.342	0.475
Unionization Rate	130713	0	45.71	13.77	13.36	1.873
Nb. Full-Time Equivalent Employees	144134	.179	250825	504	6930.6	19112.0
Value Added per Emp. (in nominal euros)	144134	.0001975	4839.218	20.60366	29.80	74.85

Table 2.A.2: Distinct Units Across Datasets

	Panel Data (DADS Panel)	Repeated Cross-Section (DADS Salariés)
Commuting Zones	304	310
4-Digit Occupations	403	413
4-Digit Industries	601	620
Firms	337 254	1 098 708
Individuals	1 005 318	

Table 2.A.3: Summary Statistics : Aggregate Level

	count	min	max	p50	mean	sd
Nb. Hires	3175710	1	7201	1	3.985	26.26
Log(Nb. Hires)	3175710	0	8.881975	0	0.529	0.895
Labor Market Concentration (Business Group)	3175710	.0008135	1	.1428571	0.259	0.287
Labor Market Concentration (Firm)	3175710	.0005867	1	.1358025	0.252	0.287
Product Market Concentration (Local)	3175710	.0011498	1	.2977242	0.388	0.295
Product Market Concentration (Global)	3175710	.00007	1	.0037354	0.0107	0.0231
Mean Age (in years)	3175710	18	67	31	33.47	11.64
Share of Men	3175710	0	1	1	0.603	0.452
Mean Log(Value Added per Employee)	3175710	-33.82034	13.00192	4.086375	4.332	1.151
Mean Log(Nb. Employees)	3175710	-4.600158	12.3627	3.198673	3.664	2.646

Note: Each observation used to construct this table is a job spell at the occupation by industry by commuting zone level (DADS Salariés). Their associated level of labor concentration is calculated based on the repeated cross-section (DADS Salariés).
Source: DADS, FICUS, and authors' calculations.

2.B Hourly Wage

2.B.1 Unionization

Table 2.B.1: Hourly Wage (OLS) : Interaction with Unionization Rate

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.0142***	-0.0112***	-0.00465***	0.00145	-0.00361*	0.00249
	(0.00185)	(0.00178)	(0.00140)	(0.00137)	(0.00218)	(0.00305)
Log(Product HHI)	-0.00882***	-0.00661***	-0.00584***	0.00372	-0.00272	0.00196
	(0.00127)	(0.00104)	(0.00117)	(0.00244)	(0.00185)	(0.00267)
Unionization Rate	0.00182***	0.00196***	0.000901***	-0.00310***	0.000445	-0.00433***
	(0.000406)	(0.000403)	(0.000345)	(0.000938)	(0.000457)	(0.000929)
Log(Labor HHI) x Unionization Rate	0.000455***	0.000451***	0.000155	-0.000452***	0.0000366	-0.000606**
	(0.000123)	(0.000131)	(0.0000960)	(0.000129)	(0.000165)	(0.000262)
Log(Product HHI) x Unionization Rate	0.000329***	0.000409***	0.000369***	-0.000135	-0.0000824	-0.000185
	(0.000108)	(0.000102)	(0.000107)	(0.000132)	(0.000203)	(0.000263)
Age (in years)	0.00383***	0.00382***	0.00373***	0.00316***		
	(0.000420)	(0.000421)	(0.000492)	(0.000614)		
Gender	0.0333***	0.0323***	0.0313***	0.0273***		
	(0.00137)	(0.00145)	(0.00198)	(0.00292)		
Log(Value Added per Employee)	0.0275***	0.0270***	0.0251***	-0.00141*	0.0132***	0.000547
	(0.00116)	(0.00127)	(0.00134)	(0.000753)	(0.000528)	(0.00113)
Log(Nb. Employees)	0.00835***	0.00791***	0.00760***	-0.00573***	0.00621***	-0.00100
	(0.000297)	(0.000325)	(0.000303)	(0.00139)	(0.000517)	(0.00185)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.577	0.580	0.604	0.720	0.786	0.839
Adjusted R^2	0.577	0.579	0.595	0.677	0.679	0.719
N. Clusters	304	304	304	304	304	304
F	639.2	382.9	349.1	130.6	178.0	14.93
Observations	1476655	1476655	1463905	1296152	994474	831496

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. These two measures of concentration are interacted with the unionization rate, as reported by the *Enquête Réponse* (2011) at the 2-digit industry level; excluding the Temporary Employment Industry. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $(-0.00361 + 10 \times 0.0000366) \times 0.1 \times 100 = -0.03244\%$ for a worker in an industry with a 10% unionization rate.

Table 2.B.2: Hourly Wage (IV) : Interaction with Unionization Rate

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.0552**	-0.0574***	-0.0471***	-0.0230*	-0.0384***	-0.0141
	(0.0222)	(0.0147)	(0.0124)	(0.0124)	(0.0114)	(0.0110)
Log(Product HHI)	0.0287	0.0291*	0.0271**	-0.0665	0.00856*	-0.108
	(0.0189)	(0.0153)	(0.0108)	(0.0538)	(0.00459)	(0.0811)
Log(Labor HHI) x Unionization Rate	0.0000654	0.000116	0.000803***	0.000675***	0.00103***	0.000536
	(0.000387)	(0.000352)	(0.000283)	(0.000259)	(0.000350)	(0.000644)
Log(Product HHI) x Unionization Rate	0.00246***	0.00253***	0.00228***	0.00509**	0.000600	0.00555
	(0.000432)	(0.000418)	(0.000351)	(0.00251)	(0.000385)	(0.00445)
Unionization Rate	0.00319**	0.00361***	0.00539***	0.0130**	0.00480**	0.0160
	(0.00128)	(0.00120)	(0.00159)	(0.00539)	(0.00185)	(0.0115)
Age (in years)	0.00390***	0.00389***	0.00377***	0.00316***		
	(0.000380)	(0.000392)	(0.000483)	(0.000615)		
Gender	0.0316***	0.0301***	0.0290***	0.0274***		
	(0.00409)	(0.00246)	(0.00239)	(0.00287)		
Log(Value Added per Employee)	0.0267***	0.0252***	0.0239***	-0.00148	0.0128***	0.000835
	(0.00221)	(0.00215)	(0.00193)	(0.000931)	(0.000595)	(0.00114)
Log(Nb. Employees)	0.00451***	0.00440***	0.00371***	-0.00619***	0.00508***	-0.000593
	(0.000983)	(0.000596)	(0.000450)	(0.00178)	(0.000414)	(0.00182)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	478.7	586.7	693.8	206.8	313.1	3.345
Observations	1476655	1476655	1463905	1296152	994474	831496

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. These two measures of concentration are interacted with the unionization rate, as reported by the *Enquête Réponse* (2011) at the 2-digit industry level; excluding the Temporary Employment Industry. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $(-0.0384 + 10 \times 0.00103) \times 0.1 \times 100 = -0.281\%$ for a worker in an industry with a 10% unionization rate.

2.B.2 Non-permanent employees

Table 2.B.3: Hourly Wage (OLS) : Interaction with Part-Time Employment

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.0155***	-0.00865***	-0.00208***	-0.00154***	-0.00174	-0.00104
	(0.00306)	(0.00150)	(0.000767)	(0.000436)	(0.00181)	(0.000950)
Log(Product HHI)	-0.00422***	-0.000163	0.000533	0.000194	-0.00128	-0.00371***
	(0.00146)	(0.00129)	(0.00156)	(0.00113)	(0.00200)	(0.00142)
Log(Labor HHI) × Part-Time Employee	0.00625***	0.00243**	-0.00284**	-0.00238**	-0.00241**	-0.00335***
	(0.000956)	(0.000974)	(0.00121)	(0.00120)	(0.00122)	(0.000975)
Log(Product HHI) × Part-Time Employee	-0.00144***	-0.00226***	-0.00318***	0.00125	-0.00241***	0.00136
	(0.000549)	(0.000569)	(0.000530)	(0.00102)	(0.000827)	(0.00146)
Age (in years)	0.00408***	0.00404***	0.00391***	0.00329***		
	(0.000464)	(0.000457)	(0.000539)	(0.000559)		
Gender	0.0320***	0.0312***	0.0298***	0.0248***		
	(0.000718)	(0.000694)	(0.00126)	(0.00185)		
Log(Value Added per Employee)	0.0229***	0.0220***	0.0193***	-0.00186**	0.00952***	0.00145
	(0.00164)	(0.00171)	(0.00189)	(0.000800)	(0.000456)	(0.00154)
Log(Nb. Employees)	0.00836***	0.00802***	0.00788***	-0.000458	0.00750***	0.00278*
	(0.000222)	(0.000196)	(0.000160)	(0.00109)	(0.000148)	(0.00146)
Age (in years) × Part-Time Employee	-0.00158***	-0.00156***	-0.00154***	-0.00144***		
	(0.000136)	(0.000130)	(0.000149)	(0.000171)		
Gender × Part-Time Employee	-0.00527***	-0.00525***	-0.00496***	-0.00258		
	(0.00158)	(0.00154)	(0.00142)	(0.00182)		
Log(Value Added per Employee) × Part-Time Employee	0.0000218	0.000500	0.00151***	0.00239***	0.00236***	0.000139
	(0.000710)	(0.000699)	(0.000575)	(0.000851)	(0.000707)	(0.00172)
Log(Nb. Employees) × Part-Time Employee	-0.000578**	-0.000363	-0.000186	0.00202*	-0.000993***	-0.000195
	(0.000248)	(0.000250)	(0.000234)	(0.00112)	(0.000177)	(0.00163)
Quarter × Year × Part-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
4-digit Occupation × Part-Time FE	Yes	Yes	No	No	No	No
Commuting Zone × Part-Time FE	No	Yes	No	No	No	No
Commuting Zone × 4-digit Occupation × Part-Time FE	No	No	Yes	Yes	Yes	Yes
Firm × Part-Time FE	No	No	No	Yes	No	Yes
Worker × Part-Time FE	No	No	No	No	Yes	Yes
R^2	0.531	0.535	0.569	0.681	0.773	0.823
Adjusted R^2	0.531	0.535	0.557	0.642	0.646	0.690
N. Clusters	304	304	304	304	304	304
F	645.3	577.7	719.0	137.3	484.4	8.090
Observations	2225008	2225008	2201604	1975245	1465483	1259925

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variables regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-(0.00166 + 0.00242) \times 0.1 \times 100 = -0.0408\%$ for an employee with part-time status.

Table 2.B.4: Hourly Wage (IV) : Interaction with Part-Time Employment

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.153**	-0.108***	-0.0882***	-0.0622***	-0.0675**	-0.0136
	(0.0659)	(0.0148)	(0.0125)	(0.00749)	(0.0278)	(0.0158)
Log(Product HHI)	0.0885***	0.0678***	0.0641***	0.0198	0.0527***	-0.0649
	(0.0119)	(0.0201)	(0.0172)	(0.0394)	(0.0170)	(0.0590)
Log(Labor HHI) × Part-Time Employee	0.0335	0.00186	-0.00571	-0.00309	-0.0228	-0.0446**
	(0.0534)	(0.0119)	(0.0110)	(0.0126)	(0.0259)	(0.0177)
Log(Product HHI) × Part-Time Employee	0.0164	0.0312	0.0327	0.00416	0.00951	0.0563
	(0.0282)	(0.0215)	(0.0324)	(0.0390)	(0.0294)	(0.0732)
Age (in years)	0.00382***	0.00397***	0.00387***	0.00328***		
	(0.000234)	(0.000453)	(0.000531)	(0.000561)		
Gender	0.0227***	0.0304***	0.0283***	0.0247***		
	(0.00647)	(0.00138)	(0.00140)	(0.00192)		
Log(Value Added per Employee)	0.0157***	0.0184***	0.0169***	-0.00226**	0.00818***	0.00145
	(0.00210)	(0.00314)	(0.00317)	(0.000966)	(0.00120)	(0.00116)
Log(Nb. Employees)	0.00708***	0.00636***	0.00551***	-0.00108	0.00602***	0.00209**
	(0.000441)	(0.000310)	(0.000355)	(0.00164)	(0.000250)	(0.00102)
Age (in years) × Part-Time Employee	-0.00121***	-0.00141***	-0.00142***	-0.00144***		
	(0.000126)	(0.000167)	(0.000189)	(0.000174)		
Gender × Part-Time Employee	-0.00295	-0.00716***	-0.00622***	-0.00258		
	(0.00471)	(0.00197)	(0.00216)	(0.00179)		
Log(Value Added per Employee) × Part-Time Employee	0.00502**	0.00211***	0.00247***	0.00224**	0.00340***	-0.000406
	(0.00250)	(0.000802)	(0.000651)	(0.000912)	(0.000660)	(0.00149)
Log(Nb. Employees) × Part-Time Employee	-0.00272***	-0.00202***	-0.00238*	0.00164	-0.00179*	-0.000815
	(0.000666)	(0.000766)	(0.00125)	(0.00107)	(0.000911)	(0.00160)
Quarter × Year × Part-Time FE	Yes	Yes	Yes	Yes	Yes	Yes
4-digit Occupation × Part-Time FE	Yes	Yes	No	No	No	No
Commuting Zone × Part-Time FE	No	Yes	No	No	No	No
Commuting Zone × 4-digit Occupation × Part-Time FE	No	No	Yes	Yes	Yes	Yes
Firm × Part-Time FE	No	No	No	Yes	No	Yes
Worker × Part-Time FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	703.8	837.8	1434.3	113.0	353.6	5.853
Observations	2225008	2225008	2201604	1975245	1465483	1259925

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variables regression using the Log(Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-(0.0679 + 0.0198) \times 0.1 \times 100 = -0.877\%$.

2.B.3 First Stage Results

Table 2.B.5: Labor Market HHI: First Stage Results

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)
Instrument : Worker	0.193**	0.301***	0.325***	0.306***	0.332***	0.320***
	(0.0814)	(0.0729)	(0.0761)	(0.0763)	(0.0606)	(0.0577)
Instrument : Firm	0.0597***	0.00873	-0.00118	-0.00298	0.00182	0.0115
	(0.0133)	(0.00563)	(0.00122)	(0.0214)	(0.00232)	(0.0393)
Age (in years)	-0.000472	0.000334***	0.0000815	-0.000158**		
	(0.00104)	(0.000122)	(0.0000578)	(0.0000639)		
Gender	-0.126***	-0.0109**	-0.00789***	-0.00249**		
	(0.0234)	(0.00538)	(0.00181)	(0.00124)		
Log(Value Added per Employee)	-0.0589***	-0.0137**	0.000843	-0.00262	0.00185	-0.00207
	(0.00836)	(0.00636)	(0.00352)	(0.00314)	(0.00323)	(0.00425)
Log(Nb. Employees)	0.0147***	0.00841***	0.00557***	-0.00984	0.00412***	-0.0101
	(0.00243)	(0.00173)	(0.000670)	(0.00864)	(0.000895)	(0.00886)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.463	0.829	0.909	0.921	0.936	0.945
Adjusted R^2	0.462	0.829	0.907	0.914	0.910	0.915
N. Clusters	304	304	304	304	304	304
F	193.0	34.13	94.36	67.18	105.3	114.5
Observations	2388557	2388557	2375830	2201999	1885103	1706624

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a first-stage linear regression using the Log(Labor HHI) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in the number of firms hiring the given occupation is correlated with a decrease in labor market concentration by approximately $-(0.332) \times 0.1 \times 100 = -3.32\%$.

Table 2.B.6: Product Market HHI: First Stage Results

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)
Instrument : Worker	0.0124	0.0811***	0.0807***	0.0809***	0.100***	0.104***
	(0.00856)	(0.0145)	(0.0126)	(0.0201)	(0.0144)	(0.0290)
Instrument : Firm	0.232***	0.197***	0.191***	0.136***	0.185**	0.148**
	(0.0777)	(0.0708)	(0.0696)	(0.0514)	(0.0855)	(0.0646)
Age (in years)	-0.000471	0.000199	-0.000180*	0.0000368		
	(0.000581)	(0.000216)	(0.000102)	(0.0000450)		
Gender	-0.0675**	0.00684**	0.0185***	0.000778		
	(0.0273)	(0.00306)	(0.00640)	(0.000611)		
Log(Value Added per Employee)	-0.0273*	0.0119**	0.0205***	0.00618***	0.0163***	0.00471
	(0.0142)	(0.00606)	(0.00665)	(0.00211)	(0.00477)	(0.00304)
Log(Nb. Employees)	0.0317**	0.0285**	0.0354**	-0.00816	0.0245**	-0.00559
	(0.0138)	(0.0114)	(0.0144)	(0.00715)	(0.0105)	(0.00742)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.273	0.518	0.619	0.934	0.780	0.958
Adjusted R^2	0.273	0.518	0.612	0.928	0.690	0.935
N. Clusters	304	304	304	304	304	304
F	10.71	30.77	145.0	18.23	176.6	22.55
Observations	2388557	2388557	2375830	2201999	1885103	1706624

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a first-stage linear regression using the Log(Labor HHI) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in the number of firms hiring in this given industry is correlated with a decrease in the product market concentration by approximately $-(0.100) \times 0.1 \times 100 = -1.00\%$.

2.B.4 Business Group Labor Market Concentration

Table 2.B.7: Hourly Wage (OLS) : Business Group Labor Market Concentration

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI - Business Group)	-0.0141***	-0.00870***	-0.00294***	-0.00233***	-0.00257*	-0.00241***
	(0.00320)	(0.00152)	(0.000859)	(0.000489)	(0.00134)	(0.000815)
Log(Product HHI)	-0.00533***	-0.000820	-0.000619	0.00159	-0.00245	-0.00188***
	(0.00123)	(0.00121)	(0.00143)	(0.00161)	(0.00166)	(0.000642)
Age (in years)	0.00339***	0.00336***	0.00327***	0.00274***		
	(0.000417)	(0.000410)	(0.000470)	(0.000484)		
Gender	0.0306***	0.0296***	0.0287***	0.0242***		
	(0.00113)	(0.00106)	(0.00167)	(0.00248)		
Log(Value Added per Employee)	0.0230***	0.0223***	0.0202***	-0.000881	0.0112***	0.000707
	(0.00166)	(0.00173)	(0.00186)	(0.000721)	(0.000580)	(0.000749)
Log(Nb. Employees)	0.00819***	0.00792***	0.00780***	0.0000999	0.00722***	0.00163**
	(0.000279)	(0.000243)	(0.000203)	(0.00137)	(0.000121)	(0.000712)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.523	0.527	0.556	0.664	0.741	0.793
Adjusted R^2	0.523	0.527	0.548	0.629	0.633	0.677
N. Clusters	304	304	304	304	304	304
F	1061.4	934.5	1189.9	230.2	1105.3	13.00
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI - Business Group) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2 but measured at the business group level. This business group is identified using the *Enquête sur les liaisons financières entre sociétés* (LIFI) dataset. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration measured at the business group lowers wages by approximately $-0.00257 \times 0.1 \times 100 = -0.0257\%$.

Table 2.B.8: Hourly Wage (IV) : Business Group Labor Market Concentration

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI - Business Group)	-0.102***	-0.0934***	-0.0750***	-0.0543***	-0.0564***	-0.0219*
	(0.0127)	(0.00957)	(0.00778)	(0.00723)	(0.0119)	(0.0116)
Log(Local Product HHI)	0.0795***	0.0773***	0.0754***	0.0263	0.0654**	-0.0267
	(0.0235)	(0.0259)	(0.0245)	(0.0252)	(0.0297)	(0.0298)
Age (in years)	0.00338***	0.00336***	0.00328***	0.00273***		
	(0.000377)	(0.000399)	(0.000459)	(0.000486)		
Gender	0.0274***	0.0287***	0.0268***	0.0241***		
	(0.00138)	(0.00202)	(0.00194)	(0.00254)		
Log(Value Added per Employee)	0.0197***	0.0190***	0.0180***	-0.00118*	0.00980***	0.000689
	(0.00214)	(0.00312)	(0.00323)	(0.000677)	(0.00167)	(0.000731)
Log(Nb. Employees)	0.00596***	0.00565***	0.00466***	-0.000343	0.00511***	0.00117
	(0.000258)	(0.000250)	(0.000268)	(0.00159)	(0.000199)	(0.000756)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	1131.3	1380.1	2004.0	208.5	701.4	9.566
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI - Business Group) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2 but measured at the business group level. This business group is identified using the *Enquête sur les liaisons financières entre sociétés* (LIFI) dataset. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration measured at the business group lowers wages by approximately $-0.0543 \times 0.1 \times 100 = -0.543\%$.

2.B.5 Cross-Section

2.B.5.1 Hourly Wage

Table 2.B.9: Hourly Wage (OLS) : Baseline with Repeated Cross-Section

	(1)	(2)	(3)	(4)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0120***	-0.00729***	-0.00204***	-0.00170***
	(0.00148)	(0.000790)	(0.000647)	(0.000573)
Log(Product HHI)	-0.000123	0.00298***	0.00332***	-0.000436
	(0.000641)	(0.000569)	(0.000712)	(0.000550)
Age (in years)	0.00423***	0.00421***	0.00417***	0.00374***
	(0.000230)	(0.000224)	(0.000263)	(0.000260)
Gender	0.0292***	0.0283***	0.0282***	0.0242***
	(0.00109)	(0.00106)	(0.000428)	(0.000751)
Log(Value Added per Employee)	0.0134***	0.0136***	0.0130***	0.00122***
	(0.00216)	(0.00207)	(0.00204)	(0.000346)
Log(Number of Employees)	0.00981***	0.00937***	0.00901***	-0.00350***
	(0.000219)	(0.000193)	(0.000170)	(0.000802)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone $\times$ Occupation FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
R^2	0.543	0.546	0.569	0.676
Adjusted R^2	0.543	0.545	0.566	0.654
N. Clusters	307	307	305	305
F	1577.7	1680.0	1648.2	753.2
Observations	11576378	11576378	11566217	11255149

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (4): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00170 \times 0.1 \times 100 = -0.0170\%$.

Table 2.B.10: Hourly Wage (IV) : Baseline with Repeated Cross-Section

	(1)	(2)	(3)	(4)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.202***	-0.183***	-0.165***	-0.119***
	(0.0675)	(0.0520)	(0.0508)	(0.0326)
Log(Product HHI)	0.116***	0.0803***	0.0685***	0.0420*
	(0.00878)	(0.0171)	(0.0179)	(0.0233)
Age (in years)	0.00452***	0.00443***	0.00431***	0.00380***
	(0.000277)	(0.000301)	(0.000307)	(0.000285)
Gender	0.0142*	0.0251***	0.0248***	0.0237***
	(0.00853)	(0.000578)	(0.000482)	(0.000738)
Log(Value Added per Employee)	0.0172***	0.0152***	0.0144***	0.000252
	(0.00112)	(0.000838)	(0.000317)	(0.000425)
Log(Number of Employees)	0.00603***	0.00822***	0.00694***	-0.00474***
	(0.000549)	(0.00144)	(0.00128)	(0.00119)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone $\times$ Occupation FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
N. Clusters	307	307	305	305
F	1136.6	1421.3	2525.2	909.9
Observations	11576378	11576378	11566217	11255149

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variables using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (4): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.119 \times 0.1 \times 100 = -1.19\%$.

2.B.5.2 Wage Bill

Table 2.B.11: Wage Bill (OLS) : Baseline

	(1)	(2)	(3)	(4)
	Log(Wage Bill)	Log(Wage Bill)	Log(Wage Bill)	Log(Wage Bill)
Log(Labor HHI)	-0.199***	-0.105***	-0.0542***	-0.142***
	(0.00536)	(0.00608)	(0.00505)	(0.00980)
Log(Product HHI)	-0.108***	-0.0897***	-0.0945***	-0.0491**
	(0.0145)	(0.0119)	(0.0121)	(0.0193)
Mean Age (in years)	0.00144***	0.00146***	0.00173***	0.00420***
	(0.000356)	(0.000337)	(0.000372)	(0.000412)
Share of Men	0.193***	0.188***	0.185***	0.135***
	(0.00349)	(0.00354)	(0.00339)	(0.00293)
Mean Log(Value Added per Employee)	-0.0710***	-0.0686***	-0.0619***	0.0324***
	(0.0103)	(0.0104)	(0.00985)	(0.00303)
Mean Log(Nb. Employees)	0.0924***	0.0908***	0.0889***	0.0610***
	(0.00383)	(0.00374)	(0.00360)	(0.00380)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.331	0.334	0.370	0.673
Adjusted R^2	0.331	0.334	0.354	0.598
N. Clusters	308	307	305	305
F	1511.2	1099.3	1416.4	1037.9
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Wage Bill) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers the wage bill by approximately $-0.0542 \times 0.1 \times 100 = -0.542\%$.

Table 2.B.12: Wage Bill (IV) : Baseline

	(1)	(2)	(3)	(4)
	Log(Wage Bill)	Log(Wage Bill)	Log(Wage Bill)	Log(Wage Bill)
Log(Labor HHI)	-0.985***	-0.740***	-0.731***	-1.098***
	(0.335)	(0.112)	(0.107)	(0.142)
Log(Product HHI)	-0.117***	-0.155***	-0.178***	4.198
	(0.0319)	(0.0364)	(0.0342)	(5.867)
Mean Age (in years)	0.00314**	0.00186***	0.00197***	0.00462***
	(0.00151)	(0.000510)	(0.000455)	(0.000504)
Share of Men	0.160***	0.197***	0.186***	0.133***
	(0.00701)	(0.00765)	(0.00396)	(0.00482)
Mean Log(Value Added per Employee)	-0.0223	-0.0686***	-0.0624***	0.0283***
	(0.0324)	(0.00920)	(0.00962)	(0.00758)
Mean Log(Nb. Employees)	0.0925***	0.0960***	0.0988***	0.0484*
	(0.00531)	(0.00525)	(0.00638)	(0.0291)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	418.7	929.0	984.7	517.4
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variable regression using the Log(Wage Bill) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers the wage bill by approximately $-0.731 \times 0.1 \times 100 = -7.31\%$.

2.C New Hires

2.C.1 Weighted by New Hires

Table 2.C.1: New Hires (OLS): Weighted by New Hires

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.487***	0.128***	0.273***	0.0777***
	(0.0315)	(0.0377)	(0.0418)	(0.0253)
Log(Product HHI)	-0.456***	-0.289***	-0.262***	-0.137**
	(0.0562)	(0.0567)	(0.0498)	(0.0575)
Mean Age (in years)	-0.00661***	-0.00917***	-0.0107***	0.000343
	(0.00117)	(0.00116)	(0.00115)	(0.000273)
Share of Men	-0.129*	-0.230***	-0.209***	-0.00613
	(0.0665)	(0.0424)	(0.0153)	(0.00532)
Mean Log(Value Added per Employee)	0.114***	0.129***	0.134***	0.00457
	(0.0180)	(0.0139)	(0.0128)	(0.00710)
Mean Log(Nb. Employees)	0.192***	0.145***	0.126***	0.0448***
	(0.0170)	(0.0151)	(0.0129)	(0.00868)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.536	0.596	0.688	0.945
Adjusted R^2	0.536	0.596	0.679	0.932
N. Clusters	308	307	305	305
F	528.4	186.7	164.5	44.38
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Each observation is weighted by the number of new hires. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.273 \times 0.1 \times 100 = 2.73\%$.

Table 2.C.2: New Hires (IV) : Weighted by New Hires

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-1.787*	-1.799*	-2.102*	-1.165
	(1.011)	(0.983)	(1.200)	(1.719)
Log(Product HHI)	-0.638	-1.026***	-1.243***	-11.48
	(0.397)	(0.105)	(0.0455)	(10.38)
Mean Age (in years)	-0.000764	-0.00108	-0.00374	0.00295
	(0.00292)	(0.00473)	(0.00393)	(0.00975)
Share of Men	-0.625*	-0.231***	-0.187***	-0.0299
	(0.372)	(0.0423)	(0.0279)	(0.0341)
Mean Log(Value Added per Employee)	0.128**	0.0874**	0.0930**	0.0738***
	(0.0519)	(0.0339)	(0.0383)	(0.0262)
Mean Log(Nb. Employees)	0.251***	0.292***	0.305***	0.151*
	(0.0135)	(0.0487)	(0.0623)	(0.0812)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	122.5	128.3	239.8	44.11
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Each observation is weighted by the number of new hires. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers hires by approximately $-2.102 \times 0.1 \times 100 = 21.02\%$.

2.C.2 Weighted by Mean New Hires

Table 2.C.3: New Hires (OLS) : Weight by Mean Market New Hires

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.558***	-0.0198	0.0191	-0.0546**
	(0.0207)	(0.0275)	(0.0275)	(0.0229)
Log(Product HHI)	-0.416***	-0.281***	-0.265***	-0.228***
	(0.0670)	(0.0675)	(0.0590)	(0.0867)
Mean Age (in years)	-0.0105***	-0.0113***	-0.0119***	-0.00314***
	(0.000752)	(0.00106)	(0.00133)	(0.000340)
Share of Men	-0.0526	-0.128**	-0.146***	-0.00818
	(0.0687)	(0.0509)	(0.0211)	(0.00798)
Mean Log(Value Added per Employee)	0.120***	0.132***	0.138***	0.00899
	(0.0147)	(0.0132)	(0.0130)	(0.00557)
Mean Log(Nb. Employees)	0.154***	0.125***	0.109***	0.0349***
	(0.0167)	(0.0157)	(0.0137)	(0.00560)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.533	0.575	0.660	0.923
Adjusted R^2	0.533	0.575	0.651	0.905
N. Clusters	308	307	305	305
F	777.4	176.1	141.6	50.10
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.0191 \times 0.1 \times 100 = 0.191\%$.

Table 2.C.4: New Hires (IV) : Weighted by Mean Market New Hires

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-4.197	-1.646**	-1.444**	-1.521*
	(5.348)	(0.797)	(0.598)	(0.844)
Log(Product HHI)	-0.108	-1.063***	-1.199***	-5.305**
	(1.511)	(0.0669)	(0.0364)	(2.573)
Mean Age (in years)	0.000299	-0.00740***	-0.00944***	0.000514
	(0.0118)	(0.00207)	(0.000771)	(0.00264)
Share of Men	-0.995	-0.0778	-0.0909***	-0.0487***
	(1.445)	(0.0691)	(0.0176)	(0.0106)
Mean Log(Value Added per Employee)	0.234	0.0913***	0.0993***	0.0484***
	(0.214)	(0.0314)	(0.0329)	(0.0166)
Mean Log(Nb. Employees)	0.186***	0.236***	0.231***	0.0941***
	(0.0467)	(0.0314)	(0.0328)	(0.0214)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	60.21	216.5	329.1	133.1
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-1.444 \times 0.1 \times 100 = -14.44\%$.

2.C.3 Poisson Regression

Table 2.C.5: New Hires (OLS) : Poisson Regression

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.544***	-0.377***	-0.431***	-0.486***
	(0.0221)	(0.0216)	(0.0223)	(0.0788)
Log(Product HHI)	-0.370***	-0.339***	-0.354***	-0.446***
	(0.0313)	(0.0278)	(0.0284)	(0.0213)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	310	310	310	310
F	998.7	160.6	237.5	356.8
Observations	22016820	22016820	22016820	22016820

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.431 \times 0.1 \times 100 = -4.31\%$.

2.C.4 First Stage

Table 2.C.6: Labor HHI: First-Stage Results (Aggregated Level)

	(1)	(2)	(3)	(4)
	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)
Instrument : Worker	0.266***	0.399***	0.420***	0.402***
	(0.0930)	(0.0490)	(0.0473)	(0.0605)
Instrument : Firm	0.00463	-0.00745***	-0.00289***	0.0215***
	(0.00342)	(0.000961)	(0.000400)	(0.00430)
Mean Age (in years)	0.00207***	0.000461**	0.000131*	0.000382***
	(0.000711)	(0.000211)	(0.0000714)	(0.0000933)
Share of Men	-0.0420***	0.0110**	-0.00300***	0.00157
	(0.0138)	(0.00552)	(0.000914)	(0.00114)
Mean Log(Value Added per Employee)	0.0616***	0.00779***	0.00791***	0.00113
	(0.00603)	(0.00129)	(0.000322)	(0.000932)
Mean Log(Nb. Employees)	-0.000726	0.000515	0.00484***	0.00674***
	(0.00214)	(0.00131)	(0.000294)	(0.000611)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.367	0.803	0.893	0.909
Adjusted R^2	0.367	0.803	0.890	0.888
N. Clusters	308	307	305	305
F	125.5	488.4	780.5	367.2
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Labor HHI) as a dependent variable, providing "first-stage" estimates from two-stage instrumental variable regression. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in the number of firms hiring the given occupation is correlated with a decrease in labor market concentration by approximately $-(0.420) \times 0.1 \times 100 = -4.2\%$.

Table 2.C.7: Product HHI : First-Stage Results (Aggregated Level)

	(1)	(2)	(3)	(4)
	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)
Instrument : Worker	-0.107***	-0.0420***	-0.0456***	0.00702***
	(0.0391)	(0.0118)	(0.0144)	(0.00189)
Instrument : Firm	0.183***	0.179***	0.181***	0.0155
	(0.0142)	(0.0132)	(0.0138)	(0.00996)
Mean Age (in years)	0.00112**	0.000331	0.000357	0.0000144
	(0.000510)	(0.000210)	(0.000228)	(0.0000397)
Mean Share of Men	0.0196*	0.0357***	0.0384***	0.000746
	(0.0110)	(0.00692)	(0.00760)	(0.000769)
Mean Log(Value Added per Employee)	-0.0724***	-0.0947***	-0.0927***	0.00112
	(0.0100)	(0.00642)	(0.00647)	(0.000828)
Mean Log(Nb. Employees)	0.0727***	0.0728***	0.0736***	0.00430***
	(0.00609)	(0.00519)	(0.00549)	(0.000320)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.144	0.303	0.350	0.939
Adjusted R^2	0.144	0.303	0.333	0.926
N. Clusters	308	307	305	305
F	843.8	898.2	996.9	45.40
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Product HHI) as a dependent variable, providing "first-stage" estimates from two-stage instrumental variable regression. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in the number of firms hiring the given occupation is correlated with an increase in the product market concentration by approximately $-(-0.0456) \times 0.1 \times 100 = 0.456\%$.

Table 2.C.8: Labor HHI : First-Stage Results Weighted by Hires (Aggregated Level)

	(1)	(2)	(3)	(4)
	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)
Instrument : Worker	0.339***	0.341***	0.301***	0.284***
	(0.114)	(0.100)	(0.0973)	(0.0950)
Instrument : Firm	0.0708***	0.00341	-0.00527***	0.0195
	(0.0209)	(0.00319)	(0.00145)	(0.0263)
Mean Age (in years)	0.00361***	0.00343***	0.00237***	0.00338***
	(0.00137)	(0.000636)	(0.000431)	(0.000747)
Share of Men	-0.358***	-0.0248	-0.0334***	-0.0276***
	(0.101)	(0.0196)	(0.00288)	(0.00338)
Mean Log(Value Added per Employee)	0.00369	-0.00476**	0.000583	0.0102***
	(0.00531)	(0.00222)	(0.000895)	(0.00244)
Mean Log(Nb. Employees)	0.0275***	0.0361***	0.0265***	0.0575***
	(0.00375)	(0.00208)	(0.00233)	(0.00754)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.442	0.825	0.913	0.922
Adjusted R^2	0.442	0.825	0.910	0.905
N. Clusters	308	307	305	305
F	370.4	113.5	433.8	183.5
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Labor HHI) as a dependent variable, providing "first-stage" estimates from two-stage instrumental variable regression. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. Each observation is weighted by the number of hires. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in the number of firms hiring the given occupation is correlated with a decrease in labor market concentration by approximately $-(0.301) \times 0.1 \times 100 = -3.01\%$.

Table 2.C.9: Product HHI (OLS) : First-Stage Results Weighted by Hires (Aggregated Level)

	(1)	(2)	(3)	(4)
	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)
Instrument : Worker	0.0124	-0.000680	-0.00634	0.0255***
	(0.0492)	(0.0345)	(0.0321)	(0.00513)
Instrument : Firm	0.237***	0.190***	0.187***	0.0188
	(0.0519)	(0.0351)	(0.0321)	(0.0207)
Mean Age (in years)	-0.000267	0.000162	-0.0000554	-0.000196*
	(0.000475)	(0.00111)	(0.00107)	(0.000106)
Share of Men	-0.156	0.0634***	0.109***	0.000736
	(0.116)	(0.0145)	(0.0119)	(0.00216)
Mean Log(Value Added per Employee)	-0.0759***	-0.0792***	-0.0758***	0.00526
	(0.0211)	(0.0227)	(0.0241)	(0.00375)
Mean Log(Nb. Employees)	0.0866***	0.0929***	0.109***	0.00351***
	(0.0152)	(0.0153)	(0.0207)	(0.000633)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.236	0.501	0.596	0.954
Adjusted R^2	0.236	0.501	0.585	0.944
N. Clusters	308	307	305	305
F	523.4	826.0	794.9	39.80
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Product HHI) as a dependent variable, providing "first-stage" estimates from two-stage instrumental variable regression. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of hires. Standard Errors are clustered at the commuting zone level. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in the number of firms hiring the given occupation is correlated with an increase in the product market concentration by approximately $-(-0.00634) \times 0.1 \times 100 = 0.0634\%$.

Table 2.C.10: Labor HHI (OLS) : First-Stage Results Weighted by Mean Hires

	(1)	(2)	(3)	(4)
	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)	Log(Labor HHI)
Instrument : Worker	0.0404	0.309***	0.352***	0.328***
	(0.134)	(0.0861)	(0.0807)	(0.0836)
Instrument : Firm	0.0562***	-0.000434	-0.00665***	0.0231
	(0.0183)	(0.00289)	(0.000640)	(0.0154)
Mean Age (in years)	0.00272***	0.00179***	0.00112***	0.00174***
	(0.000846)	(0.000452)	(0.000294)	(0.000468)
Share of Men	-0.267***	0.00385	-0.0162***	-0.0180***
	(0.0779)	(0.0162)	(0.00172)	(0.00256)
Mean Log(Value Added per Employee)	0.0195***	0.000300	0.00393***	0.00493**
	(0.00449)	(0.00177)	(0.000647)	(0.00207)
Mean Log(Nb. Employees)	0.0138***	0.0198***	0.0122***	0.0339***
	(0.00349)	(0.00136)	(0.000913)	(0.00337)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.416	0.824	0.906	0.911
Adjusted R^2	0.416	0.824	0.904	0.890
N. Clusters	308	307	305	305
F	175.8	86.76	654.9	190.2
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Labor HHI) as a dependent variable, providing "first-stage" estimates from two-stage instrumental variable regression. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. Each observation is weighted by the mean number of hires across time. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in the number of firms hiring the given occupation is correlated with a decrease in labor market concentration by approximately $-(0.352) \times 0.1 \times 100 = -3.52\%$.

Table 2.C.11: Product HHI (OLS) : First-Stage Results Weighted by Mean Hire

	(1)	(2)	(3)	(4)
	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)	Log(Product HHI)
Instrument : Worker	-0.252***	-0.0650***	-0.0536***	0.0347***
	(0.0683)	(0.0196)	(0.0198)	(0.00536)
Instrument : Firm	0.222***	0.185***	0.184***	0.0312
	(0.0464)	(0.0326)	(0.0301)	(0.0242)
Mean Age (in years)	0.000120	-0.0000550	-0.000188	0.000133
	(0.000308)	(0.000735)	(0.000673)	(0.000202)
Share of Men	-0.107	0.0604***	0.0889***	-0.00303**
	(0.0809)	(0.00685)	(0.0114)	(0.00141)
Mean Log(Value Added per Employee)	-0.0720***	-0.0816***	-0.0769***	0.00692*
	(0.0204)	(0.0201)	(0.0215)	(0.00391)
Mean Log(Nb. Employees)	0.0869***	0.0899***	0.101***	0.00270
	(0.0149)	(0.0140)	(0.0179)	(0.00168)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.226	0.480	0.568	0.940
Adjusted R^2	0.226	0.480	0.557	0.927
N. Clusters	308	307	305	305
F	354.2	345.9	315.1	36.14
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Product HHI) as a dependent variable, providing "first-stage" estimates from two-stage instrumental variable regression. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of hires across time. Standard Errors are clustered at the commuting zone level. The two instruments of relevance are denoted by *Instrument: Worker* and *Instrument: Industry*, as described in Section 2.3.2. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in the number of firms hiring the given occupation is correlated with an increase in the product market concentration by approximately $-(-0.0536) \times 0.1 \times 100 = 0.536\%$.

2.C.5 Exits

2.C.5.1 Baseline

Table 2.C.12: Exits (OLS) : Baseline

	(1)	(2)	(3)	(4)
	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)
Log(Labor HHI)	-0.102***	-0.0807***	-0.0224***	-0.0701***
	(0.00243)	(0.00472)	(0.00292)	(0.00350)
Log(Product HHI)	-0.107***	-0.102***	-0.107***	-0.0418**
	(0.0125)	(0.0128)	(0.0127)	(0.0204)
Mean Age (in years)	-0.00359***	-0.00366***	-0.00379***	-0.0000769
	(0.000129)	(0.000126)	(0.000122)	(0.000116)
Share of Men	-0.0621***	-0.0610***	-0.0577***	-0.00531**
	(0.00757)	(0.00769)	(0.00700)	(0.00265)
Mean Log(Value Added per Employee)	-0.0211***	-0.0232***	-0.0277***	0.0180***
	(0.00309)	(0.00316)	(0.00275)	(0.00167)
Mean Log(Nb. Employees)	0.0780***	0.0782***	0.0786***	0.0279***
	(0.00354)	(0.00354)	(0.00338)	(0.00150)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.169	0.172	0.217	0.675
Adjusted R^2	0.169	0.171	0.197	0.592
N. Clusters	305	304	304	304
F	2095.6	979.9	512.1	302.7
Observations	2271453	2271452	2261341	1889175

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear regression using the Log(Nb. Exits) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new quitters. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new exits) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new exits) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers exits by approximately $-0.0224 \times 0.1 \times 100 = -0.224\%$.

Table 2.C.13: Exits (IV) : Baseline

	(1)	(2)	(3)	(4)
	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)
Log(Labor HHI)	-0.216***	-0.173***	-0.197***	0.00480
	(0.0350)	(0.0288)	(0.0314)	(0.283)
Log(Product HHI)	-0.292***	-0.307***	-0.337***	-18.59
	(0.0183)	(0.0132)	(0.0134)	(14.12)
Mean Age (in years)	-0.00253***	-0.00270***	-0.00276***	0.00175**
	(0.000199)	(0.000127)	(0.000151)	(0.000728)
Share of Men	-0.0613***	-0.0553***	-0.0529***	0.0197
	(0.0144)	(0.00983)	(0.00951)	(0.0223)
Mean Log(Value Added per Employee)	-0.0316***	-0.00800**	-0.00846***	0.183
	(0.00861)	(0.00321)	(0.00298)	(0.118)
Mean Log(Nb. Employees)	0.0937***	0.0920***	0.0946***	0.125*
	(0.00506)	(0.00456)	(0.00479)	(0.0712)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	305	304	304	304
F	629.9	990.0	959.7	10.07
Observations	2271453	2271452	2261341	1889175

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from instrumental variable regression using the Log(Nb. Exits) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new quitters. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new exits) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new exits) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers exits by approximately $-0.197 \times 0.1 \times 100 = -1.97\%$.

2.C.5.2 Weighted by New Hires

Table 2.C.14: Exits (OLS) : Weighted by Market Exits

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.583***	-0.209***	-0.187***	-0.173***
	(0.0254)	(0.0195)	(0.0240)	(0.0178)
Log(Product HHI)	-0.341***	-0.240***	-0.241***	-0.108**
	(0.0613)	(0.0610)	(0.0521)	(0.0436)
Mean Age (in years)	-0.00130	-0.00411**	-0.00865***	0.00338***
	(0.00220)	(0.00160)	(0.000868)	(0.000994)
Share of Men	-0.182**	-0.276***	-0.307***	-0.0287***
	(0.0923)	(0.0718)	(0.0194)	(0.00710)
Mean Log(Value Added per Employee)	-0.102**	-0.138***	-0.149***	0.0690***
	(0.0402)	(0.0401)	(0.0290)	(0.00824)
Mean Log(Nb. Employees)	0.203***	0.185***	0.171***	0.0844***
	(0.0219)	(0.0199)	(0.0205)	(0.00792)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.631	0.655	0.747	0.948
Adjusted R^2	0.630	0.654	0.740	0.935
N. Clusters	305	304	304	304
F	571.2	216.3	203.5	121.8
Observations	2271453	2271452	2261341	1889175

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear regression using the Log(Nb. Exits) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of exits. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new quitters. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new exits) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new exits) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers exits by approximately $-0.187 \times 0.1 \times 100 = -1.87\%$.

Table 2.C.15: Exits (IV) : Weighted by Market Exits

	(1)	(2)	(3)	(4)
	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)
Log(Labor HHI)	-0.690***	-0.617***	-0.553***	-1.396
	(0.177)	(0.0757)	(0.0627)	(1.166)
Log(Product HHI)	-0.631***	-0.718***	-0.983***	4.011
	(0.0713)	(0.0788)	(0.0693)	(8.778)
Mean Age (in years)	0.00193	0.00209	-0.000344	0.00203
	(0.00153)	(0.00195)	(0.00150)	(0.00214)
Share of Men	-0.267***	-0.243***	-0.285***	-0.0362**
	(0.0637)	(0.0810)	(0.0303)	(0.0142)
Mean Log(Value Added per Employee)	-0.0979	-0.0926**	-0.0518	0.0174
	(0.0616)	(0.0446)	(0.0318)	(0.134)
Mean Log(Nb. Employees)	0.228***	0.234***	0.248***	0.0650
	(0.0234)	(0.0303)	(0.0400)	(0.0625)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	305	304	304	304
F	65.74	65.25	115.3	25.04
Observations	2271453	2271452	2261341	1889175

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from instrumental variable regression using the Log(Nb. Exits) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of exits. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new quitters. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new exits) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new exits) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers exits by approximately $-0.553 \times 0.1 \times 100 = -5.53\%$.

2.C.5.3 Weighted by Mean New Hires

Table 2.C.16: Exits (OLS) : Weighted by Mean Market Exits

	(1)	(2)	(3)	(4)
	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)
Log(Labor HHI)	-0.517***	-0.148***	-0.191***	-0.213***
	(0.0202)	(0.0167)	(0.0124)	(0.0149)
Log(Product HHI)	-0.374***	-0.270***	-0.263***	-0.173**
	(0.0768)	(0.0717)	(0.0614)	(0.0670)
Mean Age (in years)	-0.00687***	-0.00821***	-0.0111***	-0.00168**
	(0.00103)	(0.000818)	(0.000901)	(0.000675)
Share of Men	-0.0449	-0.100**	-0.135***	-0.0105
	(0.0658)	(0.0486)	(0.0122)	(0.00890)
Mean Log(Value Added per Employee)	-0.133***	-0.157***	-0.139***	0.0651***
	(0.0292)	(0.0306)	(0.0221)	(0.00558)
Mean Log(Nb. Employees)	0.158***	0.142***	0.128***	0.0583***
	(0.0194)	(0.0169)	(0.0155)	(0.00355)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.544	0.572	0.655	0.874
Adjusted R^2	0.544	0.572	0.647	0.842
N. Clusters	305	304	304	304
F	643.5	156.8	168.1	196.7
Observations	2271453	2271452	2261341	1889175

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Exits) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Each observation is weighted by the mean number of new exits across time for a given combination of industry, occupation, and commuting zone. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new exits. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new exits) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new exits) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new exits by approximately $-0.191 \times 0.1 \times 100 = -1.91\%$.

Table 2.C.17: Exits (IV) : Weighted by Mean Market Exits

	(1)	(2)	(3)	(4)
	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)	Log(Nb. Exits)
Log(Labor HHI)	-0.465***	-0.463***	-0.681***	-3.165
	(0.0957)	(0.0736)	(0.148)	(3.107)
Log(Product HHI)	-0.718***	-0.710***	-0.881***	20.05
	(0.0628)	(0.0651)	(0.0623)	(37.22)
Mean Age (in years)	-0.00411***	-0.00456***	-0.00684***	-0.00929
	(0.000893)	(0.000839)	(0.000593)	(0.00961)
Share of Men	-0.0693	-0.0786	-0.120***	-0.0649
	(0.0624)	(0.0567)	(0.0189)	(0.0976)
Mean Log(Value Added per Employee)	-0.114***	-0.125***	-0.0734***	-0.105
	(0.0310)	(0.0334)	(0.0211)	(0.329)
Mean Log(Nb. Employees)	0.187***	0.183***	0.185***	-0.0168
	(0.0248)	(0.0249)	(0.0306)	(0.154)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	305	304	304	304
F	54.58	89.25	134.4	5.815
Observations	2271453	2271452	2261341	1889175

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Exits) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Each observation is weighted by the mean number of new exits across time for a given combination of industry, occupation, and commuting zone. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new exits. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new exits) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new exits) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new exits by approximately $-1.444 \times 0.1 \times 100 = -14.44\%$.

2.C.5.4 Poisson Regression

Table 2.C.18: Exits (OLS) : Poisson Regression

	(1)	(2)	(3)	(4)
	Nb. Exits	Nb. Exits	Nb. Exits	Nb. Exits
Log(Labor HHI)	-0.518***	-0.484***	-0.551***	-0.664***
	(0.00759)	(0.0118)	(0.0232)	(0.0455)
Log(Product HHI)	-0.263***	-0.259***	-0.270***	-0.384***
	(0.0223)	(0.0216)	(0.0220)	(0.0224)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	305	305	305	305
F	2451.0	836.4	329.2	193.0
Observations	15259580	15259580	15259580	15259580

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Exits as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new exits by approximately $-0.551 \times 0.1 \times 100 = -5.51\%$.

2.C.6 Net Employment

2.C.6.1 Baseline

Table 2.C.19: Net Employment (OLS) : Baseline

	(1)	(2)	(3)	(4)
	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)
Log(Labor HHI)	-0.114***	-0.106***	-0.0118***	-0.0274***
	(0.0120)	(0.0151)	(0.00130)	(0.00500)
Log(Product HHI)	-0.159***	-0.154***	-0.160***	-0.0477**
	(0.0142)	(0.0136)	(0.0159)	(0.0205)
Mean Age (in years)	-0.00322***	-0.00327***	-0.00318***	-0.00272***
	(0.000325)	(0.000338)	(0.000219)	(0.0000635)
Share of Men	-0.119***	-0.116***	-0.110***	-0.000937
	(0.00352)	(0.00424)	(0.00404)	(0.00157)
Mean Log(Value Added per Employee)	0.0743***	0.0703***	0.0653***	-0.00662***
	(0.00407)	(0.00364)	(0.00309)	(0.00127)
Mean Log(Nb. Employees)	0.102***	0.101***	0.100***	0.0425***
	(0.00412)	(0.00413)	(0.00404)	(0.00219)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.122	0.124	0.168	0.881
Adjusted R^2	0.121	0.124	0.163	0.863
N. Clusters	306	305	304	304
F	1297.6	1322.1	1434.2	733.5
Observations	11750180	11750179	11743710	11330634

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear regression using the Log(Net Number of Employees) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the employees and the mean age (in years) of the employees. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across employees) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across employees) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers net employment by approximately $-0.0118 \times 0.1 \times 100 = -0.118\%$.

Table 2.C.20: Net Employment (IV) : Baseline

	(1)	(2)	(3)	(4)
	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)
Log(Labor HHI)	0.00167	-0.0247***	-0.0613***	-0.104***
	(0.0230)	(0.00686)	(0.00622)	(0.00962)
Log(Local Product HHI)	-0.308***	-0.312***	-0.346***	-1.276***
	(0.0216)	(0.0233)	(0.0313)	(0.369)
Mean Age (in years)	-0.00304***	-0.00298***	-0.00271***	-0.00264***
	(0.000339)	(0.000303)	(0.000135)	(0.0000754)
Share of Men	-0.110***	-0.109***	-0.0994***	-0.0000168
	(0.00395)	(0.00500)	(0.00588)	(0.00182)
Mean Log(Value Added per Employee)	0.1000***	0.0868***	0.0847***	0.00193
	(0.0122)	(0.00423)	(0.00402)	(0.00247)
Mean Log(Nb. Employees)	0.119***	0.118***	0.120***	0.0525***
	(0.00596)	(0.00598)	(0.00653)	(0.00202)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	306	305	304	304
F	1119.8	1283.0	1004.9	473.9
Observations	11750180	11750179	11743710	11330634

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from instrumental variable regression using the Log(Net Number of Employees) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the employees and the mean age (in years) of the employees. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across employees) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across employees) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers net employment by approximately $-0.0613 \times 0.1 \times 100 = -0.613\%$.

2.C.6.2 Weighted by Employment

Table 2.C.21: Net Employment (OLS) : Weighted by Market Employment

	(1)	(2)	(3)	(4)
	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)
Log(Labor HHI)	-0.450***	-0.215***	-0.0314***	-0.0436***
	(0.0167)	(0.0105)	(0.00444)	(0.00847)
Log(Product HHI)	-0.505***	-0.432***	-0.420***	-0.110**
	(0.0832)	(0.0882)	(0.0808)	(0.0449)
Mean Age (in years)	-0.0196***	-0.0208***	-0.0209***	-0.00572***
	(0.00380)	(0.00413)	(0.00379)	(0.000206)
Share of Men	-0.376***	-0.411***	-0.399***	-0.0114*
	(0.0890)	(0.0774)	(0.0514)	(0.00601)
Mean Log(Value Added per Employee)	0.0696***	0.0272	0.00242	0.0105
	(0.0208)	(0.0199)	(0.0138)	(0.0116)
Mean Log(Nb. Employees)	0.292***	0.275***	0.234***	0.107***
	(0.0203)	(0.0209)	(0.0219)	(0.00895)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.517	0.533	0.625	0.980
Adjusted R^2	0.517	0.533	0.623	0.977
N. Clusters	306	305	304	304
F	1961.8	602.0	535.2	465.6
Observations	11750180	11750179	11743710	11330634

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear regression using the Log(Nb. of Employees) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of employees. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new quitters. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across employees) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across employees) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers net employment by approximately $-0.0314 \times 0.1 \times 100 = -0.314\%$.

Table 2.C.22: Net Employment (IV): Weighted by Market Employment

	(1)	(2)	(3)	(4)
	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)
Log(Labor HHI)	-0.0755	-0.180***	-0.208***	-0.223***
	(0.0619)	(0.0595)	(0.0604)	(0.0796)
Log(Product HHI)	-0.983***	-0.898***	-0.933***	-0.748***
	(0.0283)	(0.0515)	(0.0591)	(0.0778)
Mean Age (in years)	-0.0172***	-0.0169***	-0.0166***	-0.00511***
	(0.00329)	(0.00347)	(0.00279)	(0.000204)
Share of Men	-0.317***	-0.358***	-0.315***	-0.00894
	(0.0776)	(0.0855)	(0.0672)	(0.00743)
Mean Log(Value Added per Employee)	0.161***	0.0860***	0.0752***	0.0196**
	(0.0392)	(0.0321)	(0.0234)	(0.00968)
Mean Log(Nb. Employees)	0.368***	0.347***	0.319***	0.123***
	(0.0198)	(0.0249)	(0.0318)	(0.0113)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	306	305	304	304
F	922.8	398.4	560.7	348.2
Observations	11750180	11750179	11743710	11330634

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from instrumental variables regression using the Log(Nb. of Employees) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of employees. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new quitters. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across employees) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across employees) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers net employment by approximately $-0.208 \times 0.1 \times 100 = -2.08\%$.

2.C.6.3 Weighted by Mean Employment

Table 2.C.23: Net Employment (OLS) : Weighted by Mean Market Employment

	(1)	(2)	(3)	(4)
	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)
Log(Labor HHI)	-0.451***	-0.225***	-0.0403***	-0.0736***
	(0.0158)	(0.0108)	(0.00620)	(0.0156)
Log(Product HHI)	-0.512***	-0.441***	-0.431***	-0.187**
	(0.0877)	(0.0919)	(0.0852)	(0.0748)
Mean Age (in years)	-0.0177***	-0.0186***	-0.0183***	-0.00483***
	(0.00358)	(0.00388)	(0.00345)	(0.000387)
Share of Men	-0.374***	-0.399***	-0.380***	-0.0261**
	(0.0766)	(0.0677)	(0.0456)	(0.0132)
Mean Log(Value Added per Employee)	0.0798***	0.0374**	0.0108	0.0167
	(0.0189)	(0.0185)	(0.0135)	(0.0174)
Mean Log(Nb. Employees)	0.289***	0.274***	0.235***	0.165***
	(0.0218)	(0.0225)	(0.0232)	(0.0165)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.498	0.513	0.602	0.964
Adjusted R^2	0.498	0.513	0.599	0.959
N. Clusters	306	305	304	304
F	1709.0	595.2	631.4	203.6
Observations	11750180	11750179	11743710	11330634

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear regression using the Log(Nb. of Employees) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of employees across time for a given combination of industry, occupation, and commuting zone. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new exits and the mean age (in years) of the new quitters. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across employees) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across employees) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers net employment by approximately $-0.0403 \times 0.1 \times 100 = -0.403\%$.

Table 2.C.24: Net Employment (IV) : Weighted by Mean Market Employment

	(1)	(2)	(3)	(4)
	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)	Log(Nb. Employees)
Log(Labor HHI)	-0.109*	-0.195***	-0.222***	-0.313***
	(0.0562)	(0.0458)	(0.0475)	(0.0856)
Log(Product HHI)	-0.996***	-0.927***	-0.958***	-0.668***
	(0.0350)	(0.0629)	(0.0703)	(0.0781)
Mean Age (in years)	-0.0152***	-0.0150***	-0.0145***	-0.00420***
	(0.00298)	(0.00316)	(0.00250)	(0.000297)
Share of Men	-0.321***	-0.345***	-0.306***	-0.0279**
	(0.0684)	(0.0756)	(0.0594)	(0.0123)
Mean Log(Value Added per Employee)	0.161***	0.0983***	0.0843***	0.0145
	(0.0366)	(0.0270)	(0.0195)	(0.0117)
Mean Log(Nb. Employees)	0.361***	0.344***	0.316***	0.172***
	(0.0226)	(0.0273)	(0.0329)	(0.0162)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	306	305	304	304
F	802.9	438.6	570.8	113.9
Observations	11750180	11750179	11743710	11330634

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from an instrumental variable regression using the Log(Nb. of Employees) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of employees across time for a given combination of industry, occupation, and commuting zone. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the employees and the mean age (in years) of employees. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across employees) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across employees) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers net employment by approximately $-0.222 \times 0.1 \times 100 = -2.22\%$.

2.C.6.4 Poisson Regression

Table 2.C.25: Net Employment (OLS) : Poisson Regression

	(1)	(2)	(3)	(4)
	Net Employment	Net Employment	Net Employment	Net Employment
Log(Labor HHI)	-0.618***	-0.731***	-0.784***	-0.395***
	(0.0228)	(0.0235)	(0.0473)	(0.0497)
Log(Product HHI)	-0.360***	-0.377***	-0.384***	-0.269***
	(0.0356)	(0.0358)	(0.0374)	(0.00926)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	306	306	306	306
F	16039.7	7273.1	14531.4	1088.2
Observations	37389080	37389080	37389080	37389080

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression using Net Employment as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers net employment by approximately $-0.784 \times 0.1 \times 100 = -7.84\%$.

2.C.7 Panel

2.C.7.1 Baseline

Table 2.C.26: New Hires (OLS) : Baseline with Panel Data

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.122***	-0.0120*	0.0440***	0.00443
	(0.00583)	(0.00708)	(0.00344)	(0.00488)
Log(Product HHI)	-0.0851***	-0.0648***	-0.0737***	-0.00800***
	(0.0109)	(0.00926)	(0.00920)	(0.00250)
Mean Age (in years)	-0.000753**	-0.000612**	-0.000861***	-0.000579***
	(0.000302)	(0.000259)	(0.000309)	(0.000113)
Share of Men	-0.0176**	-0.0237***	-0.0191***	-0.00544**
	(0.00783)	(0.00726)	(0.00639)	(0.00246)
Mean Log(Value Added per Employee)	-0.0535***	-0.0562***	-0.0628***	-0.00625***
	(0.00639)	(0.00545)	(0.00653)	(0.00158)
Mean Log(Nb. Employees)	0.0516***	0.0509***	0.0525***	0.00221
	(0.00393)	(0.00414)	(0.00365)	(0.00179)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.163	0.182	0.276	0.728
Adjusted R^2	0.162	0.182	0.234	0.674
N. Clusters	304	304	304	304
F	367.3	188.0	269.7	20.01
Observations	746669	746669	733037	624303

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.0440 \times 0.1 \times 100 = -0.44\%$.

Table 2.C.27: New Hires (IV) : Baseline with Panel Data

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	0.0664	-0.0252*	-0.0575***	-0.338***
	(0.0461)	(0.0132)	(0.0140)	(0.120)
Log(Product HHI)	-0.158***	-0.106***	-0.164***	-0.0926
	(0.0263)	(0.0182)	(0.0196)	(0.0619)
Mean Age (in years)	-0.000849***	-0.000588**	-0.000845***	-0.000591***
	(0.000260)	(0.000269)	(0.000325)	(0.000126)
Share of Men	-0.00362	-0.0227***	-0.0170**	-0.00549**
	(0.0118)	(0.00784)	(0.00740)	(0.00246)
Mean Log(Value Added per Employee)	-0.0334**	-0.0541***	-0.0574***	-0.00439**
	(0.0155)	(0.00565)	(0.00683)	(0.00189)
Mean Log(Nb. Employees)	0.0503***	0.0528***	0.0575***	0.00420**
	(0.00301)	(0.00452)	(0.00443)	(0.00187)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	183.3	131.4	231.5	30.94
Observations	746669	746669	733037	624303

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-0.0575 \times 0.1 \times 100 = -0.575\%$.

2.C.7.2 Weighted by New Hires

Table 2.C.28: New Hires (OLS) : Weighted by New Hires

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.440***	0.0782***	0.137***	0.0484**
	(0.0439)	(0.0277)	(0.0139)	(0.0204)
Log(Product HHI)	-0.364***	-0.208***	-0.176***	-0.0385***
	(0.0414)	(0.0504)	(0.0490)	(0.00343)
Mean Age (in years)	0.00708**	0.00554***	0.00106	-0.000975***
	(0.00280)	(0.00190)	(0.00158)	(0.000193)
Share of Men	-0.0121	-0.0966**	-0.0999***	-0.0235***
	(0.0639)	(0.0394)	(0.0101)	(0.00472)
Mean Log(Value Added per Employee)	-0.0933***	-0.109***	-0.127***	-0.00153
	(0.0145)	(0.0138)	(0.0109)	(0.00515)
Mean Log(Nb. Employees)	0.105***	0.0883***	0.0865***	0.00917
	(0.0161)	(0.0147)	(0.0119)	(0.00625)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.541	0.619	0.723	0.937
Adjusted R^2	0.541	0.618	0.707	0.925
N. Clusters	304	304	304	304
F	211.1	36.71	119.1	46.37
Observations	746669	746669	733037	624303

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Panel (2011-2015). Each observation is weighted by the number of new hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.137 \times 0.1 \times 100 = 1.37\%$.

Table 2.C.29: Net Hires (IV) : Weighted by Market New Hires

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.116	-0.358	-0.408	-0.668*
	(0.315)	(0.276)	(0.254)	(0.382)
Log(Product HHI)	-0.513***	-0.357***	-0.557***	-0.534
	(0.101)	(0.0385)	(0.0737)	(0.370)
Mean Age (in years)	0.00749**	0.00628***	0.00150	-0.000844***
	(0.00314)	(0.00220)	(0.00189)	(0.000315)
Share of Men	0.0779**	-0.108***	-0.0924***	-0.0315***
	(0.0313)	(0.0354)	(0.0135)	(0.00597)
Mean Log(Value Added per Employee)	-0.0647***	-0.114***	-0.103***	0.0238*
	(0.0226)	(0.0163)	(0.0101)	(0.0136)
Mean Log(Nb. Employees)	0.107***	0.103***	0.121***	0.0180*
	(0.0170)	(0.0185)	(0.0174)	(0.00950)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	100.7	60.23	191.0	89.50
Observations	746669	746669	733037	624303

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Panel (2011-2015). Each observation is weighted by the number of new hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-0.408 \times 0.1 \times 100 = -4.08\%$.

2.C.7.3 Weighted by Mean New Hires

Table 2.C.30: New Hires (OLS) : Weighted by Mean Market Hires with Panel Data

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.469***	0.0329	0.0867***	0.0418**
	(0.0436)	(0.0263)	(0.0117)	(0.0199)
Log(Product HHI)	-0.339***	-0.190***	-0.153***	-0.0303***
	(0.0469)	(0.0526)	(0.0471)	(0.00439)
Mean Age (in years)	0.00467**	0.00397**	0.000300	-0.00130***
	(0.00223)	(0.00157)	(0.00121)	(0.000316)
Share of Men	0.0190	-0.0519	-0.0707***	-0.0281***
	(0.0535)	(0.0336)	(0.00848)	(0.00720)
Mean Log(Value Added per Employee)	-0.106***	-0.118***	-0.125***	0.00265
	(0.0106)	(0.0131)	(0.00984)	(0.00467)
Mean Log(Nb. Employees)	0.0935***	0.0809***	0.0766***	0.00657
	(0.0180)	(0.0170)	(0.0134)	(0.00562)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.524	0.594	0.700	0.921
Adjusted R^2	0.524	0.594	0.682	0.906
N. Clusters	304	304	304	304
F	200.5	29.04	73.37	12.58
Observations	746669	746669	733037	624303

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Panel (2011-2015). Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.0867 \times 0.1 \times 100 = 0.867\%$.

Table 2.C.31: New Hires (IV) : Weighted by Mean Market Hires with Panel Data

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	0.255**	-0.349	-0.350**	-0.834*
	(0.115)	(0.212)	(0.170)	(0.451)
Log(Product HHI)	-0.577***	-0.315***	-0.493***	-0.331
	(0.0475)	(0.0439)	(0.0573)	(0.331)
Mean Age (in years)	0.00551	0.00426**	0.000353	-0.00137***
	(0.00356)	(0.00168)	(0.00135)	(0.000363)
Share of Men	0.192	-0.0541	-0.0608***	-0.0343***
	(0.120)	(0.0335)	(0.00976)	(0.00916)
Mean Log(Value Added per Employee)	-0.0441**	-0.122***	-0.107***	0.0170**
	(0.0203)	(0.0153)	(0.0111)	(0.00830)
Mean Log(Nb. Employees)	0.0920***	0.0916***	0.104***	0.0147*
	(0.0166)	(0.0193)	(0.0180)	(0.00762)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	87.20	58.74	144.8	21.38
Observations	746669	746669	733037	624303

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Panel (2011-2015). Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $-0.350 \times 0.1 \times 100 = -3.50\%$.

2.C.7.4 Poisson Regression

Table 2.C.32: New Hires (OLS) : Poisson Regression with Panel Data

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.495***	-0.351***	-0.409***	-0.509***
	(0.0188)	(0.0244)	(0.0174)	(0.104)
Log(Product HHI)	-0.284***	-0.253***	-0.261***	-0.302***
	(0.0171)	(0.0169)	(0.0159)	(0.103)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	353.4	118.7	1586.0	28.39
Observations	4509900	4509900	4509900	4509900

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Panel (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new exits by approximately $-0.409 \times 0.1 \times 100 = -4.09\%$.

2.D Constant Sample

2.D.1 Hourly Wage

Table 2.D.1: Hourly Wage (OLS) : Constant Sample

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0130***	-0.00767***	-0.00301**	-0.00215***	-0.00224*	-0.00203***
	(0.00352)	(0.00170)	(0.00118)	(0.000681)	(0.00126)	(0.000758)
Log(Product HHI)	-0.00697***	-0.00262	-0.00248	0.00185	-0.00320	-0.00185***
	(0.00258)	(0.00231)	(0.00260)	(0.00152)	(0.00218)	(0.000652)
Age (in years)	0.00289***	0.00286***	0.00276***	0.00233***		
	(0.000391)	(0.000386)	(0.000440)	(0.000435)		
Gender	0.0271***	0.0261***	0.0257***	0.0221***		
	(0.00184)	(0.00184)	(0.00238)	(0.00287)		
Log(Value Added per Employee)	0.0188***	0.0179***	0.0163***	-0.000226	0.0105***	0.000777
	(0.00125)	(0.00133)	(0.00136)	(0.000705)	(0.000526)	(0.000753)
Log(Nb. Employees)	0.00688***	0.00666***	0.00669***	0.000427	0.00661***	0.00161**
	(0.000160)	(0.000146)	(0.000136)	(0.00118)	(0.000125)	(0.000723)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.511	0.516	0.549	0.653	0.742	0.793
Adjusted R^2	0.511	0.515	0.540	0.619	0.634	0.677
N. Clusters	304	304	304	304	304	304
F	1077.5	844.5	1055.8	103.1	914.6	11.32
Observations	1582456	1582456	1582456	1582456	1582456	1582456

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. The sample is kept constant across specifications, using column (6) as providing a baseline. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00203 \times 0.1 \times 100 = -0.0203\%$.

Table 2.D.2: Hourly Wage (IV) : Constant Sample

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0897***	-0.0812***	-0.0702***	-0.0316***	-0.0575***	-0.0194**
	(0.0144)	(0.0126)	(0.0115)	(0.00982)	(0.00972)	(0.00972)
Log(Product HHI)	0.0790***	0.0770**	0.0807**	-0.00638	0.0696*	-0.0273
	(0.0290)	(0.0299)	(0.0320)	(0.0283)	(0.0363)	(0.0285)
Age (in years)	0.00285***	0.00284***	0.00278***	0.00233***		
	(0.000324)	(0.000357)	(0.000420)	(0.000438)		
Gender	0.0235***	0.0251***	0.0240***	0.0220***		
	(0.00210)	(0.00258)	(0.00276)	(0.00291)		
Log(Value Added per Employee)	0.0162***	0.0154***	0.0143***	-0.000376	0.00899***	0.000698
	(0.00195)	(0.00277)	(0.00282)	(0.000754)	(0.00169)	(0.000722)
Log(Nb. Employees)	0.00517***	0.00472***	0.00370***	-0.000133	0.00467***	0.00110
	(0.000451)	(0.000431)	(0.000430)	(0.00135)	(0.000227)	(0.000790)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	965.4	1132.0	1500.5	124.8	508.3	9.036
Observations	1582456	1582456	1582456	1582456	1582456	1582456

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrument variables using the Log(Gross Hourly Wage) as a dependent variable. The sample is kept constant across specifications, using column (6) as providing a baseline. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0575 \times 0.1 \times 100 = -0.575\%$.

2.D.2 New Hires

2.D.2.1 Baseline

Table 2.D.3: New Hires (OLS) : Baseline with Constant Sample

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.137***	-0.0727***	-0.0198***	-0.0937***
	(0.00249)	(0.00652)	(0.00430)	(0.00675)
Log(Product HHI)	-0.113***	-0.102***	-0.109***	-0.0429**
	(0.0129)	(0.0124)	(0.0130)	(0.0208)
Mean Age (in years)	-0.00382***	-0.00382***	-0.00370***	-0.000557***
	(0.0000759)	(0.0000742)	(0.0000761)	(0.0000586)
Share of Men	-0.0492***	-0.0536***	-0.0478***	0.00218
	(0.00725)	(0.00743)	(0.00683)	(0.00146)
Mean Log(Value Added per Employee)	0.0722***	0.0746***	0.0788***	-0.00502***
	(0.00420)	(0.00417)	(0.00398)	(0.00130)
Mean Log(Nb. Employees)	0.0507***	0.0499***	0.0490***	0.0132***
	(0.00234)	(0.00237)	(0.00220)	(0.00154)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.161	0.164	0.224	0.742
Adjusted R^2	0.161	0.164	0.202	0.683
N. Clusters	305	305	305	305
F	2652.0	1112.1	932.7	225.3
Observations	2620737	2620737	2620737	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. The sample is kept constant across specifications, using column (4) as providing a baseline. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.0198 \times 0.1 \times 100 = -0.198\%$.

Table 2.D.4: New Hires (IV) : Baseline with Constant Sample

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.546**	-0.388***	-0.408***	-0.585***
	(0.268)	(0.103)	(0.105)	(0.187)
Log(Product HHI)	-0.323***	-0.324***	-0.353***	-3.096***
	(0.0295)	(0.0194)	(0.0158)	(0.798)
Mean Age (in years)	-0.00316***	-0.00358***	-0.00350***	-0.000240
	(0.000606)	(0.000179)	(0.000160)	(0.000261)
Share of Men	-0.0744***	-0.0430***	-0.0401***	0.00523
	(0.00724)	(0.0121)	(0.00886)	(0.00360)
Mean Log(Value Added per Employee)	0.0945***	0.0604***	0.0634***	-0.00117
	(0.0294)	(0.00575)	(0.00548)	(0.00384)
Mean Log(Nb. Employees)	0.0739***	0.0698***	0.0726***	0.0290***
	(0.00870)	(0.00384)	(0.00443)	(0.00481)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	305	305	305	305
F	299.3	794.4	678.1	47.89
Observations	2620737	2620737	2620737	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. The sample is kept constant across specifications, using column (4) as providing a baseline. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.408 \times 0.1 \times 100 = -4.08\%$.

2.D.2.2 Weighted by New Hires

Table 2.D.5: New Hires (OLS) : Weighted by New Hires with Constant Sample

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.499***	0.124***	0.267***	0.0777***
	(0.0319)	(0.0395)	(0.0424)	(0.0253)
Log(Product HHI)	-0.446***	-0.281***	-0.254***	-0.137**
	(0.0573)	(0.0594)	(0.0524)	(0.0575)
Mean Age (in years)	-0.00628***	-0.00948***	-0.0113***	0.000343
	(0.00127)	(0.00140)	(0.00134)	(0.000273)
Share of Men	-0.117*	-0.240***	-0.209***	-0.00613
	(0.0700)	(0.0436)	(0.0168)	(0.00532)
Mean Log(Value Added per Employee)	0.103***	0.120***	0.127***	0.00457
	(0.0197)	(0.0150)	(0.0135)	(0.00710)
Mean Log(Nb. Employees)	0.186***	0.139***	0.120***	0.0448***
	(0.0189)	(0.0173)	(0.0146)	(0.00868)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.521	0.584	0.678	0.945
Adjusted R^2	0.521	0.584	0.669	0.932
N. Clusters	305	305	305	305
F	376.8	131.6	144.2	44.38
Observations	2620737	2620737	2620737	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. The sample is kept constant across specifications, using column (4) as providing a baseline. Each observation is weighted by the number of new hires. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.267 \times 0.1 \times 100 = 2.67\%$.

Table 2.D.6: New Hires (IV) : Weighted by New Hires with Constant Sample

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-1.977*	-1.995*	-2.359*	-1.165
	(1.160)	(1.118)	(1.378)	(1.719)
Log(Product HHI)	-0.702*	-1.098***	-1.330***	-11.48
	(0.395)	(0.114)	(0.0677)	(10.38)
Mean Age (in years)	-0.00260	-0.000765	-0.00342	0.00295
	(0.00186)	(0.00548)	(0.00522)	(0.00975)
Share of Men	-0.820	-0.266***	-0.191***	-0.0299
	(0.549)	(0.0387)	(0.0357)	(0.0341)
Mean Log(Value Added per Employee)	0.136*	0.0822**	0.0879**	0.0738***
	(0.0694)	(0.0401)	(0.0434)	(0.0262)
Mean Log(Nb. Employees)	0.274***	0.309***	0.329***	0.151*
	(0.0285)	(0.0584)	(0.0704)	(0.0812)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	305	305	305	305
F	100.3	97.86	191.2	44.11
Observations	2620737	2620737	2620737	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. The sample is kept constant across specifications, using column (4) as providing a baseline. Each observation is weighted by the number of new hires. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-2.359 \times 0.1 \times 100 = 23.59\%$.

2.D.2.3 Weighted by Mean New Hires

Table 2.D.7: New Hires (OLS) : Weighted by New Hires with Constant Sample

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.579***	-0.0337	0.0102	-0.0546**
	(0.0209)	(0.0292)	(0.0267)	(0.0229)
Log(Product HHI)	-0.414***	-0.280***	-0.265***	-0.228***
	(0.0670)	(0.0690)	(0.0606)	(0.0867)
Mean Age (in years)	-0.0114***	-0.0124***	-0.0130***	-0.00314***
	(0.000850)	(0.00124)	(0.00145)	(0.000340)
Share of Men	-0.0544	-0.142***	-0.147***	-0.00818
	(0.0694)	(0.0505)	(0.0229)	(0.00798)
Mean Log(Value Added per Employee)	0.118***	0.131***	0.137***	0.00899
	(0.0156)	(0.0135)	(0.0135)	(0.00557)
Mean Log(Nb. Employees)	0.152***	0.123***	0.108***	0.0349***
	(0.0180)	(0.0173)	(0.0150)	(0.00560)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.526	0.570	0.654	0.923
Adjusted R^2	0.526	0.570	0.644	0.905
N. Clusters	305	305	305	305
F	620.0	151.9	132.8	50.10
Observations	2620737	2620737	2620737	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. The sample is kept constant across specifications, using column (4) as providing a baseline. Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.0102 \times 0.1 \times 100 = 0.102\%$.

Table 2.D.8: New Hires (IV) : Weighted by New Hires with Constant Sample

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-4.364	-1.710**	-1.562**	-1.521*
	(5.223)	(0.816)	(0.654)	(0.844)
Log(Product HHI)	-0.340	-1.178***	-1.307***	-5.305**
	(1.242)	(0.0835)	(0.0582)	(2.573)
Mean Age (in years)	-0.00503	-0.00897***	-0.0107***	0.000514
	(0.00446)	(0.00205)	(0.00104)	(0.00264)
Share of Men	-1.233	-0.0976	-0.0871***	-0.0487***
	(1.714)	(0.0653)	(0.0187)	(0.0106)
Mean Log(Value Added per Employee)	0.262	0.0912**	0.0992***	0.0484***
	(0.251)	(0.0379)	(0.0380)	(0.0166)
Mean Log(Nb. Employees)	0.234***	0.252***	0.248***	0.0941***
	(0.0225)	(0.0341)	(0.0338)	(0.0214)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	305	305	305	305
F	52.80	194.1	317.2	133.1
Observations	2620737	2620737	2620737	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. The sample is kept constant across specifications, using column (4) as providing a baseline. Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $-1.562 \times 0.1 \times 100 = 15.62\%$.

2.D.2.4 Poisson Regression

Table 2.D.9: New Hires (OLS): Poisson Regression with Constant Sample

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.544***	-0.377***	-0.431***	-0.486***
	(0.0221)	(0.0216)	(0.0223)	(0.0788)
Log(Product HHI)	-0.370***	-0.339***	-0.354***	-0.446***
	(0.0313)	(0.0278)	(0.0284)	(0.0213)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	310	310	310	310
F	998.7	160.6	237.5	356.8
Observations	22016820	22016820	22016820	22016820

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. The sample is kept constant across specifications, using column (4) as providing a baseline. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new exits by approximately $-0.409 \times 0.1 \times 100 = -4.09\%$.

2.E Only Labor Market Concentration

2.E.1 Hourly Wage

Table 2.E.1: Hourly Wage (OLS) : Only Labor Market Concentration

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0156***	-0.00797***	-0.00311***	-0.00227***	-0.00247**	-0.00227***
	(0.00240)	(0.00120)	(0.000792)	(0.000457)	(0.00115)	(0.000851)
Age (in years)	0.00339***	0.00336***	0.00327***	0.00274***		
	(0.000413)	(0.000409)	(0.000470)	(0.000484)		
Gender	0.0304***	0.0295***	0.0287***	0.0242***		
	(0.00109)	(0.00105)	(0.00169)	(0.00248)		
Log(Value Added per Employee)	0.0230***	0.0223***	0.0202***	-0.000878	0.0112***	0.000699
	(0.00170)	(0.00177)	(0.00191)	(0.000722)	(0.000616)	(0.000748)
Log(Nb. Employees)	0.00799***	0.00788***	0.00778***	0.0000801	0.00714***	0.00164**
	(0.000238)	(0.000264)	(0.000247)	(0.00140)	(0.000120)	(0.000713)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.523	0.527	0.556	0.664	0.741	0.793
Adjusted R^2	0.523	0.527	0.548	0.629	0.633	0.677
N. Clusters	304	304	304	304	304	304
F	741.1	757.2	923.2	232.0	1404.6	10.81
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00247 \times 0.1 \times 100 = -0.0247\%$.

Table 2.E.2: Hourly Wage (IV) : Only Labor Market Concentration

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0757***	-0.0555***	-0.0403***	-0.0389***	-0.0258***	-0.0304***
	(0.0155)	(0.0137)	(0.0116)	(0.00850)	(0.00835)	(0.00989)
Age (in years)	0.00334***	0.00338***	0.00327***	0.00274***		
	(0.000358)	(0.000403)	(0.000470)	(0.000486)		
Gender	0.0227***	0.0290***	0.0284***	0.0241***		
	(0.00200)	(0.00137)	(0.00183)	(0.00254)		
Log(Value Added per Employee)	0.0199***	0.0216***	0.0202***	-0.00107	0.0112***	0.000549
	(0.00219)	(0.00221)	(0.00202)	(0.000742)	(0.000695)	(0.000741)
Log(Nb. Employees)	0.00909***	0.00830***	0.00797***	-0.000477	0.00723***	0.00122
	(0.000301)	(0.000177)	(0.000192)	(0.00158)	(0.000125)	(0.000758)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	710.0	757.1	980.1	257.2	1345.8	7.926
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variables using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0258 \times 0.1 \times 100 = -0.258\%$.

2.E.2 New Hires

2.E.2.1 Baseline

Table 2.E.3: New Hires (OLS) : Baseline with Only Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.158***	-0.0612***	-0.00373	-0.0945***
	(0.0132)	(0.00681)	(0.00396)	(0.00658)
Mean Age (in years)	-0.00348***	-0.00345***	-0.00338***	-0.000558***
	(0.0000841)	(0.0000880)	(0.0000804)	(0.0000584)
Share of Men	-0.0541***	-0.0569***	-0.0552***	0.00215
	(0.00491)	(0.00539)	(0.00464)	(0.00144)
Mean Log(Value Added per Employee)	0.0821***	0.0833***	0.0855***	-0.00506***
	(0.00381)	(0.00372)	(0.00348)	(0.00127)
Mean Log(Nb. Employees)	0.0413***	0.0413***	0.0398***	0.0130***
	(0.000879)	(0.000843)	(0.000802)	(0.00147)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.157	0.164	0.215	0.741
Adjusted R^2	0.157	0.164	0.194	0.683
N. Clusters	308	307	305	305
F	1863.7	1764.6	1318.9	179.7
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.00373 \times 0.1 \times 100 = -0.0373\%$.

Table 2.E.4: New Hires (IV) : Baseline with Only Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.330**	-0.290***	-0.300***	-0.655***
	(0.128)	(0.0695)	(0.0672)	(0.198)
Mean Age (in years)	-0.00311***	-0.00333***	-0.00331***	-0.000264
	(0.000329)	(0.0000747)	(0.0000722)	(0.000160)
Share of Men	-0.0614***	-0.0544***	-0.0561***	0.00302*
	(0.00276)	(0.00718)	(0.00466)	(0.00171)
Mean Log(Value Added per Employee)	0.0929***	0.0852***	0.0882***	-0.00463***
	(0.0104)	(0.00388)	(0.00367)	(0.00171)
Mean Log(Nb. Employees)	0.0412***	0.0412***	0.0410***	0.0162***
	(0.000883)	(0.000853)	(0.000813)	(0.00278)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	1111.1	1435.4	1130.2	100.7
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.3 \times 0.1 \times 100 = -3\%$.

2.E.2.2 Weighted by New Hires

Table 2.E.5: New Hires (OLS) : Weighted by New Hires with Only Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.723***	0.0852*	0.266***	0.0725**
	(0.0925)	(0.0508)	(0.0449)	(0.0296)
Mean Age (in years)	-0.00629***	-0.00947***	-0.0110***	0.000385
	(0.00117)	(0.00145)	(0.00150)	(0.000277)
Share of Men	-0.140*	-0.249***	-0.236***	-0.00639
	(0.0826)	(0.0436)	(0.0116)	(0.00507)
Mean Log(Value Added per Employee)	0.134***	0.143***	0.146***	0.00390
	(0.0139)	(0.0124)	(0.0124)	(0.00644)
Mean Log(Nb. Employees)	0.153***	0.117***	0.0941***	0.0447***
	(0.00736)	(0.00567)	(0.00320)	(0.00878)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.494	0.581	0.678	0.944
Adjusted R^2	0.494	0.581	0.669	0.932
N. Clusters	308	307	305	305
F	804.9	295.1	210.3	51.57
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.266 \times 0.1 \times 100 = 2.66\%$.

Table 2.E.6: New Hires (IV) : Weighted by New Hires with Only Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-1.892**	-1.907**	-2.227*	-2.345**
	(0.919)	(0.938)	(1.152)	(1.123)
Mean Age (in years)	-0.00127	-0.00231	-0.00484**	0.00913
	(0.00289)	(0.00341)	(0.00229)	(0.00559)
Share of Men	-0.559*	-0.297***	-0.320***	-0.0715**
	(0.337)	(0.0624)	(0.0322)	(0.0298)
Mean Log(Value Added per Employee)	0.153***	0.137***	0.147***	0.0246
	(0.0210)	(0.0161)	(0.0143)	(0.0152)
Mean Log(Nb. Employees)	0.189***	0.188***	0.158***	0.179**
	(0.0353)	(0.0384)	(0.0347)	(0.0872)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	-0.293	-0.674	-0.679	-3.830
Adjusted R^2	-0.293	-0.674	-0.679	-3.830
N. Clusters	308	307	305	305
F	92.95	135.3	187.7	6.387
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-2.227 \times 0.1 \times 100 = -22.27\%$.

2.E.2.3 Weighted by Mean New Hires

Table 2.E.7: New Hires (OLS) : Weighted by Mean New Hires with Only Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.764***	-0.0562	0.0106	-0.0721**
	(0.0835)	(0.0421)	(0.0371)	(0.0348)
Mean Age (in years)	-0.0104***	-0.0115***	-0.0121***	-0.00314***
	(0.000798)	(0.00130)	(0.00156)	(0.000359)
Share of Men	-0.0609	-0.145***	-0.168***	-0.00782
	(0.0770)	(0.0488)	(0.0148)	(0.00811)
Mean Log(Value Added per Employee)	0.142***	0.147***	0.151***	0.00748
	(0.0133)	(0.0128)	(0.0135)	(0.00499)
Mean Log(Nb. Employees)	0.116***	0.0973***	0.0796***	0.0349***
	(0.00561)	(0.00545)	(0.00356)	(0.00623)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.498	0.561	0.650	0.922
Adjusted R^2	0.498	0.561	0.641	0.904
N. Clusters	308	307	305	305
F	1019.1	273.5	186.5	52.63
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of hires across time. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.0106 \times 0.1 \times 100 = 0.106\%$.

Table 2.E.8: New Hires (IV) : Weighted by Mean New Hires with Only Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-3.680	-1.508**	-1.353**	-2.185**
	(10.97)	(0.723)	(0.536)	(0.874)
Mean Age (in years)	-0.00145	-0.00878***	-0.0105***	0.000932
	(0.0315)	(0.00112)	(0.000722)	(0.00226)
Share of Men	-0.844	-0.141*	-0.190***	-0.0449***
	(3.075)	(0.0736)	(0.0109)	(0.0112)
Mean Log(Value Added per Employee)	0.224	0.149***	0.156***	0.0143*
	(0.308)	(0.0140)	(0.0129)	(0.00814)
Mean Log(Nb. Employees)	0.166	0.125***	0.0948***	0.102***
	(0.201)	(0.0181)	(0.00878)	(0.0391)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	44.30	205.1	154.8	58.44
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of hires across time. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-1.353 \times 0.1 \times 100 = -13.53\%$.

2.E.2.4 Poisson Regression

Table 2.E.9: New Hires (OLS) : Poisson Regression with Only Labor Market Concentration

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.685***	-0.391***	-0.460***	-0.532***
	(0.0620)	(0.0269)	(0.0349)	(0.0681)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	310	310	310	310
F	122.0	211.5	174.0	60.86
Observations	22016820	22016820	22016820	22016820

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.460 \times 0.1 \times 100 = -4.6\%$.

2.F Global Product Market Concentration

2.F.1 Hourly Wages

Table 2.F.1: Hourly Wage (OLS) : Global Product Market HHI

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0156***	-0.00792***	-0.00300***	-0.00240***	-0.00205**	-0.00220***
	(0.00239)	(0.00116)	(0.000734)	(0.000535)	(0.000967)	(0.000825)
Log(Global Product HHI)	-0.000562	-0.00102	-0.00248**	0.00379*	-0.00898***	-0.00175*
	(0.00127)	(0.00127)	(0.00101)	(0.00203)	(0.000474)	(0.00103)
Age (in years)	0.00339***	0.00336***	0.00327***	0.00274***		
	(0.000413)	(0.000409)	(0.000469)	(0.000484)		
Gender	0.0304***	0.0295***	0.0287***	0.0242***		
	(0.00109)	(0.00105)	(0.00169)	(0.00248)		
Log(Value Added per Employee)	0.0230***	0.0224***	0.0204***	-0.000880	0.0116***	0.000698
	(0.00179)	(0.00187)	(0.00197)	(0.000726)	(0.000633)	(0.000747)
Log(Nb. Employees)	0.00795***	0.00781***	0.00760***	0.000138	0.00649***	0.00162**
	(0.000187)	(0.000200)	(0.000197)	(0.00137)	(0.000122)	(0.000713)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.523	0.527	0.556	0.664	0.742	0.793
Adjusted R^2	0.523	0.527	0.548	0.629	0.633	0.677
N. Clusters	304	304	304	304	304	304
F	647.0	695.3	798.2	204.9	1259.6	10.40
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the year by industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00205 \times 0.1 \times 100 = -0.0205\%$.

Table 2.F.2: Hourly Wages (IV) : Global Product Market HHI

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0757***	-0.0555***	-0.0404***	-0.0413***	-0.0252***	-0.0315***
	(0.0154)	(0.0136)	(0.0117)	(0.00724)	(0.00758)	(0.0116)
Log(Global Product HHI)	-0.000400	0.000110	-0.00201**	0.00941**	-0.00860***	0.00381
	(0.00145)	(0.00145)	(0.000973)	(0.00410)	(0.000470)	(0.00496)
Age (in years)	0.00334***	0.00338***	0.00327***	0.00274***		
	(0.000358)	(0.000404)	(0.000469)	(0.000486)		
Gender	0.0227***	0.0290***	0.0284***	0.0241***		
	(0.00199)	(0.00136)	(0.00183)	(0.00254)		
Log(Value Added per Employee)	0.0199***	0.0216***	0.0204***	-0.00109	0.0116***	0.000544
	(0.00229)	(0.00233)	(0.00208)	(0.000764)	(0.000698)	(0.000740)
Log(Nb. Employees)	0.00906***	0.00831***	0.00783***	-0.000366	0.00660***	0.00125*
	(0.000299)	(0.000183)	(0.000159)	(0.00156)	(0.000130)	(0.000751)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	607.1	658.7	836.2	270.9	1146.6	6.177
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variables using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the year by industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0252 \times 0.1 \times 100 = -0.252\%$.

2.F.2 New Hires

2.F.2.1 Baseline

Table 2.F.3: New Hires (OLS) : Baseline Global Product HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.170***	-0.0683***	-0.00803**	-0.0945***
	(0.0151)	(0.00727)	(0.00390)	(0.00658)
Log(Global Product HHI)	-0.0934***	-0.0960***	-0.107***	-0.0205***
	(0.00529)	(0.00610)	(0.00664)	(0.00773)
Mean Age (in years)	-0.00330***	-0.00326***	-0.00318***	-0.000558***
	(0.0000771)	(0.0000799)	(0.0000729)	(0.0000584)
Share of Men	-0.0523***	-0.0550***	-0.0530***	0.00217
	(0.00462)	(0.00514)	(0.00426)	(0.00145)
Mean Log(Value Added per Employee)	0.0509***	0.0511***	0.0508***	-0.00502***
	(0.00322)	(0.00299)	(0.00287)	(0.00128)
Mean Log(Nb. Employees)	0.0423***	0.0424***	0.0407***	0.0130***
	(0.000911)	(0.000952)	(0.000842)	(0.00146)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.171	0.178	0.232	0.741
Adjusted R^2	0.171	0.178	0.212	0.683
N. Clusters	308	307	305	305
F	1511.9	1417.3	973.0	198.1
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman Index for the product market, measured at the year by industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-0.00803 \times 0.1 \times 100 = -0.0803\%$.

Table 2.F.4: New Hires (IV) : Baseline with Global Product HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.344**	-0.303***	-0.316***	-0.655***
	(0.133)	(0.0729)	(0.0712)	(0.197)
Log(Global Product HHI)	-0.112***	-0.100***	-0.109***	-0.0116***
	(0.0228)	(0.00756)	(0.00691)	(0.00213)
Mean Age (in years)	-0.00289***	-0.00312***	-0.00311***	-0.000264
	(0.000373)	(0.0000789)	(0.0000701)	(0.000160)
Share of Men	-0.0592***	-0.0524***	-0.0539***	0.00303*
	(0.00275)	(0.00699)	(0.00427)	(0.00171)
Mean Log(Value Added per Employee)	0.0556***	0.0517***	0.0530***	-0.00460***
	(0.00483)	(0.00298)	(0.00299)	(0.00171)
Mean Log(Nb. Employees)	0.0423***	0.0423***	0.0420***	0.0162***
	(0.000993)	(0.000871)	(0.000971)	(0.00278)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	918.0	1144.3	853.9	91.04
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman Index for the product market, measured at the year by industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-0.316 \times 0.1 \times 100 = -3.16\%$.

2.F.2.2 Weighted by New Hires

Table 2.F.5: New Hires (OLS) : Weighted by New Hires with Global Product HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.727***	0.0867*	0.268***	0.0731**
	(0.0948)	(0.0511)	(0.0458)	(0.0291)
Log(Global Product HHI)	-0.258***	-0.275***	-0.313***	-0.0834***
	(0.0278)	(0.0307)	(0.0270)	(0.0170)
Mean Age (in years)	-0.00553***	-0.00872***	-0.0102***	0.000372
	(0.00128)	(0.00128)	(0.00118)	(0.000276)
Share of Men	-0.121	-0.230***	-0.211***	-0.00604
	(0.0786)	(0.0391)	(0.0105)	(0.00521)
Mean Log(Value Added per Employee)	0.0306**	0.0331**	0.0238*	0.00428
	(0.0144)	(0.0145)	(0.0143)	(0.00668)
Mean Log(Nb. Employees)	0.156***	0.120***	0.0968***	0.0443***
	(0.00741)	(0.00573)	(0.00316)	(0.00875)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.506	0.595	0.694	0.944
Adjusted R^2	0.506	0.595	0.686	0.932
N. Clusters	308	307	305	305
F	885.0	262.4	197.9	51.43
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman Index for the product market, measured at the year by industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.268 \times 0.1 \times 100 = 2.68\%$.

Table 2.F.6: New Hires (IV) : Weighted by New Hires with Global Product HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-1.968**	-1.977**	-2.294*	-2.346**
	(0.985)	(0.991)	(1.216)	(1.131)
Log(Global Product HHI)	-0.289***	-0.270***	-0.308***	0.00649
	(0.0675)	(0.0308)	(0.0275)	(0.0969)
Mean Age (in years)	-0.000126	-0.00131	-0.00394	0.00913
	(0.00345)	(0.00388)	(0.00280)	(0.00563)
Share of Men	-0.564	-0.281***	-0.297***	-0.0715**
	(0.369)	(0.0582)	(0.0380)	(0.0303)
Mean Log(Value Added per Employee)	0.0383**	0.0284	0.0266	0.0246*
	(0.0168)	(0.0196)	(0.0162)	(0.0149)
Mean Log(Nb. Employees)	0.195***	0.193***	0.162***	0.179**
	(0.0381)	(0.0405)	(0.0364)	(0.0881)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	79.75	113.2	212.2	31.13
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman Index for the product market, measured at the year by industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-2.294 \times 0.1 \times 100 = 22.94\%$.

2.F.2.3 Weighted by Mean New Hires

Table 2.F.7: New Hires (OLS) : Weighted by Mean New Hires with Global Product HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.772***	-0.0614	0.00555	-0.0704**
	(0.0858)	(0.0415)	(0.0346)	(0.0335)
Log(Global Product HHI)	-0.294***	-0.302***	-0.329***	-0.110***
	(0.0325)	(0.0366)	(0.0349)	(0.0403)
Mean Age (in years)	-0.00974***	-0.0108***	-0.0115***	-0.00314***
	(0.000763)	(0.00118)	(0.00138)	(0.000359)
Share of Men	-0.0448	-0.129***	-0.147***	-0.00747
	(0.0739)	(0.0454)	(0.0126)	(0.00836)
Mean Log(Value Added per Employee)	0.0250	0.0270	0.0241	0.00818
	(0.0168)	(0.0181)	(0.0190)	(0.00528)
Mean Log(Nb. Employees)	0.119***	0.100***	0.0820***	0.0345***
	(0.00566)	(0.00555)	(0.00356)	(0.00612)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.514	0.578	0.669	0.922
Adjusted R^2	0.514	0.578	0.660	0.904
N. Clusters	308	307	305	305
F	1128.9	226.4	148.7	44.48
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of hires across time. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman Index for the product market, measured at the year by industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.00555 \times 0.1 \times 100 = 0.0555\%$.

Table 2.F.8: New Hires (IV) : Weighted by Mean New Hires with Global Product HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-4.216	-1.579**	-1.424**	-2.190**
	(13.25)	(0.758)	(0.568)	(0.886)
Log(Global Product HHI)	-0.433	-0.313***	-0.334***	0.0463
	(0.609)	(0.0396)	(0.0338)	(0.117)
Mean Age (in years)	0.00117	-0.00798***	-0.00976***	0.000940
	(0.0396)	(0.00131)	(0.000548)	(0.00228)
Share of Men	-0.960	-0.124*	-0.170***	-0.0451***
	(3.694)	(0.0712)	(0.00982)	(0.0117)
Mean Log(Value Added per Employee)	0.0659	0.0246	0.0277	0.0141*
	(0.134)	(0.0200)	(0.0170)	(0.00762)
Mean Log(Nb. Employees)	0.179	0.130***	0.0979***	0.103**
	(0.247)	(0.0190)	(0.00920)	(0.0398)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	47.74	165.3	149.3	89.17
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of hires across time. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman Index for the product market, measured at the year by industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-1.424 \times 0.1 \times 100 = 14.24\%$.

2.F.2.4 Poisson Regression

Table 2.F.9: New Hires (OLS) : Poisson Regression with Global Product HHI

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.764***	-0.408***	-0.462***	-0.482***
	(0.0637)	(0.0220)	(0.0278)	(0.0627)
Log(Global Product HHI)	-0.457***	-0.471***	-0.493***	-0.842***
	(0.00857)	(0.00940)	(0.00859)	(0.112)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	310	310	310	310
F	1551.4	1456.2	1686.1	29.72
Observations	22016820	22016820	22016820	22016820

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Global Product HHI) is the logarithm of the Herfindalh-Hirschman Index for the product market, measured at the year by industry level. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.462 \times 0.1 \times 100 = -4.62\%$.

2.G Tradeable Sector

2.G.1 Hourly Wage

Table 2.G.1: Hourly Wages (OLS) : Tradeable Sector

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0213***	-0.00884***	-0.00432**	-0.00232**	-0.00250**	-0.00295**
	(0.00230)	(0.00163)	(0.00174)	(0.00114)	(0.00122)	(0.00116)
Log(Global Product HHI)	0.00606***	0.00561***	0.00391**	0.00883***	-0.00299*	0.00392**
	(0.00203)	(0.00208)	(0.00167)	(0.00185)	(0.00171)	(0.00188)
Age (in years)	0.00497***	0.00497***	0.00480***	0.00436***		
	(0.000587)	(0.000601)	(0.000730)	(0.000867)		
Gender	0.0446***	0.0439***	0.0420***	0.0384***		
	(0.00137)	(0.00133)	(0.00194)	(0.00397)		
Log(Value Added per Employee)	0.0248***	0.0232***	0.0215***	-0.00246***	0.0104***	0.000424
	(0.00177)	(0.00144)	(0.000972)	(0.000884)	(0.00111)	(0.00272)
Log(Nb. Employees)	0.00888***	0.00834***	0.00757***	-0.0105***	0.00661***	-0.00131
	(0.000673)	(0.000543)	(0.000478)	(0.00129)	(0.000883)	(0.00356)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.609	0.613	0.642	0.729	0.804	0.838
Adjusted R^2	0.609	0.613	0.628	0.692	0.708	0.741
N_clust	304	304	303	303	301	301
F	531.6	422.7	620.8	84.18	65.06	2.721
Observations	518225	518225	507071	464662	272000	242700

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00250 \times 0.1 \times 100 = -0.0250\%$.

Table 2.G.2: Hourly Wage (IV) : Tradeable Sector

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.00823	-0.0108	0.00390	-0.0114	-0.0108	-0.0148
	(0.0567)	(0.0420)	(0.0425)	(0.0350)	(0.0232)	(0.0224)
Log(Global Product HHI)	0.00539	0.00565**	0.00375	0.0102**	-0.00265	0.00695
	(0.00488)	(0.00274)	(0.00245)	(0.00418)	(0.00257)	(0.00556)
Age (in years)	0.00491***	0.00497***	0.00479***	0.00437***		
	(0.000372)	(0.000566)	(0.000704)	(0.000861)		
Gender	0.0449***	0.0439***	0.0421***	0.0384***		
	(0.00222)	(0.00126)	(0.00206)	(0.00396)		
Log(Value Added per Employee)	0.0259***	0.0232***	0.0215***	-0.00254***	0.0104***	0.000244
	(0.00361)	(0.00117)	(0.000987)	(0.000964)	(0.00108)	(0.00290)
Log(Nb. Employees)	0.00859***	0.00838***	0.00747***	-0.0107***	0.00669***	-0.00166
	(0.00197)	(0.00120)	(0.000944)	(0.00133)	(0.000747)	(0.00391)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	303	303	301	301
F	476.6	433.9	590.4	109.4	61.01	1.369
Observations	518225	518225	507071	464662	272000	242700

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from instrumental variables using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0108 \times 0.1 \times 100 = -0.108\%$.

2.G.2 New Hires

2.G.2.1 Baseline

Table 2.G.3: New Hires (OLS) : Baseline with Tradeable Sector

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.137***	-0.0310*	0.0320***	-0.00669
	(0.0290)	(0.0160)	(0.00791)	(0.00996)
Log(Global Product HHI)	-0.0531***	-0.0565***	-0.0689***	-0.0241***
	(0.0119)	(0.0134)	(0.0161)	(0.00809)
Mean Age (in years)	-0.00155***	-0.00144***	-0.00112***	-0.000223**
	(0.000166)	(0.000166)	(0.000156)	(0.0000931)
Share of Men	-0.0542***	-0.0520***	-0.0488***	0.00580***
	(0.00411)	(0.00582)	(0.00471)	(0.00182)
Mean Log(Value Added per Employee)	-0.0111**	-0.0186***	-0.0220***	-0.0104***
	(0.00527)	(0.00348)	(0.00263)	(0.00196)
Mean Log(Nb. Employees)	0.0506***	0.0503***	0.0454***	0.0269***
	(0.00119)	(0.00102)	(0.00141)	(0.00134)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.139	0.158	0.248	0.645
Adjusted R^2	0.138	0.157	0.212	0.545
N. Clusters	304	304	304	304
F	861.1	693.2	343.3	183.4
Observations	1284755	1284755	1270979	988646

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.0320 \times 0.1 = -0.320\%$.

Table 2.G.4: New Hires (IV) : Baseline with Tradeable Sector

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.341	-0.289**	-0.308***	-0.794**
	(0.227)	(0.115)	(0.112)	(0.353)
Log(Global Product HHI)	-0.0598**	-0.0593***	-0.0698***	-0.0191***
	(0.0237)	(0.0151)	(0.0164)	(0.00419)
Mean Age (in years)	-0.000749	-0.00124***	-0.00102***	0.000126
	(0.000921)	(0.0000985)	(0.000120)	(0.000291)
Share of Men	-0.0471***	-0.0466***	-0.0483***	0.0111***
	(0.0170)	(0.00885)	(0.00522)	(0.00315)
Mean Log(Value Added per Employee)	-0.0316	-0.0195***	-0.0218***	-0.0113***
	(0.0242)	(0.00290)	(0.00278)	(0.00256)
Mean Log(Nb. Employees)	0.0484***	0.0500***	0.0468***	0.0311***
	(0.00309)	(0.00122)	(0.00110)	(0.00433)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	552.5	640.1	511.3	129.0
Observations	1284755	1284755	1270979	988646

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.308 \times 0.1 \times 100 = -3.08\%$.

2.G.2.2 Weighted by New Hires

Table 2.G.5: New Hires (OLS) : Weighted by New Hires in Tradeable Sector

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.409***	0.316***	0.404***	0.197***
	(0.0880)	(0.0866)	(0.0663)	(0.0492)
Mean Log(Product HHI)	-0.172**	-0.171**	-0.213***	-0.0913***
	(0.0678)	(0.0685)	(0.0704)	(0.0270)
Mean Age (in years)	-0.000596	-0.00147	-0.000726	0.00147***
	(0.00191)	(0.00157)	(0.000747)	(0.000431)
Share of Men	-0.277***	-0.261***	-0.160***	0.0206***
	(0.0424)	(0.0290)	(0.0202)	(0.00746)
Mean Log(Value Added per Employee)	0.00232	-0.0657**	-0.0933***	-0.0267***
	(0.0288)	(0.0268)	(0.0230)	(0.00681)
Mean Log(Nb. Employees)	0.206***	0.176***	0.138***	0.0788***
	(0.0103)	(0.00864)	(0.00893)	(0.0119)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.508	0.616	0.745	0.929
Adjusted R^2	0.507	0.616	0.733	0.909
N. Clusters	304	304	304	304
F	471.5	339.1	275.4	138.3
Observations	1284755	1284755	1270979	988646

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Each observation is weighted by the number of hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.404 \times 0.1 \times 100 = 4.04\%$.

Table 2.G.6: New Hires (IV) : Weighted by New Hires in Tradeable Sector

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-3.231	-4.280	-5.265	-8.692
	(3.773)	(6.114)	(8.204)	(18.70)
Log(Global Product HHI)	0.0214	-0.125***	-0.201***	0.332
	(0.196)	(0.0263)	(0.0554)	(1.140)
Mean Age (in years)	0.0279	0.0148	0.0134	0.0248
	(0.0423)	(0.0230)	(0.0218)	(0.0534)
Share of Men	-0.198	-0.198	-0.194***	0.153
	(0.191)	(0.160)	(0.0365)	(0.303)
Mean Log(Value Added per Employee)	-0.468	-0.137	-0.0983***	-0.0883
	(0.763)	(0.127)	(0.0292)	(0.122)
Mean Log(Nb. Employees)	0.257***	0.305*	0.271	0.485
	(0.0881)	(0.180)	(0.216)	(0.978)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	38.30	38.21	79.28	1.820
Observations	1284755	1284755	1270979	988646

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Each observation is weighted by the number of hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-5.265 \times 0.1 \times 100 = -52.65\%$.

2.G.2.3 Weighted by Mean New Hires

Table 2.G.7: New Hires (OLS) : Weighted by Mean New Hires in Tradeable Sector

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.453***	0.132*	0.179***	0.134**
	(0.0815)	(0.0781)	(0.0557)	(0.0564)
Log(Global Product HHI)	-0.175**	-0.177**	-0.212***	-0.126***
	(0.0701)	(0.0733)	(0.0783)	(0.0460)
Mean Age (in years)	-0.00446***	-0.00397***	-0.00299***	-0.00280***
	(0.00147)	(0.00146)	(0.00105)	(0.000550)
Share of Men	-0.162***	-0.154***	-0.110***	0.00658
	(0.0366)	(0.0290)	(0.0205)	(0.0147)
Mean Log(Value Added per Employee)	-0.0143	-0.0707***	-0.0864***	-0.0209
	(0.0253)	(0.0262)	(0.0256)	(0.0151)
Mean Log(Nb. Employees)	0.144***	0.134***	0.109***	0.0742***
	(0.00732)	(0.00895)	(0.00958)	(0.0126)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.492	0.571	0.689	0.876
Adjusted R^2	0.492	0.571	0.675	0.842
N. Clusters	304	304	304	304
F	473.5	247.9	174.8	51.13
Observations	1284755	1284755	1270979	988646

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $0.179 \times 0.1 \times 100 = 1.79\%$.

Table 2.G.8: New Hires (IV) : Weighted by Mean New Hires in Tradeable Industry

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	3.062	-3.176	-2.641	-5.048
	(3.623)	(3.483)	(2.215)	(5.003)
Log(Global Product HHI)I	-0.356***	-0.161***	-0.203***	0.242
	(0.110)	(0.0495)	(0.0618)	(0.527)
Mean Age (in years)	-0.0288	0.00111	-0.000301	0.00257
	(0.0181)	(0.00550)	(0.00184)	(0.00675)
Share of Men	-0.280***	-0.0515	-0.0891*	0.0905
	(0.0833)	(0.174)	(0.0481)	(0.118)
Mean Log(Value Added per Employee)	0.548	-0.104	-0.0800***	-0.00906
	(0.429)	(0.0684)	(0.0200)	(0.0281)
Mean Log(Nb. Employees)	0.117***	0.172***	0.138***	0.208
	(0.0174)	(0.0492)	(0.0376)	(0.178)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	53.61	88.01	96.09	12.26
Observations	1284755	1284755	1270979	988646

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Each observation is weighted by the mean number of new hires across time for a given combination of industry, occupation, and commuting zone. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration increases new hires by approximately $2.641 \times 0.1 \times 100 = 26.41\%$.

2.G.2.4 Poisson Regression

Table 2.G.9: New Hires (OLS) : Poisson Regression in Tradeable Sector

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.733***	-0.355***	-0.349***	-0.502***
	(0.0958)	(0.0572)	(0.0483)	(0.131)
Log(Global Product HHI)	-0.323***	-0.349***	-0.396***	-0.831***
	(0.0596)	(0.0663)	(0.0649)	(0.105)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	304	304	304	304
F	43.22	19.32	27.60	105.2
Observations	10854560	10854560	10854560	10854560

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015) in an industry considered to be exposed to international trade. These industries were selected if at least 5% of the industry's revenue was obtained from exports during our initial year of data (2011). Missing values were replaced with zeros, to provide a balanced panel. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new exits by approximately $-0.349 \times 0.1 \times 100 = -3.49\%$.

2.H Local Time Varying Industry FE

2.H.1 Hourly Wage

Table 2.H.1: Hourly Wage (OLS) : Local Time Varying Industry FE

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.00865***	-0.00865***	-0.00274***	-0.00118**	-0.000297	-0.000166
	(0.00169)	(0.00169)	(0.000828)	(0.000543)	(0.000556)	(0.000677)
Age (in years)	0.00323***	0.00323***	0.00316***	0.00268***		
	(0.000455)	(0.000455)	(0.000499)	(0.000519)		
Gender	0.0280***	0.0280***	0.0271***	0.0239***		
	(0.00169)	(0.00169)	(0.00205)	(0.00273)		
Log(Value Added per Employee)	0.0194***	0.0194***	0.0185***	-0.000567	0.0111***	0.00123**
	(0.00182)	(0.00182)	(0.00178)	(0.00109)	(0.000574)	(0.000539)
Log(Nb. Employees)	0.00580***	0.00580***	0.00576***	-0.00106	0.00469***	0.000545
	(0.000150)	(0.000150)	(0.000129)	(0.00181)	(0.000143)	(0.000805)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.568	0.568	0.588	0.681	0.761	0.806
Adjusted R^2	0.551	0.551	0.564	0.636	0.643	0.681
N_clust	304	304	304	304	302	302
F	685.0	685.0	680.2	187.3	786.5	2.147
Observations	2185387	2185387	2172391	2003504	1685952	1516526

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.000297 \times 0.1 \times 100 = -0.00297\%$.

Table 2.H.2: Hourly Wage (IV) : Local Time Varying Industry FE

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0445***	-0.0445***	-0.0288**	-0.0225*	0.00941	0.0127
	(0.0116)	(0.0116)	(0.0125)	(0.0115)	(0.0107)	(0.0120)
Age (in years)	0.00323***	0.00323***	0.00315***	0.00268***		
	(0.000458)	(0.000458)	(0.000500)	(0.000521)		
Gender	0.0280***	0.0280***	0.0270***	0.0238***		
	(0.00178)	(0.00178)	(0.00212)	(0.00277)		
Log(Value Added per Employee)	0.0194***	0.0194***	0.0185***	-0.000478	0.0111***	0.00116**
	(0.00186)	(0.00186)	(0.00178)	(0.00112)	(0.000554)	(0.000538)
Log(Nb. Employees)	0.00611***	0.00611***	0.00584***	-0.000983	0.00467***	0.000486
	(0.000157)	(0.000157)	(0.000133)	(0.00185)	(0.000156)	(0.000832)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	302	302
F	690.7	690.7	675.8	225.6	710.4	2.038
Observations	2185387	2185387	2172391	2003504	1685952	1516526

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration raises wages by approximately $0.00941 \times 0.1 \times 100 = 0.0941\%$.

2.H.2 New Hires

2.H.2.1 Baseline

Table 2.H.3: New Hires (OLS) : Baseline with Local Time Varying Industry FE

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.0704***	-0.0704***	0.000687	-0.0519***
	(0.00842)	(0.00842)	(0.00308)	(0.00857)
Mean Age (in years)	-0.00344***	-0.00344***	-0.00322***	-0.00116***
	(0.0000693)	(0.0000693)	(0.0000750)	(0.0000545)
Share of Men	-0.0346***	-0.0346***	-0.0280***	0.000171
	(0.00647)	(0.00647)	(0.00535)	(0.00233)
Mean Log(Value Added per Employee)	0.0113***	0.0113***	0.00878***	-0.00494***
	(0.00216)	(0.00216)	(0.00145)	(0.000943)
Mean Log(Nb. Employees)	0.0283***	0.0283***	0.0298***	0.00597***
	(0.00367)	(0.00367)	(0.00329)	(0.00187)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.350	0.350	0.404	0.812
Adjusted R^2	0.207	0.207	0.248	0.707
N. Clusters	306	306	304	304
F	762.6	762.6	640.2	215.8
Observations	2802104	2802104	2787540	2217501

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $0.000687 \times 0.1 \times 100 = 0.00687\%$.

Table 2.H.4: New Hires (IV) : Baseline with Local Time Varying Industry FE

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.414***	-0.414***	-0.418***	-0.668***
	(0.105)	(0.105)	(0.0993)	(0.210)
Mean Age (in years)	-0.00327***	-0.00327***	-0.00315***	-0.000992***
	(0.0000966)	(0.0000966)	(0.0000857)	(0.0000887)
Share of Men	-0.0295***	-0.0295***	-0.0291***	-0.000394
	(0.00952)	(0.00952)	(0.00520)	(0.00233)
Mean Log(Value Added per Employee)	0.00978***	0.00978***	0.00895***	-0.00484***
	(0.00159)	(0.00159)	(0.00145)	(0.00111)
Mean Log(Nb. Employees)	0.0297***	0.0297***	0.0312***	0.00881***
	(0.00369)	(0.00369)	(0.00356)	(0.00278)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	306	306	304	304
F	780.1	780.1	668.4	100.8
Observations	2802104	2802104	2787540	2217501

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.418 \times 0.1 \times 100 = -4.18\%$.

2.H.2.2 Weighted by New Hires

Table 2.H.5: New Hires (OLS) : Weighted by New Hires with Local Time Varying Industry FE

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	0.0701***	0.0701***	0.209***	0.101***
	(0.0205)	(0.0205)	(0.0144)	(0.00865)
Mean Age (in years)	-0.0106***	-0.0106***	-0.00874***	-0.00262***
	(0.000877)	(0.000877)	(0.000605)	(0.000128)
Share of Men	-0.165***	-0.165***	-0.0964***	-0.0101
	(0.0309)	(0.0309)	(0.0287)	(0.0107)
Mean Log(Value Added per Employee)	0.0419***	0.0419***	0.0363***	0.000267
	(0.00918)	(0.00918)	(0.00654)	(0.00565)
Mean Log(Nb. Employees)	0.0537***	0.0537***	0.0665***	0.0151
	(0.0161)	(0.0161)	(0.0163)	(0.00961)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.815	0.815	0.850	0.970
Adjusted R^2	0.774	0.774	0.811	0.953
N. Clusters	306	306	304	304
F	272.3	272.3	333.7	259.6
Observations	2802104	2802104	2787540	2217501

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each observation is weighted by the number of new hires. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $0.209 \times 0.1 \times 100 = 2.09\%$.

Table 2.H.6: New Hires (IV) : Weighted by New Hires with Local Time Varying Industry FE

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-1.822	-1.822	-1.914	-2.323
	(1.206)	(1.206)	(1.284)	(1.652)
Mean Age (in years)	-0.00982***	-0.00982***	-0.00862***	-0.00166*
	(0.000552)	(0.000552)	(0.000467)	(0.000943)
Share of Men	-0.170***	-0.170***	-0.143***	-0.0565***
	(0.0496)	(0.0496)	(0.0101)	(0.0214)
Mean Log(Value Added per Employee)	0.0504***	0.0504***	0.0511***	0.0263
	(0.0125)	(0.0125)	(0.0160)	(0.0229)
Mean Log(Nb. Employees)	0.103**	0.103**	0.0978***	0.0763
	(0.0500)	(0.0500)	(0.0372)	(0.0556)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	306	306	304	304
F	216.2	216.2	214.3	21.24
Observations	2802104	2802104	2787540	2217501

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each observation is weighted by the number of new hires. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $-1.914 \times 0.1 \times 100 = -19.14\%$.

2.H.2.3 Weighted by Mean New Hires

Table 2.H.7: New Hires (OLS) : Weighted by Mean New Hires with Local Time Varying Industry FE

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.0144	-0.0144	0.0906***	0.0622***
	(0.0159)	(0.0159)	(0.0103)	(0.0107)
Mean Age (in years)	-0.00946***	-0.00946***	-0.00813***	-0.00345***
	(0.000990)	(0.000990)	(0.000790)	(0.000258)
Share of Men	-0.116***	-0.116***	-0.0767***	-0.0114
	(0.0293)	(0.0293)	(0.0247)	(0.0105)
Mean Log(Value Added per Employee)	0.0329***	0.0329***	0.0297***	0.00341
	(0.00641)	(0.00641)	(0.00445)	(0.00631)
Mean Log(Nb. Employees)	0.0519***	0.0519***	0.0598***	0.0170**
	(0.0157)	(0.0157)	(0.0159)	(0.00776)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.789	0.789	0.826	0.958
Adjusted R^2	0.742	0.742	0.780	0.934
N. Clusters	306	306	304	304
F	206.6	206.6	246.2	100.0
Observations	2802104	2802104	2787540	2217501

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each observation is weighted by the mean number of new hires. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $0.0906 \times 0.1 \times 100 = 0.906\%$.

Table 2.H.8: New Hires (IV) : Weighted by Mean New Hirew with Local Time Varying Industry FE

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-1.717*	-1.717*	-1.461**	-2.040*
	(0.992)	(0.992)	(0.703)	(1.099)
Mean Age (in years)	-0.00856***	-0.00856***	-0.00795***	-0.00307***
	(0.000322)	(0.000322)	(0.000669)	(0.000212)
Share of Men	-0.0977*	-0.0977*	-0.0954***	-0.0395***
	(0.0567)	(0.0567)	(0.0193)	(0.00741)
Mean Log(Value Added per Employee)	0.0354***	0.0354***	0.0358***	0.0170
	(0.00885)	(0.00885)	(0.00642)	(0.0132)
Mean Log(Nb. Employees)	0.0872**	0.0872**	0.0761***	0.0542*
	(0.0376)	(0.0376)	(0.0238)	(0.0286)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	306	306	304	304
F	262.6	262.6	299.2	49.61
Observations	2802104	2802104	2787540	2217501

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Each observation is weighted by the mean number of new hires. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-1.461 \times 0.1 \times 100 = -14.61\%$.

2.H.2.4 Poisson Regression

Table 2.H.9: New Hires (OLS) : Poisson Regression with Local Time Varying Industry FE

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.429***	-0.429***	-0.485***	-0.550***
	(0.0159)	(0.0159)	(0.0108)	(0.112)
Quarter x Year x CZ x Industry FE	Yes	Yes	Yes	Yes
Quarter $\times$ Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	307	307	304	304
F	727.3	727.3	2024.1	24.05
Observations	17734872	17734872	17734749	17732685

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Each specification of this regression also controls for time varying local industry effect through the inclusion of dummy variables at the time by industry by commuting zone level. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.485 \times 0.1 \times 100 = -4.85\%$.

2.I Employment Weighted Product Market Concentration

2.I.1 Hourly Wage

Table 2.I.1: Hourly Wage (OLS) : Baseline with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0152***	-0.00812***	-0.00304***	-0.00259***	-0.00201*	-0.00226***
	(0.00230)	(0.00122)	(0.000793)	(0.000602)	(0.00108)	(0.000818)
Log(Emp. Adj. Local Product HHI)	-0.000679	0.000980**	0.0000835	0.00305**	-0.00319***	0.000178
	(0.000564)	(0.000432)	(0.000388)	(0.00120)	(0.000601)	(0.000640)
Age (in years)	0.00315***	0.00313***	0.00303***	0.00256***		
	(0.000259)	(0.000255)	(0.000308)	(0.000350)		
Gender	0.0303***	0.0294***	0.0286***	0.0242***		
	(0.00123)	(0.00119)	(0.00184)	(0.00264)		
Log(Value Added per Employee)	0.0227***	0.0219***	0.0199***	-0.000779	0.0114***	0.000773
	(0.00157)	(0.00163)	(0.00175)	(0.000697)	(0.000568)	(0.000757)
Log(Nb. Employees)	0.00802***	0.00781***	0.00776***	0.0000928	0.00730***	0.00163**
	(0.000229)	(0.000233)	(0.000223)	(0.00139)	(0.000120)	(0.000726)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.521	0.525	0.554	0.663	0.742	0.793
Adjusted R^2	0.521	0.525	0.546	0.628	0.633	0.677
N. Clusters	304	304	304	304	304	304
F	640.1	644.3	800.3	193.6	1155.2	9.581
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00201 \times 0.1 \times 100 = -0.0201\%$.

Table 2.I.2: Hourly Wage (IV) : Baseline with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0585***	-0.0573***	-0.0413***	-0.0397***	-0.0318**	-0.0262***
	(0.0154)	(0.0110)	(0.00971)	(0.00823)	(0.0129)	(0.00707)
Log(Emp. Adj. Local Product HHI)	0.0506***	0.0498***	0.0483***	0.0145	0.0395***	-0.0165
	(0.00725)	(0.00519)	(0.00346)	(0.0135)	(0.00446)	(0.0168)
Age (in years)	0.00311***	0.00310***	0.00300***	0.00256***		
	(0.000266)	(0.000253)	(0.000310)	(0.000352)		
Gender	0.0294***	0.0292***	0.0284***	0.0241***		
	(0.00270)	(0.00132)	(0.00165)	(0.00269)		
Log(Value Added per Employee)	0.0174***	0.0169***	0.0159***	-0.00107	0.00867***	0.000694
	(0.00328)	(0.00275)	(0.00237)	(0.000705)	(0.00121)	(0.000745)
Log(Nb. Employees)	0.00539***	0.00537***	0.00496***	-0.000543	0.00533***	0.00132*
	(0.000281)	(0.000266)	(0.000303)	(0.00152)	(0.000207)	(0.000793)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	590.4	722.2	981.0	255.2	881.3	8.170
Observations	2225026	2225026	2212203	2044008	1734623	1563889

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the number of reported full-time equivalent number of workers in the firm over the year. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0318 \times 0.1 \times 100 = -0.318\%$.

2.I.2 New Hires

2.I.2.1 Baseline

Table 2.I.3: New Hires (OLS) : Baseline with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.112***	-0.0589***	-0.00575	-0.0936***
	(0.00690)	(0.00687)	(0.00359)	(0.00673)
Log(Emp. Adj. Local Product HHI)	-0.0683***	-0.0631***	-0.0661***	-0.0291**
	(0.00465)	(0.00426)	(0.00465)	(0.0120)
Mean Age (in years)	-0.00330***	-0.00329***	-0.00321***	-0.000557***
	(0.0000690)	(0.0000707)	(0.0000714)	(0.0000586)
Share of Men	-0.0517***	-0.0535***	-0.0516***	0.00213
	(0.00582)	(0.00633)	(0.00563)	(0.00145)
Mean Log(Value Added per Employee)	0.0745***	0.0756***	0.0781***	-0.00484***
	(0.00422)	(0.00409)	(0.00383)	(0.00134)
Mean Log(Nb. Employees)	0.0545***	0.0534***	0.0527***	0.0134***
	(0.00198)	(0.00197)	(0.00183)	(0.00160)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.175	0.178	0.230	0.742
Adjusted R^2	0.175	0.178	0.209	0.683
N. Clusters	308	307	305	305
F	1561.6	1367.2	1282.6	221.7
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-0.00575 \times 0.1 \times 100 = -0.0575\%$.

Table 2.I.4: New Hires (IV) : Baseline with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.413**	-0.307***	-0.320***	-0.626***
	(0.189)	(0.0791)	(0.0793)	(0.183)
Log(Emp. Adj. Local Product HHI)	-0.118***	-0.125***	-0.136***	-0.804**
	(0.00843)	(0.0107)	(0.0112)	(0.329)
Mean Age (in years)	-0.00245***	-0.00300***	-0.00296***	-0.000251
	(0.000641)	(0.000126)	(0.000111)	(0.000203)
Share of Men	-0.0639***	-0.0474***	-0.0487***	0.00238
	(0.00389)	(0.00940)	(0.00675)	(0.00214)
Mean Log(Value Added per Employee)	0.0899***	0.0703***	0.0732***	0.00147
	(0.0161)	(0.00440)	(0.00430)	(0.00437)
Mean Log(Nb. Employees)	0.0639***	0.0652***	0.0676***	0.0262***
	(0.00332)	(0.00357)	(0.00427)	(0.00693)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	-0.086	0.008	0.005	-0.615
Adjusted R^2	-0.086	0.008	0.005	-0.615
N. Clusters	308	307	305	305
F	493.8	921.9	699.4	69.47
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-0.320 \times 0.1 \times 100 = -3.2\%$.

2.I.2.2 Weighted by New Hires

Table 2.I.5: New Hires (OLS) : Weighted by New Hires using Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.447***	0.147***	0.273***	0.0791***
	(0.0510)	(0.0418)	(0.0423)	(0.0248)
Log(Emp. Adj. Local Product HHI)	-0.290***	-0.204***	-0.188***	-0.104***
	(0.0196)	(0.0188)	(0.0174)	(0.0372)
Mean Age (in years)	-0.00569***	-0.00854***	-0.0102***	0.000330
	(0.00118)	(0.00113)	(0.00114)	(0.000275)
Share of Men	-0.120	-0.219***	-0.203***	-0.00626
	(0.0743)	(0.0464)	(0.0163)	(0.00520)
Mean Log(Value Added per Employee)	0.133***	0.141***	0.146***	0.00530
	(0.0208)	(0.0164)	(0.0149)	(0.00701)
Mean Log(Nb. Employees)	0.221***	0.170***	0.149***	0.0465***
	(0.0190)	(0.0163)	(0.0138)	(0.00936)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.548	0.606	0.695	0.945
Adjusted R^2	0.548	0.605	0.687	0.932
N. Clusters	308	307	305	305
F	580.7	234.6	363.3	41.19
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. Each observation is weighted by the number of new hires. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-0.00575 \times 0.1 \times 100 = -0.0575\%$.

Table 2.I.6: New Hires (IV) : Weighted by New Hires with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-1.769*	-1.745*	-2.048*	-1.824**
	(0.960)	(0.899)	(1.095)	(0.764)
Log(Emp. Adj. Local Product HHI)	-0.286**	-0.433***	-0.526***	-2.269***
	(0.144)	(0.0224)	(0.0409)	(0.576)
Mean Age (in years)	-0.0000489	-0.000450	-0.00314	0.00656
	(0.00241)	(0.00402)	(0.00314)	(0.00417)
Share of Men	-0.594*	-0.234***	-0.223***	-0.0584***
	(0.346)	(0.0637)	(0.0155)	(0.0170)
Mean Log(Value Added per Employee)	0.154***	0.133***	0.147***	0.0518***
	(0.0326)	(0.0248)	(0.0287)	(0.0197)
Mean Log(Nb. Employees)	0.261***	0.299***	0.308***	0.198***
	(0.0140)	(0.0528)	(0.0659)	(0.0703)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	103.4	117.7	254.7	9.452
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. Each observation is weighted by the number of new hires. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration decreases new hires by approximately $-2.048 \times 0.1 \times 100 = -20.48\%$.

2.I.2.3 Weighted by Mean New Hires

Table 2.I.7: New Hires (OLS) : Weighted by Mean New Hires with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.522***	-0.00579	0.0145	-0.0521**
	(0.0379)	(0.0305)	(0.0283)	(0.0224)
Log(Emp. Adj. Local Product HHI)	-0.264***	-0.195***	-0.186***	-0.166***
	(0.0258)	(0.0248)	(0.0226)	(0.0589)
Mean Age (in years)	-0.00977***	-0.0108***	-0.0115***	-0.00314***
	(0.000719)	(0.00103)	(0.00130)	(0.000346)
Share of Men	-0.0508	-0.123**	-0.143***	-0.00879
	(0.0733)	(0.0534)	(0.0215)	(0.00768)
Mean Log(Value Added per Employee)	0.137***	0.143***	0.149***	0.0107*
	(0.0174)	(0.0149)	(0.0141)	(0.00577)
Mean Log(Nb. Employees)	0.179***	0.147***	0.131***	0.0376***
	(0.0184)	(0.0169)	(0.0148)	(0.00688)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.543	0.584	0.667	0.923
Adjusted R^2	0.543	0.583	0.659	0.905
N. Clusters	308	307	305	305
F	705.9	244.2	239.7	45.83
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. Each observation is weighted by the mean number of new hires across time. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $0.0145 \times 0.1 \times 100 = 0.145\%$.

Table 2.I.8: New Hires (IV) : Weighted by Mean New Hires with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-4.122	-1.581**	-1.396**	-1.851***
	(6.278)	(0.748)	(0.552)	(0.669)
Log(Emp. Adj. Local Product HHI)	-0.0556	-0.444***	-0.505***	-1.273***
	(0.756)	(0.0224)	(0.0409)	(0.336)
Mean Age (in years)	0.000207	-0.00680***	-0.00885***	0.000613
	(0.0128)	(0.00169)	(0.000488)	(0.00205)
Share of Men	-0.974	-0.0912	-0.124***	-0.0492***
	(1.699)	(0.0832)	(0.0244)	(0.0112)
Mean Log(Value Added per Employee)	0.237	0.139***	0.152***	0.0384***
	(0.176)	(0.0237)	(0.0253)	(0.0142)
Mean Log(Nb. Employees)	0.188***	0.241***	0.234***	0.117***
	(0.0719)	(0.0371)	(0.0381)	(0.0336)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	-4.131	-0.437	-0.279	-2.506
Adjusted R^2	-4.132	-0.437	-0.279	-2.506
N. Clusters	308	307	305	305
F	59.26	203.7	370.4	70.05
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. Each observation is weighted by the mean number of new hires across time. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There are two firm level control variables: Mean Log(Value Added per Worker) and Mean Log(Number of Employees). The former is the Mean (across new hires) of the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The latter is the mean (across new hires) of the number of reported full-time equivalent number of workers in the firms if the respective employees over the year. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-1.396 \times 0.1 \times 100 = -13.96\%$.

2.I.2.4 Poisson Regression

Table 2.I.9: New Hires (OLS) : Poisson Regression with Employment Weighted Product Market HHI

	(1)	(2)	(3)	(4)
	Nb. Hires	Nb. Hires	Nb. Hires	Nb. Hires
Log(Labor HHI)	-0.528***	-0.379***	-0.432***	-0.462***
	(0.0359)	(0.0229)	(0.0234)	(0.0775)
Log(Emp. Adj. Local Product HHI)	-0.211***	-0.197***	-0.207***	-0.367***
	(0.0129)	(0.0104)	(0.0108)	(0.0170)
Quarter × Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	310	310	310	310
F	147.1	179.3	187.4	371.5
Observations	22016820	22016820	22016820	22016820

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the output from a Poisson Regression Nb. Hires as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Missing values were replaced with zeros, to provide a balanced panel. Log(Emp. Adj. Local Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market measured at the commuting zone by industry per year, where firms' revenues where weighted by the number of employees in the commuting zone. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. The exponential specification was used because the data presents a relationship of this form. It lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.432 \times 0.1 \times 100 = -4.32\%$.

2.J No Firm Size

2.J.1 Hourly Wage

Table 2.J.1: Hourly Wage (OLS) : Baseline without Controlling for Firm Size

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0136***	-0.00683***	-0.00196	-0.00247***	-0.00162	-0.00204***
	(0.00333)	(0.00184)	(0.00147)	(0.000619)	(0.00168)	(0.000763)
Log(Product HHI)	-0.00188	0.00235	0.00358*	0.00173	0.000437	-0.00186***
	(0.00201)	(0.00180)	(0.00197)	(0.00168)	(0.00197)	(0.000650)
Age (in years)	0.00302***	0.00300***	0.00293***	0.00256***		
	(0.000266)	(0.000259)	(0.000311)	(0.000350)		
Gender	0.0293***	0.0284***	0.0275***	0.0242***		
	(0.00125)	(0.00121)	(0.00179)	(0.00264)		
Log(Value Added per Employee)	0.0175***	0.0171***	0.0153***	-0.000867	0.00650***	-0.000516
	(0.00190)	(0.00195)	(0.00212)	(0.000535)	(0.000989)	(0.000446)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
R^2	0.514	0.518	0.548	0.663	0.738	0.793
Adjusted R^2	0.514	0.518	0.540	0.628	0.629	0.677
N. Clusters	304	304	304	304	304	304
F	1337.0	924.1	700.9	207.6	121.7	13.92
Observations	2250464	2250464	2237656	2069051	1753604	1582456

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There is a single firm level control variables: Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00162 \times 0.1 \times 100 = -0.0162\%$.

Table 2.J.2: Hourly Wage (IV) : Baseline without Controlling for Firm Size

	(1)	(2)	(3)	(4)	(5)	(6)
	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)	Log(Gross Hourly Wage)
Log(Labor HHI)	-0.0973***	-0.0906***	-0.0711***	-0.0457***	-0.0616***	-0.0195**
	(0.0136)	(0.0126)	(0.0109)	(0.00976)	(0.0101)	(0.00971)
Log(Product HHI)	0.103***	0.101***	0.0960***	0.0255	0.0901**	-0.0275
	(0.0339)	(0.0359)	(0.0337)	(0.0260)	(0.0427)	(0.0285)
Age (in years)	0.00305***	0.00304***	0.00299***	0.00256***		
	(0.000222)	(0.000240)	(0.000293)	(0.000352)		
Gender	0.0263***	0.0271***	0.0258***	0.0241***		
	(0.00169)	(0.00225)	(0.00223)	(0.00272)		
Log(Value Added per Employee)	0.0164***	0.0156***	0.0149***	-0.000864	0.00644***	-0.000181
	(0.00270)	(0.00372)	(0.00366)	(0.000710)	(0.00221)	(0.000542)
Quarter x Year FE	Yes	Yes	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No	No	No
Commuting Zone FE	No	Yes	No	No	No	No
Commuting Zone x Occupation FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	No	Yes
Worker FE	No	No	No	No	Yes	Yes
N. Clusters	304	304	304	304	304	304
F	1135.8	1364.4	1604.9	225.1	232.1	11.34
Observations	2250464	2250464	2237656	2069051	1753604	1582456

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Gross Hourly Wage) as a dependent variable. Each observation is a new hire labor contract, as provided in the DADS Panel (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There is a single firm level control variables: Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The gender fixed-effect cannot be identified in specification (v) and (iv) by collinearity with individual fixed-effects. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (5): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0616 \times 0.1 \times 100 = -0.616\%$.

2.J.2 New Hires

2.J.2.1 Baseline

Table 2.J.3: New Hires (OLS) : Baseline without Controlling for Firm Size

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.129***	-0.0601***	0.000737	-0.0931***
	(0.00477)	(0.00718)	(0.00341)	(0.00677)
Log(Product HHI)	-0.0887***	-0.0765***	-0.0820***	-0.0417**
	(0.0120)	(0.0107)	(0.0112)	(0.0207)
Mean Age (in years)	-0.00416***	-0.00409***	-0.00398***	-0.000651***
	(0.0000699)	(0.0000731)	(0.0000671)	(0.0000568)
Share of Men	-0.0637***	-0.0664***	-0.0638***	0.00124
	(0.00566)	(0.00591)	(0.00540)	(0.00136)
Mean Log(Value Added per Employee)	0.0967***	0.0979***	0.0993***	-0.00700***
	(0.00368)	(0.00364)	(0.00331)	(0.00144)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.154	0.158	0.211	0.741
Adjusted R^2	0.154	0.158	0.191	0.683
N. Clusters	308	307	305	305
F	2139.3	1244.0	1042.2	89.36
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There is a single firm level control variables: Mean Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $0.000737 \times 0.1 \times 100 = 0.00737\%$.

Table 2.J.4: New Hires (IV) : Baseline without Controlling for Firm Size

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.445**	-0.335***	-0.348***	-0.592***
	(0.202)	(0.0842)	(0.0840)	(0.191)
Log(Product HHI)	-0.196***	-0.210***	-0.233***	-3.259***
	(0.0146)	(0.0183)	(0.0168)	(0.847)
Mean Age (in years)	-0.00344***	-0.00398***	-0.00397***	-0.000444*
	(0.000670)	(0.000108)	(0.0000871)	(0.000248)
Share of Men	-0.0768***	-0.0610***	-0.0623***	0.00324
	(0.00369)	(0.00891)	(0.00615)	(0.00335)
Log(Value Added per Employee)	0.114***	0.0940***	0.0965***	-0.00544
	(0.0176)	(0.00432)	(0.00433)	(0.00427)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	700.3	1220.8	1018.4	51.61
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There is a single firm level control variables: Mean Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-0.348 \times 0.1 \times 100 = -3.48\%$.

2.J.2.2 Weighted by New Hires

Table 2.J.5: New Hires (OLS) : Weighted by New Hires without Controlling for Firm Size

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.502***	0.170***	0.324***	0.0888***
	(0.0388)	(0.0403)	(0.0407)	(0.0243)
Log(Product HHI)	-0.387***	-0.222***	-0.191***	-0.137**
	(0.0541)	(0.0532)	(0.0451)	(0.0580)
Mean Age (in years)	-0.0107***	-0.0122***	-0.0132***	0.0000763
	(0.00113)	(0.00139)	(0.00144)	(0.000283)
Share of Men in Firm	-0.219***	-0.307***	-0.282***	-0.00954**
	(0.0661)	(0.0388)	(0.0138)	(0.00464)
Log(Value Added per Employee)	0.158***	0.164***	0.168***	-0.00687
	(0.0169)	(0.0127)	(0.0118)	(0.00641)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.505	0.579	0.677	0.944
Adjusted R^2	0.505	0.579	0.668	0.932
N. Clusters	308	307	305	305
F	261.9	90.76	164.5	33.93
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the number of new hires. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There is a single firm level control variables: Mean Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $0.324 \times 0.1 \times 100 = 3.24\%$.

Table 2.J.6: New Hires (IV) : Weighted by New Hires without Controlling for Firm Size

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-2.034*	-2.066*	-2.395*	-1.304
	(1.038)	(1.056)	(1.309)	(1.769)
Log(Product HHI)	-0.212	-0.517***	-0.745***	-11.41
	(0.338)	(0.0879)	(0.0374)	(9.744)
Mean Age (in years)	-0.00508*	-0.00596	-0.00855***	0.00264
	(0.00278)	(0.00371)	(0.00265)	(0.00949)
Share of Men	-0.757**	-0.403***	-0.396***	-0.0470
	(0.368)	(0.0412)	(0.0630)	(0.0433)
Mean Log(Value Added per Employee)	0.196***	0.168***	0.181***	0.0339
	(0.0386)	(0.0262)	(0.0326)	(0.0272)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	54.33	63.98	184.6	48.63
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Standard Errors are clustered at the commuting zone level. Each observation is weighted by the number of new hires. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There is a single firm level control variables: Mean Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-2.395 \times 0.1 \times 100 = -23.95\%$.

2.J.2.3 Weighted by Mean New Hires

Table 2.J.7: New Hires (OLS) : Weighted by Mean New Hires without Controlling for Firm Size

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-0.580***	-0.000623	0.0388	-0.0485**
	(0.0262)	(0.0299)	(0.0284)	(0.0228)
Log(Product HHI)	-0.352***	-0.219***	-0.202***	-0.228***
	(0.0634)	(0.0629)	(0.0539)	(0.0879)
Mean Age (in years)	-0.0136***	-0.0136***	-0.0139***	-0.00341***
	(0.000885)	(0.00134)	(0.00161)	(0.000381)
Share of Men	-0.120*	-0.189***	-0.202***	-0.0109
	(0.0655)	(0.0457)	(0.0163)	(0.00746)
Mean Log(Value Added per Employee)	0.163***	0.167***	0.172***	0.000766
	(0.0146)	(0.0124)	(0.0122)	(0.00596)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
R^2	0.511	0.561	0.651	0.922
Adjusted R^2	0.511	0.561	0.642	0.905
N. Clusters	308	307	305	305
F	583.7	122.9	136.7	37.95
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a linear regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of new hires across time. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There is a single firm level control variables: Mean Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration raises new hires by approximately $0.0388 \times 0.1 \times 100 = 0.388\%$.

Table 2.J.8: New Hires (IV) : Weighted by Mean New Hires without Controlling for Firm Size

	(1)	(2)	(3)	(4)
	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)	Log(Nb. Hires)
Log(Labor HHI)	-3.942	-1.734**	-1.544**	-1.590*
	(5.111)	(0.823)	(0.624)	(0.862)
Log(Product HHI)	0.105	-0.655***	-0.815***	-5.307**
	(1.321)	(0.0418)	(0.0275)	(2.492)
Mean Age (in years)	-0.00415	-0.0115***	-0.0131***	-0.0000799
	(0.0122)	(0.00140)	(0.000592)	(0.00251)
Share of Men	-0.976	-0.195***	-0.220***	-0.0578***
	(1.348)	(0.0632)	(0.0131)	(0.0127)
Mean Log(Value Added per Employee)	0.279	0.167***	0.175***	0.0259
	(0.185)	(0.0263)	(0.0280)	(0.0174)
Quarter x Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x Occupation FE	No	No	Yes	No
CZ x Occ. x Industry FE	No	No	No	Yes
N. Clusters	308	307	305	305
F	65.30	184.4	361.0	151.1
Observations	3175710	3175709	3165195	2620737

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variable regression using the Log(Nb. Hires) as a dependent variable. Each observation is measured at the Occupation by Industry by Commuting Zone by Quarter level, as provided in the DADS Salariés (2011-2015). Each observation is weighted by the mean number of new hires across time. Standard Errors are clustered at the commuting zone level. Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an occupation, and through quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two employee level control variables: share of men among the new hires and the mean age (in years) of the new hires. There is a single firm level control variables: Mean Log(Value Added per Worker) : the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (3): *ceteris paribus*, a 10% increase in labor market concentration lowers new hires by approximately $-1.544 \times 0.1 \times 100 = -15.44\%$.

2.K Stock

2.K.1 Occupation Based Labor Market Concentration

Table 2.K.1: Hourly Wage (OLS) : Baseline with the Employment Stock

	(1)	(2)	(3)	(4)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.0173***	-0.00621***	-0.00511***	-0.00317***
	(0.0000724)	(0.0000956)	(0.000270)	(0.000238)
Log(Product HHI)	-0.0121***	-0.000553***	-0.00151***	-0.00760***
	(0.0000698)	(0.0000799)	(0.0000889)	(0.000198)
Gender	0.106***	0.105***	0.104***	0.0930***
	(0.000151)	(0.000150)	(0.000150)	(0.000140)
Age (in years)	0.00692***	0.00692***	0.00693***	0.00644***
	(0.00000544)	(0.00000540)	(0.00000536)	(0.00000512)
Log(Value Added per Worker)	0.0941***	0.0894***	0.0853***	0.0127***
	(0.0000872)	(0.0000872)	(0.0000886)	(0.000350)
Log(Number of Employees)	0.0211***	0.0202***	0.0195***	-0.0210***
	(0.0000256)	(0.0000260)	(0.0000264)	(0.000518)
Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x 4-Digit Occupation FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
R^2	0.643	0.648	0.664	0.766
Adjusted R^2	0.643	0.648	0.663	0.754
F	692641.2	629200.0	601937.9	335249.9
Observations	20506462	20506462	20500096	20367963

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an ordinary least squares regression using the Log(Hourly Wage) as a dependent variable. Each observation is an employment contract. The sample includes all workers employed in 2014 and 2015 on January 1st (i.e, the employment stock). Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is defined over a commuting zone, a 4-digit occupation, and through the year. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The latter is the number of reported full-time equivalent number of workers in the firm over the year. The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (4): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.00317 \times 0.1 \times 100 = -0.0317\%$.

Table 2.K.2: Hourly Wages (IV) : Baseline with the Employment Stock

	(1)	(2)	(3)	(4)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	-0.0520***	-0.0227***	-0.0111***	-0.0303***
	(0.000703)	(0.00276)	(0.00371)	(0.00282)
Log(Product HHI)	0.0219***	0.0239***	0.0224***	-0.000725
	(0.000177)	(0.000149)	(0.000142)	(0.00129)
Gender	0.107***	0.106***	0.105***	0.0930***
	(0.000154)	(0.000151)	(0.000151)	(0.000140)
Age (in years)	0.00689***	0.00689***	0.00692***	0.00644***
	(0.00000562)	(0.00000551)	(0.00000537)	(0.00000512)
Log(Value Added per Worker)	0.0908***	0.0870***	0.0830***	0.0128***
	(0.000114)	(0.0000916)	(0.0000894)	(0.000351)
Log(Number of Employees)	0.0210***	0.0198***	0.0185***	-0.0205***
	(0.0000278)	(0.0000986)	(0.0000276)	(0.000523)
Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x 4-Digit Occupation FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
Observations	20506462	20506462	20500096	20367963

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from a two stage least squares regression using the Log(Hourly Wage) as a dependent variable. The instrument is described in section 2.3.2. Each observation is an employment contract. The sample includes all workers employed in 2014 and 2015 on January 1st (i.e, the employment stock). Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is defined over a commuting zone, a 4-digit occupation, and through time quarters. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The latter is the number of reported full-time equivalent number of workers in the firm over the year. The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (4): *ceteris paribus*, a 10% increase in labor market concentration lowers wages by approximately $-0.0303 \times 0.1 \times 100 = -0.303\%$.

2.K.2 Industry Based Labor Market Concentration

Table 2.K.3: Hourly Wages (OLS) : Baseline with the Employment Stock and Industry based Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	0.0123***	0.0137***	0.0167***	0.00434***
	(0.0000749)	(0.0000775)	(0.0000872)	(0.000198)
Log(Product HHI)	-0.0334***	-0.0126***	-0.0171***	-0.00920***
	(0.0000883)	(0.000103)	(0.000121)	(0.000211)
Gender: Male Dummy	0.106***	0.105***	0.104***	0.0930***
	(0.000151)	(0.000150)	(0.000150)	(0.000140)
Age (in years)	0.00690***	0.00690***	0.00691***	0.00644***
	(0.00000544)	(0.00000540)	(0.00000536)	(0.00000512)
Log(Value Added per Worker)	0.0949***	0.0884***	0.0841***	0.0126***
	(0.0000872)	(0.0000874)	(0.0000887)	(0.000350)
Log(Number of Employees)	0.0202***	0.0190***	0.0185***	-0.0211***
	(0.0000261)	(0.0000263)	(0.0000268)	(0.000518)
Year FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x 4-Digit Occupation FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
R^2	0.642	0.649	0.665	0.766
Adjusted R^2	0.642	0.649	0.663	0.754
F	686628.1	634495.5	609015.0	335305.7
Observations	20506462	20506462	20500096	20367963

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an ordinary least squares regression using the Log(Hourly Wage) as a dependent variable. Each observation is an employment contract. The sample includes all workers employed in 2014 and 2015 on January 1st (i.e, the employment stock). Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an industry, and through the year. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The latter is the number of reported full-time equivalent number of workers in the firm over the year. The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (4): *ceteris paribus*, a 10% increase in labor market concentration increases wages by approximately $0.00434 \times 0.1 \times 100 = 0.0434\%$.

Table 2.K.4: Hourly Wage (IV) : Baseline with the Employment Stock and Industry based Labor Market Concentration

	(1)	(2)	(3)	(4)
	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)	Log(Hourly Wage)
Log(Labor HHI)	0.109***	0.240***	-0.156***	-1.077***
	(0.00153)	(0.0374)	(0.0593)	(0.220)
Log(Product HHI)	-0.157***	-0.376***	0.284***	1.691***
	(0.00239)	(0.0622)	(0.0996)	(0.345)
Gender	0.106***	0.104***	0.107***	0.0908***
	(0.000163)	(0.000296)	(0.000877)	(0.000534)
Age (in years)	0.00686***	0.00672***	0.00705***	0.00644***
	(0.00000578)	(0.0000290)	(0.0000508)	(0.0000112)
Log(Value Added per Worker)	0.0871***	0.0853***	0.0834***	0.00684***
	(0.000157)	(0.000322)	(0.000177)	(0.00140)
Log(Number of Employees)	0.0144***	0.00984***	0.0225***	-0.0357***
	(0.0000859)	(0.00143)	(0.00151)	(0.00317)
Time FE	Yes	Yes	Yes	Yes
4-Digit Occupation FE	Yes	Yes	No	No
Commuting Zone FE	No	Yes	No	No
Commuting Zone x 4-Digit Occupation FE	No	No	Yes	Yes
Firm FE	No	No	No	Yes
Observations	20506462	20506462	20500096	20367963

Standard errors in parentheses

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Note: This table presents the regression output from an instrumental variables regression using the Log(Hourly Wage) as a dependent variable. Each observation is an employment contract. The sample includes all workers employed in 2014 and 2015 on January 1st (i.e, the employment stock). Log(Labor HHI) is the logarithm of the Herfindalh-Hirschman index for the labor market, as described in equation 2.2.2. A labor market is here defined over a commuting zone, an industry, and through the year. Log(Product HHI) is the logarithm of the Herfindalh-Hirschman index for the product market. It is defined as described in equation 2.2.4. The product market is defined over a commuting zone and at the industry level. There are two individual level control variables: gender (equal to one if the worker is a man, zero otherwise) and age (in years). There are two firm level control variables: Log(Value Added per Worker) and Log(Number of Employees). The latter is the number of reported full-time equivalent number of workers in the firm over the year. The former is the log of total value added (revenues minus intermediary costs) over a year divided by the number of full-time equivalent employees. The log-log specification was used because the data presents a linear relationship under this form. The log-log specification lends itself to the following interpretation of the main coefficient of interest in column (4): *ceteris paribus*, a 10% increase in labor market concentration lowers hourly wages by approximately $-1.077 \times 0.1 \times 100 = -10.77\%$.

2.L Simulation

Table 2.L.1: Simulation Summary Statistics (2015)

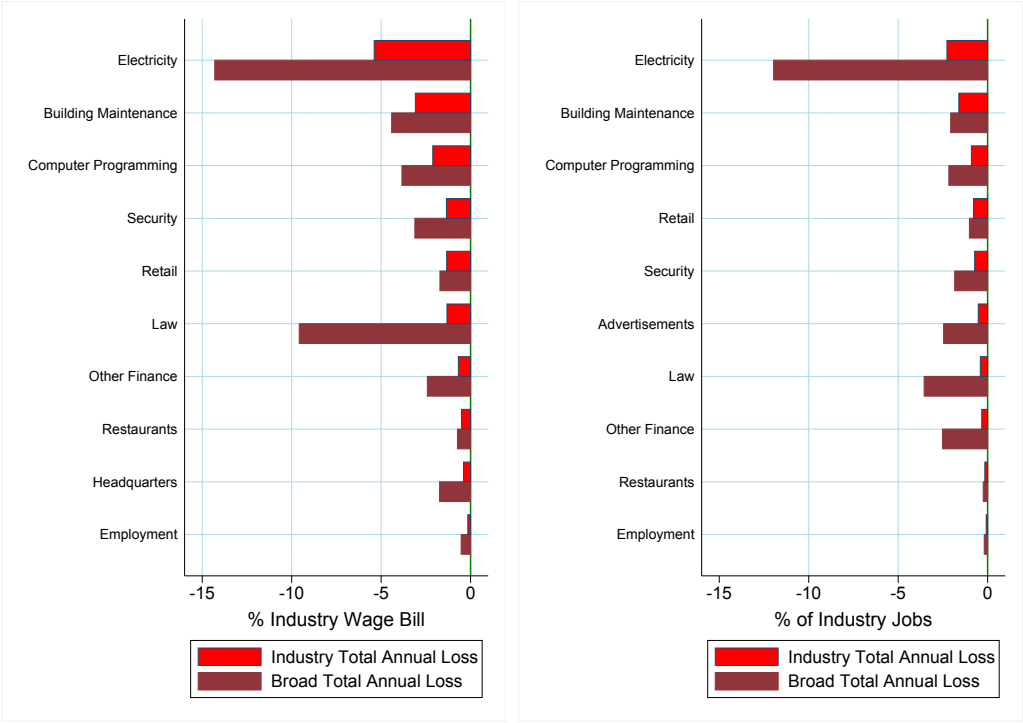
	count	min	max	p25	p50	p75	mean	sd
Original Labor Market HHI	2534847	.0006489	1	.0243952	.0688776	.2066116	0.175	0.247
Post Merger Labor Market HHI	2534847	0006489	1	.0248725	.07	.2073094	0.176	0.247

Note: The line *Original Labor Market HHI* presents the descriptive statistics for the observed level of labor market concentration for each individual. The line *Post Merger Labor Market HHI* presents the descriptive statistics for the level of labor market concentration for a worker after simulating the merger in her own industry (i.e,if a worker is in the car repair industry, then the reported HHI is the one this individual would have after the two largest companies in the car repair industry merge). This table is calculated based on our repeated cross-section (DADS Salariés) taken at the individual level.
Source: DADS Salariés and authors' calculations.

(a) (b)

Figure 2.L.1: Industry Employment Loss in Blue and White Collar Jobs

Graph 2.1(a) : Each line represents the annual expected new blue collar jobs lost for workers across France induced from a merger in *that* industry (i.e, a merger in the Retail industry would reduce annual recruitment of blue collars by 3800 jobs across France). It is calculated based on equation 2.4.2. A white collar job is defined as having an occupation number starting with 5 and 6 in the French *Professions et catégories socioprofessionnelles* occupation classification system.
Graph 2.1(b): Each line represents the annual expected new white collar jobs lost for workers across France induced from a merger in *that* industry (i.e, a merger in the Computer Programming industry would reduce annual recruitment of white collar workers by 600 jobs across France). It is calculated based on equation 2.4.2. A white collar job is defined as having an occupation number starting with 3 in the French *Professions et catégories socioprofessionnelles* occupation classification system.



(a) (b)

Figure 2.L.2: Total Wage and Employment Effects (Relative to Industry Values)

Figure 2.2(a) : Each line represents the sum of annual expected wage bill loss for new hires across France induced from a merger, reported in relation to the initial wage bill. It is calculated based on equation 2.4.1. Industry Total Annual Loss is calculated so as to include the loss to workers in the industry that merged (i.e, the impact on car repairers of a merger in the car repair industry). So, the merger in the retail industry would lead to a 2% reduction in that industry's wage bill. Broad Total Annual Loss is calculated so as to include the loss to all workers in the economy, including those in the industry that merged. So, the merger in the retail industry would lead to 2.2% reduction in annual wage bill loss across the economy, once the broad effects taken into account.
Figure 2.2(b): Each line in light red represents the annual expected new loss to new hires for workers in *that* industry (i.e, a merger in the Building Maintenance industry would reduce annual recruitment by 4.4% in the Building Maintenance industry). It is calculated based on equation 2.4.2. Each line in dark red represents the annual expected new jobs lost for workers across France induced from a merger in *that* industry (i.e, a merger in the Building Maintenance industry would reduce annual recruitment by 4.5% relative to the number of new hires in that industry). It is calculated based on equation 2.4.2.
Source: DADS Salariés and authors' calculations.

2.M Unionization Rates

Table 2.M.1: Unionization Rates (Enquête Réponse 2011)

APE Code	Unionization Rate	Number of Respondents	Industry Label
8	4,92%	61	Other mining and quarrying
10	12,48%	561	Manufacture of food products
11	5,41%	74	Manufacture of beverages
13	11,36%	88	Manufacture of textiles
14	7,14%	70	Manufacture of wearing apparel
15	12,20%	41	Manufacture of leather and related products
16	5,75%	87	Manufacture of wood and of products of wood and cork, except furniture; manufacture of articles of straw and plaiting materials
17	11,70%	94	Manufacture of paper and paper products
18	11,11%	90	Printing and reproduction of recorded media
19	26,09%	23	Manufacture of coke and refined petroleum products
20	11,33%	300	Manufacture of chemicals and chemical products
21	11,66%	163	Manufacture of basic pharmaceutical products and pharmaceutical preparations
22	12,93%	379	Manufacture of rubber and plastic products
23	18,27%	197	Manufacture of other non-metallic mineral products
24	18,38%	185	Manufacture of basic metals
25	9,98%	581	Manufacture of fabricated metal products, except machinery and equipment
26	13,10%	229	Manufacture of computer, electronic and optical products
27	11,52%	217	Manufacture of electrical equipment
28	9,12%	351	Manufacture of machinery and equipment n.e.c.
29	16,34%	202	Manufacture of motor vehicles, trailers and semi-trailers
30	20,22%	178	Manufacture of other transport equipment
31	5,15%	97	Manufacture of furniture
32	11,43%	70	Other manufacturing
33	9,95%	191	Repair and installation of machinery and equipment
35	16,22%	148	Electricity, gas, steam and air conditioning supply
36	29,82%	57	Water collection, treatment and supply
37	0,00%	15	Sewerage
38	14,66%	116	Waste collection, treatment and disposal activities; materials recovery
41	5,70%	158	Construction of buildings
42	8,45%	284	Civil engineering
43	4,09%	1051	Specialised construction activities
45	2,92%	343	Wholesale and retail trade and repair of motor vehicles and motorcycles
46	5,65%	1238	Wholesale trade, except of motor vehicles and motorcycles
47	7,94%	1436	Retail trade, except of motor vehicles and motorcycles
49	13,25%	1094	Land transport and transport via pipelines
50	21,74%	23	Water transport
51	45,71%	35	Air transport
52	16,31%	564	Warehousing and support activities for transportation
55	11,50%	113	Accommodation
56	10,32%	281	Food and beverage service activities
58	11,22%	205	Publishing activities
59	11,63%	43	Motion picture, video and television programme production, sound recording and music publishing activities
60	7,69%	26	Programming and broadcasting activities
61	13,76%	109	Telecommunications
62	6,75%	385	Computer programming, consultancy and related activities
63	7,29%	96	Information service activities
64	0,00%	9	Financial service activities, except insurance and pension funding
66	8,62%	58	Activities auxiliary to financial services and insurance activities
68	14,29%	189	Real estate activities
69	3,17%	284	Legal and accounting activities
70	4,21%	285	Activities of head offices; management consultancy activities
71	5,75%	452	Architectural and engineering activities; technical testing and analysis
72	9,68%	93	Scientific research and development
73	16,00%	150	Advertising and market research
74	7,89%	38	Other professional, scientific and technical activities
75	0,00%	5	Veterinary activities
77	2,30%	87	Rental and leasing activities
78	25,00%	32	Employment activities
79	4,84%	62	Travel agency, tour operator and other reservation service and related activities
80	17,89%	123	Security and investigation activities
81	13,77%	334	Services to buildings and landscape activities
82	6,05%	215	Office administrative, office support and other business support activities
85	3,85%	78	Education
86	13,74%	393	Human health activities
87	5,47%	128	Residential care activities
88	7,35%	68	Social work activities without accommodation
90	7,69%	13	Creative, arts and entertainment activities
91	0,00%	10	Libraries, archives, museums and other cultural activities
92	23,33%	30	Gambling and betting activities
93	0,00%	14	Sports activities and amusement and recreation activities
94	20,00%	30	Activities of membership organisations
95	0,00%	18	Repair of computers and personal and household goods
96	10,00%	70	Other personal service activities

Chapter 3

What Would Wages be Like Without Antitrust Law?

Abstract: The U.S. Supreme Court exempted Major League Baseball from the Sherman Antitrust Act. As a result, debuting players are still precluded from switching teams, rendering owners de facto monopsonies. By how much does this lower wages? Using a quasi-random discontinuity in the rule determining eligibility for Arbitration, by which a third party determines the player's wage to a level commensurate with his market value, this exemption is found to have lowered wages by at least 32% compared to its market rate.

3.1 Introduction

The fall of the labor share of income has re-affirmed the role of Antitrust law in safeguarding workers' income, as reflected by President Biden's Executive Order on *Promoting Competition in the American Economy* (Council of Economic Advisors, 2016; Posner, 2021). The relationship between both has never been clearer than in Major League Baseball (MLB). Indeed, the 1922 U.S. Supreme Court decision[1] to exempt MLB from the Sherman Antitrust Act has allowed anti-competitive practices to proliferate, perpetuate, and exacerbate. The most infamous of these practices is the so called *Reserve Clause*, according to which a player is *bound* to his owner during his whole career. As a result of negocitations between the players' union and the league,[2] this clause now effectively only lasts only six years[3].

The current system functions as a *Covenant not to Compete*[4] (CNC) pushed to its limit. Indeed, this clause is long-lasting, inexpensively enforced, and covers the complete labor market. This context offers an ideal case to study the effects of such impediments to workers' mobility on their bargaining power (Kahn, 2000).

[1] Federal Baseball Club v. National League, 259 U.S. 200 (1922)

[2] Banner (2015) suggests that the reserve clause has remained for inexperienced players as a result of senior players conspiring through the MLBPA to keep the salaries of inexperienced players low in hopes of sustaining high salaries for themselves.

[3] The 1975 *Seitz Decision* declared the *Reserve Clause* to be null, leading to a new Collective Bargaining Agreement with the Major League Baseball Players Association (MLBPA). This new agreement set the *status quo* : players became free to negotiate with other teams after six years of service time as *Free Agents*. Despite the Curt Flood Act (1998), which declared players protected by Antitrust law to the same degree as any other professional athlete, this limited form of the original *Reserve Clause* has survived. In this sense, and despite the restriction being no longer binding, it is still possible to observe the lingering effects of the Antitrust exemption. See Berri and Krautmann (2019), Kahn (2000), and Hylton (1999) for additional details on the relationship between the antitrust exemption and its implications today.

[4] Also known more colloquially as a *Non-Compete Clause*.

In particular, it allows us to quantify the importance of imperfect competition in determining wages. First, to study the effect of such a clause, one would need to observe how wages change as a result of its sudden disappearance. As we show below, the MLB's "super two" cut-off offers a case of workers having the sudden opportunity to join a semi-competitive labor market based on their accrued experience in this labor market. Second, we are in the rare situation of being able to access high-quality data for both compensation and measures of productivity for these players. Finally, it is possible to follow them across time and firm thanks to easily available linked employer-employee data for almost all members of this labor force.

To this end, we exploit a quasi-random discontinuity in the rule determining eligibility for *Arbitration*, by which a third party can determine a player's wage, through a state-of-the-art non-parametric methodology based on kernel weighted local-polynomials (Calonico et al., 2014). This rule, called the *Super Two*, is a random threshold based on the accrued work-experience of players. Players with enough work-experience to cross the threshold suddenly leave a monopsonistic labor market. Their wage is then determined by mutual agreement, with the threat of going through Final Offer Arbitration (FOA) in case of disagreement. In this case, which is seldom used in practice, both the player and the team submit a salary-proposition to an arbitrator who then selects the one which she finds most commensurate with the player's talent.

We find that wages increase by at least 45% when players become entitled to arbitration and this estimate is shown to be robust to an array of placebo and falsification tests. We posit the new post-*Arbitration* wage to be commensurate with the player's market value on the basis that (a) wages do not markedly evolve when players reach *Free Agency*, when players become free to negotiate their salary and change teams. Moreover, (b) we show this result to be consist with the theoretical literature on Final-Offer Arbitration by revisiting the model of Brams and Merrill (1983). (iii) This allows us to conclude that the the inability of players to change teams without the agreement of their team owner results in wages being depressed by at least 32% below their market value.

These results contribute to several literatures. The first includes research on the prevalence and impact of monopsony in labor markets. Recent examples include observational studies such as Azar et al. (2017b), which measures labor market power through a proxy *Herfindalh Hirschman Index* (HHI) or, Azar et al. (2019), which uses the same online job board data but posits structural modelling assumptions. These works are complemented by experimental studies such as Dube et al. (2020), who uses the Amazon's Mechanical Turk Platform to evaluate workers' labor supply elasticity. In general, these papers detect significant wage markdowns (i.e. wages are found to be below the Marginal Revenue Product). This has led some to question the role and responsibility of Antitrust policy in the promulgation of monopsony power (Naidu et al., 2018b; Posner, 2021). This paper contributes to this literature by establishing a *direct* link between monopsony power and antitrust law, by providing additional evidence for the capacity for monopsony to lower wages, and by complementing the methodological approaches of past studies by using state-of-the-art quasi-experimental methods (Calonico et al., 2014).

Second, this article contributes to the literature on *Covenants not to Compete*. This literature builds on the observation that these covenants are pervasive among the American workforce and even among low income workers (see Starr et al. (2015)). This suggests that their real purpose is to lower worker mobility and provide additional bargaining power to employers. These covenants have been found to lower wages, as recently evidenced by Balasubramanian et al. (2017) who measured changes in technology workers' wages when Hawaii banned these clauses. However, as noted by Council of Economic Advisors (2016) more research is needed. This paper examines an extreme version of these covenants and, in this sense, not only provides further

evidence, but also an *upper bound estimate* of the effect of these clauses on wages; suggesting that non-competes can have potentially deleterious effects.

Third, this paper contributes to the literature studying alternative ways to determine wages. It shows that the threat of Final Offer Arbitration can lead wages to effectively match those of semi-competitive labor market (Ashenfelter and Hyslop, 2001). This opens the question of the more generalized application of Final Offer Arbitration as a threat which can be used by both workers and firms to settle wages in situations where market forces may be inadequate (Mas, 2006).

Fourth, this paper also adds to *sports economics* which uses the industry as a "labor market laboratory" Kahn (2000). In particular, outside of baseball, the *Reserve Clause* can be seen as similar to the *Bosman Ruling* (1995) from the European Courts of Justice, which allowed soccer players to move across European countries to play in different national leagues and teams. Research such as Binder and Findlay (2012) argued that this decision resulted in significant improvement in the performance of *Champions' League* teams. This suggests that, perhaps as in more usual industries, competition in the labor market improves the quality of the services produced. Within baseball, this paper contributes to the numerous bodies of research probing the relationship between players' wages and their performance Kahn (2000). This literature builds on Scully (1974)'s seminal paper, which attempted to use team-level accounting data along with match-level data to recover players' marginal revenue product. More recently, Berri and Krautmann (2019) provided an alternative approach based on using the labor share of income within baseball (when known) and in other sports (such as in basketball) to induce the loss of income of players due to the *Reserve Clause*. They suggest that the labor share rose from 20% in the 1950s to perhaps 60% in the early 1980s. In contrast, this paper provides a novel way to study wage suppression in Major League Baseball through the use of the *Super Two* cut-off[5]. This novel method is based on transparent identification conditions, avoids the selection bias which plagued previous studies, and can be subjected to several rigorous placebo and falsification tests.

This paper develops these points in the following three sections. Section 3.2 presents the data along with the research design upon which this paper is built. Section 3.3 provides estimation results and robustness tests. Finally, Section 3.4 discusses the interpretation of these results.

3.2 Research Design

3.2.1 Data

We rely on freely accessible data on Major League Baseball players, their service time, and their salaries. The available sample covers the period 2010 to 2018. The players' salaries and service time are available from the website *Baseball-Reference.com*. This website is an exhaustive repository of information on Major League players and one of the only one which also keeps records of players' service times. Players' performance ratings and salaries from the freely accessible *Lahman Database* are used to complement the former. This database is well known within sabermetrics and has already been used within the context of economic research in articles such as Hakes and Sauer (2006). This dataset is complemented with the annual *Super Two* cut-off dates found

[5]We are not, however, the first to use this cut-off within the economic literature. Papps (2010)'s study of *efficiency wages* uses the cut-off to instrument for players' income. This allowed him to conclude that raising income also raised players' performance.

on the website *MLBTradeRumors.com* and are reported in Table 3.B.2.[6]

For the purposes of this study, players without salary or known service time were dropped from the sample. However, all players included in the Lahman database are present in the final sample. Summary statistics for the final sample are provided in Table 3.B.1. Given that active rosters are comprised of 25 players and 30 teams, the nearly 800 players available each year in our sample can be said to cover the population of interest. This is confirmed by considering Table 3.B.3 which displays the share of players in already two years of accrued service time (explained below). The *Super-Two* cut-off is calculated such that 22% of these players become eligible for arbitration. We observe an average of 24% which suggests that our sample does not significantly depart from the overall population.

This paper exploits the following variables throughout:

- Yearly income : this variable is provided in nominal U.S. dollars and does not include potential one-off bonuses. This variable is used in this paper after being transformed using the natural logarithm. This variable is available for 7603 observations. The mean income approaches 4 million U.S. dollars (for reference, the contractual minimum wage in the MLB is of 545 000 U.S. dollars for 2018). Below the *Super Two* threshold, the mean salary is of only USD 697 741. Above, this same amount has a mean of USD 6 074 523. This reflects not only the effect of *Arbitration* but also the experience accrued by and attrition of players above the threshold.

- Weight : measured in pounds during the rookie year. This variable is available for 7407 observations. The mean weight during the rookie year is 213 pounds with 211 for those below and 214 for those above.

- Rookie Year: calendar year during which a player accrued at least 45 days of service time or exceeded 130 At Bats (AB), approximately the number of opportunities to bat, or 50 innings pitched in Major League Baseball. The mean rookie year is 2009, with those below the *Super Two* cut-off at 2012 (more recent players) and those above the cut-off at 2006 (older players).

- At Bats per Home Run (ABHR) : the ratio of the number of At Bat opportunities to the number of Home Runs hit (a home run is the most valuable hit allowing all players on base to score for their team).[7] With the Runs Batted In (RBI), it constitutes a measure of player performance.[8] This variable is available for only 3335 observations which is to be expected given that many players never hit any home runs. The mean ABHR is at 51 but at 54 below the *Super Two* threshold and 50 above.

- Runs Batted In (RBI) : This variable attributes runs to players whose batting caused the run to occur (i.e, reach a new base). With the At Bats per Home Run (ABHR), this variable constitutes a measure of player performance. This variable is available for 7015 observations. The average RBI is 21. Below the *Super Two* cut-off, the mean RBI is 17. Above, it is 23.

[6] This data was collected based on the code available at: https://github.com/jason-sa/baseball_lin_regression.git which comes with a MIT License to "use, copy, modify, merge, publish".

[7] Many players never hit a home-run. We cannot substitute their value with a zero as it would require including a zero in a denominator.

[8] This measure of performance is transparent and easy to interpret. However, more sophisticated measures of performance such as Wins Above Average (WAA) or Wins Above Replacement (WAR) may more accurately capture players' contribution to a team. These measures rely on proprietary formula and are harder to interpret. Although unreported, we have replicated results which rely on Ats Bats per Home Run (ABHR) and Runs Batted In (RBI) using both WAA and WAR as measured by Baseball-Reference. In all cases, we find no qualitative change in results relative to those in the main body of the text.

– Days to *Super Two* cut-off: this is the running variable used in the regression discontinuity design implemented in this paper. The service time is re-centered as the number of days difference to the *Super Two* cut-off. The service time is the number of days a player spends on a MLB's active roster of 25 players. A player accrues a year of service for at least 172 days out of the 183 spent on this roster. It is recorded in the format YY.DDD to reflect that days spent above 172 do not extend to service time in the following year. Player's above the *Super Two* cut-off (0 in our data) can request *Salary Arbitration* and, above six years of service, can attempt to change teams as a *Free Agent*. This variable is available for 7603 observations. The mean days of service time if 319. Below the cut-off, the mean is at -252 days. Above, it is at 683 days.

– Within +/- 10 days of cut-off: this is a dummy variable equal to one if an observation is within 10 days of the cut-off. 2% (or 126 observations) of the sample is within ten days from the cut-off, composed of 2% (or 55 observations) of observations below the cut-off and 2% (or 71 observations) of observations above the cut-off.

As a robustness measure, the main specifications were also run on an alternative dataset which includes a better coverage of players. This data is the freely accessible *Cot's Baseball Contracts* database.[9] It is slightly larger (8690 observations) but could not be easily linked to the Lahman database for lack of player identification numbers. This meant that some of the robustness checks in the paper would not have been possible using solely this data. However, it is used to run the main specification (reported in Table 3.B.5 and in Figure 3.A.6).

3.2.2 Research Context and Identification Strategy

Our first goal is to measure the degree to which wages rise when players become eligible for *Arbitration*.[10] In Section 3.4, we will argue that this effect is akin to the one of accessing a competitive market. *Arbitration* is a provision appearing in its modern form in the 1970 Collective Bargaining Agreement (CBA) negotiated by the Major League Baseball Players' Association (MLBPA). A player eligible for *Arbitration* can be heard by a third party arbitrator to resolve a salary dispute. In this case, both the team and the player submit a wage proposition and the arbitrator chooses one of the two after having made an appraisal of the player's performance in relation to his peers and their respective salaries. In practice, however, the threat of *Arbitration* often leads the parties to reach an agreement beforehand.

Players become eligible for *Arbitration* by accruing service time, as a result of having been placed on a team's active roster for a given number of days. Players with already two years of service time and having accrued an amount of days above the *Super Two* cut-off (described below) are eligible for Arbitration. One cannot directly compare players with and without eligibility for arbitration because teams will be selective in which players they allow to become potentially eligible. As shown in Figure 3.A.1, which displays the amount of service time accrued by players the year of their drop-out from Major League Baseball (or of our dataset), there are two significant moments: when players become eligible for *Arbitration* (red bar, in year 3) and when players become *Free Agents* (green bar, at the end of year 6). In turn, this means that players are selected are likely to differ in potentially unobserved ways from those which dropped out earlier in their career.

[9] This data was used in Papps (2010).

[10] A potential extension of this paper could include looking at the effects on player mobility and career length.

In turn, it would be misleading to compare the incomes of those eligible and non-eligible players since they are likely to differ in their characteristics. This can be seen by consider the descriptive statistics of Table 3.B.1. The Run batted In (RBI) of players with arbitration is 23.23 whereas it is only of 17.65 for those non-eligible. Similarly, the mean At Bats per Home Run (ABHR) of the former is lower (50.28) than the latter (54.18). As such, there is evidence that selection effects would generate a composition effects : eligible players are also, on average, better players than non-eligible players.

This problem can be alleviated by comparing who are randomly selected into eligibility.[11] We do so by considering players below and above the so-called *Super Two* threshold. This threshold determines a player's eligibility for *Arbitration* and is the result of the Major League Baseball Players Association collective bargaining agreement of 1997 (Prospectus, 2021). To calculate it, the MLB league selects all players with at least 86 days of service on the active roster (or injured list) that year. It then chooses the cut-off such that it includes the top 22% of players in terms of accrued service time. Although teams can avoid eligibility for *Arbitration* by keeping players off the active roster, the cut-off is largely considered to be unpredictable and a source of injustice among both owners and players. The historical thresholds are presented in Table 3.B.2 whilst the distribution of players with already two years of accrued service time is displayed in Figure 3.A.3. We observe that the distribution of players who are eligible for arbitration takes no noticeable pattern. These irregular conditional distributions suggest, informally speaking, that the value of the random threshold is indeed random.

More formally, this identification condition can be tested thanks to the McCrary (2008) test. The test supposes that, in the absence of manipulation on behalf of team owners, there should be evidence of smoothness in the density function of players in terms of service time. In contrast, the density should be discontinuous at the cut-off if team owners are able to predict the cut-off and assure that the minimal number of their players become eligible for arbitration. As displayed in Table 3.B.1, 126 player-years in our data are within ten days of the cut-off (55 below and 71 above). Informally speaking, one would expect more observations below the threshold than above if teams could predict the cut-off date. More formally, we can consider a graphical representation of this test and density function, as provided in Figure 3.A.7. Table 3.B.7 shows the formal statistical test from assuming the null hypothesis of continuity in the density function. Despite testing six different parametric forms (polynomial approximation of degree one to six), discontinuity in the density at the cut-off is always rejected at the 5% level.[12]

In order to assess the reliability of the McCrary (2008) test, we also apply it to the three-year service time cut-off. One could legitimately expect that owners let go of players before the arbitration process awards them a higher salary. Some evidence of this behavior is detected by the test, as shown in Figure 3.A.8. The 95% confidence intervals for the approximated density around the three-year cut-off (516 days) do not overlap. The figure is consistent with our priors because the density of players below the cut-off is statistically greater than the density of players above the cut-off (and therefore, eligible for *Arbitration*). Considering the evidence through different levels of parametric flexibility, as reported in Table 3.B.8, there is some evidence suggesting that, for a sufficiently smooth density (which can be well approximated by a low degree local kernel-weighted polynomial), players' service time becomes discontinuous at the three-year cut-off. So, players' service time is likely to be

[11] A related issue is attrition bias. A player who is above the cut-off is more likely to continue playing than one who falls below. A potential solution discussed in Dong (2019) involves controlling for the inverse mills' ratio promoted by Heckman (1979a), as implemented by McCrary and Royer (2011). We implemented this approach using age fixed effects, years fixed effects, and the player's rookie service time as instrument in the probit (i.e, selection) equation. Results were not qualitatively changed.

[12] Note, this result is not surprising given that the threshold is determined by the aggregate behavior of the different teams.

subject to manipulation around the three-year cut-off[13]. This suggests that the *McCrary test* is reliable and, in turn, that the *Super Two* cut-off is not subject to manipulation.

Having considered the identification condition, we can now turn to descriptive evidence of the impact of *Arbitration* on income. This can be appreciated graphically. Figure 3.A.2 displays the median log-income of players within 12 months worth of service time around the cut-off[14]. There is a clear jump in the income of players who have passed the threshold. This effect is particularly strong for those who are within 30 extra days of the cut-off. This jump is grossly equal to one log point ($\exp(14-13)-1 \approx 172\%$) in median income. However, more sophisticated econometric tools are needed in order to assess the impact of arbitration. By considering only players near the cut-off, the jump will reflect the effect of *Arbitration* on comparable players, that is, on players with equal chances of dropping out from the league. This will provide us with an estimate which is robust to the underlying attrition bias[15].

3.2.3 Econometric Method

We implement our research design by use of the non-parametric local kernel-weighted polynomial approximation methodology developed by Calonico, Cattaneo, Farrell, Jansson, Ma, Titiunik and Vazquez-Bare. Their approach is flexible, well-documented, and has received acceptance within the profession (e.g, see Pons and Tricaud (2018)). As this methodology is non-parametric in nature, it prevents inference based on arbitrary parametric specifications. This methodology takes into account the bias which is naturally present when employing the regression discontinuity design, and provides a formula for robust standard errors at the same time. Moreover, this methodology is well known by practitioners thanks to its availability in popular econometric software (Calonico, 2017).

For completeness, this paper provides a brief and intuitive overview of the methodology. A more complete introduction can be found in Calonico (2017) and the full details in Calonico et al. (2014), Calonico et al. (2015), and Calonico et al. (2019). Assuming a cross-section for simplicity[16], the observations are indexed by $i = 1$ to $i = N$. We are interested in an outcome denoted by Y_i. The treatment is administered if the running variable X_i is above the threshold x^*.

Ideally, we would want to compare the value of Y_i above and below the cut-off but, for a given individual, we only observe one or the other. This requires us to make further assumptions on the rule determining the attribution of the treatment so that we recover a consistent estimate of the Average Treatment Effect (ATE). This assumption is akin to stating that, near the cut-off, the running variable is not correlated with any unobservable. We denote this assumption as the *Identification Condition*. We can then formulate the treatment effect τ as

$$\tau = \lim_{x \uparrow x^*} \mathbb{E}(Y_i|X_i = x) - \lim_{x \downarrow x^*} \mathbb{E}(Y_i|X_i = x) \tag{3.2.1}$$

To obtain a sample counterpart to τ, which we will call $\widehat{\tau}$, we need to estimate the conditional expectation function $\mathbb{E}(Y_i|X_i = x)$ above and below the cut-off x^*. We approximate this function by a local kernel-weighted polynomial function. More precisely, we can solve for the vector of parameters β, above and below

[13] Although unreported in this paper, the same exercise can be run at the year six cut-off, when players become *Free Agents*. A mass of players drop out at the cut-off and there is some evidence of discontinuity in the density function at the cut-off.

[14] This result is invariant to using means instead of medians.

[15] As expected and discussed above, the effect of *Arbitration* net of attrition bias is much lower, as revealed by Table 3.B.4.

[16] We treat our data as a repeated cross-section in effect.

the threshold, which provides the best fit. We using the following formulas:

$$\widehat{\boldsymbol{\beta}_+} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{N} 1(X_i \geq x^*) \left[Y_i - \boldsymbol{r}_p(X_i - x^*)\boldsymbol{\beta}\right]^2 \mathbb{K}_{h_n}(X_i - x^*) \tag{3.2.2}$$

and

$$\widehat{\boldsymbol{\beta}_-} = \underset{\boldsymbol{\beta}\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{N} 1(X_i < x^*) \left[Y_i - \boldsymbol{r}_p(X_i - x^*)\boldsymbol{\beta}\right]^2 \mathbb{K}_{h_n}(X_i - x^*) \tag{3.2.3}$$

where $\mathbb{K}_{h_n}$ is a kernel function with bandwidth h_n and $\boldsymbol{r}_p$ is the polynomial expansion of degree p of the variable x. That is, $\boldsymbol{r}_p(x) = (1, x, ..., x^p)$.

Let's explain the intuition. Restricting ourselves to one side of the cut-off (implemented by the indicator function $1(\cdot)$, taking a value of one if its argument is true, in the formulas), these formulas solve a least-squares problem for a vector of parameters $\boldsymbol{\beta}$ which approximate the conditional expectation function with a polynomial expansion of degree p around the cut-off x^*. However, the values of Y for values of X far away from the cut-off are likely to become increasingly determined by factors other than the treatment status. We therefore desire to discount the observations with X far away from x^*. On the other hand, there are probably few observations that are very close to x^*. So, the econometrician is faced with a bias-variance trade-off. This trade-off is translated into our estimation problem by the use of kernel functions $\mathbb{K}$ which discount observations at a distance from x^* at a speed commensurate with the bandwidth h_n. In our case, we use the Epanechnikov (parabolic) kernel defined by :

$$\mathbb{K}^E_{h_n}(X) = \frac{3}{4h_n}(1 - (X/h_n)^2) \text{ with } |X/h_n| \leq 1$$

We select the bandwidth minimizing the theoretical Mean Squared Error (MSE), as shown in Calonico (2017). We can then write the conditional expectation function as

$$\lim_{x\uparrow x^*} \widehat{\mathbb{E}(Y_i|X_i = x)} = \boldsymbol{e}_0'\widehat{\boldsymbol{\beta}_+} \tag{3.2.4}$$

and

$$\lim_{x\downarrow x^*} \widehat{\mathbb{E}(Y_i|X_i = x)} = \boldsymbol{e}_0'\widehat{\boldsymbol{\beta}_-} \tag{3.2.5}$$

with $\boldsymbol{e}_0 = (1, 0, ..., 0) \in \mathbb{R}^{p+1}$ being the counter-part of $\boldsymbol{r}_p(x) = (1, x, ..., x^p)$ as $x \to 0$.

The regression tables in this paper will usually present three cases based on these estimates. The first is the *conventional* estimate. It is based on ignoring the bias introduced by kernel weighting and, therefore, includes observations for which the local conditional Independence assumption may not hold. The second, called the *bias-corrected* takes this bias into consideration by estimating a higher order polynomial of degree q, using it to measure the bias, and subtracting this measure from our original p degree polynomial approximation. The third, denoted as the *robust* estimator, also adapts the estimators' standard errors to take into account the bias. Indeed, given that the bias must be estimated, the confidence intervals need to reflect the additional variability induced by this step. More details on the asymptotic behavior of these estimators can be found in the original publications (Calonico et al., 2014, 2015, 2019).

3.3 Empirical Results

3.3.1 Main Result

We now present the main specification of this research. We look at the discontinuity in log-income above and below the *Super Two* cut-off. There is very strong evidence that income rises once players become eligible for *Arbitration*. The result of this exercise is presented in Figure 3.A.5 using a kernel weighted local polynomial of degree four. There is clearly a large jump in income at the cut-off and, despite the high degree of flexibility in the parametric form, there is no sign that the polynomial below the threshold is increasing as it reaches the cut-off. Moreover, income continues to rise with service time above the cut-off, as one would expect when players develop experience in the game and are no longer subject to restrictions to competition.

These results are confirmed more formally when considering the regression output. This is provided in Table 3.B.4. Each column presents a different polynomial degree, from one to six. Although the estimated effect from an increase in wages ranges from 47% (column 4) to 92% (column 1), the effect does not seem to fall as we increase the flexibility of the polynomial. Indeed, the effect is up to 58% with a polynomial of degree six (column 6). In all cases, whether we consider the conventional or robust estimator, we find highly statistically significant effects at the 0.1% level. All in all, this provides strong evidence that wages are suppressed by the *Reservation Clause*.

The results point towards the existence of non-random attrition among players. We expect that players who survive within MLB may do so based on some unobserved characteristic. Failing to realise this would lead the econometrician to conflate the effect of *Arbitration* with the effect of surviving to the point of being eligible for *Arbitration*. This is akin to the *Ability Bias* within the context of the returns to education, where the econometrician conflates the underlying ability to succeed, in both education and in the workplace, with the effect of having received an education. In both cases, we expect the returns to *Arbitration*/education to be overestimated as a result. Compared to the 1 log-point effect found in the previous section, the effect reported in this section based on the regression discontinuity design can is nearly half, at 0.47 log-points. Therefore, the use of the discontinuity design is justified.

3.3.2 Heterogeneity

The analysis of these results can be deepened by considering the heterogeneity in treatment effects. We do so by considering the *Quantile Treatment Effects in Regression Discontinuity* model of Frandsen et al. (2012). By considering the quantiles, we can evaluate the heterogeneity in treatment effect in terms of the player's place in the income distribution. This method is akin to the one developed by Calonico (2017) but focuses on quantiles of the dependant variable. Intuitively speaking, it compares the θ-th quantile of the dependant variable above and below the cut-off. These quantiles are estimated using a local-linear kernel weighted approximation (i.e, polynomial of degree one). of the conditional expectation function and using these estimates to recover the density function, and its respective quantile function, above and below the cut-off. The results are provided in Table 3.B.11 and, for clarity, in graphical form in Figure 3.A.19. There is clear evidence of heterogeneous treatment effects. The *Arbitration* process appears to increase the income of all players, but in particular for those with initially higher income. This suggests that the *Reserve Clause* is particularly damaging for high

performance and pay individuals.

3.3.3 Robustness

The previous estimates are now supplemented with five robustness tests.

Measurement Error. We test the main specification using the *Cot's Baseball Contracts* dataset. This allows us to assess the reliability of our results to potential outliers or mismeasured wages. This larger and more complete dataset provides nearly identical results. Figure 3.A.6 shows the same discontinuity and the same shaped conditional expectation function. However, it reports different outliers (red dots). Its respective regression output, provided in Table 3.B.5 reveals even higher effects from arbitration. Now, the maximum effect if of 97% and the lowest effect at 69%, with 72% found for the most flexible specification using a polynomial of degree six.

Falsification using previous income. If the threshold is truly locally random, there should be no detectable effect at the cut-off on the previous year's income. As expected, we fail to detect any jump in the previous year's income around the *Super Two* cut-off. This is shown graphically in Figure 3.A.4 which displays the relationship between players' service time and their previous year's log-income. There is no discernible discontinuity at the threshold. This is confirmed more formally by considering its respective regression output, in Table 3.B.6. This table presents six specifications, where each column provides the estimates for a polynomial approximation of degree one to six. There is no evidence of a statistically significant discontinuity at the 5% level when considering conventional or robust estimators.

Falsification using irrelevant variables. Similar falsification tests can be run on variables which, *a priori*, should not be impacted at the cut-off. We use two measures : Year and Weight of the player when he became a rookie. Graphical representation of the tests are provided in Figure 3.A.14 using a polynomial of degree 4. No clear discontinuity at the cut-off can be detected. Similarly, the p-values reported in column (1) and column (2) of Table 3.B.9 reveal no statistically significant effect.

Falsification using relevant variables. A more demanding falsification test involves testing for a discontinuity in variables which should be in interest of team owners to manipulate. We use two measures : the At Bats per Home Run (ABHR) and the Runs Batted In (RBI). Whilst the latter is likely doing a disservice to pitchers, the former will only concern hitters. Again, Figure 3.A.17 reveals an absence of statistically significant discontinuity; perhaps reflecting owners keeping exceptional players away from the cut-off to avoid the risk of *Arbitration*. The same can be said in terms of ABHR, where the polynomial approximation ends at a lower point below the threshold than it restarts above the cut-off. These findings are confirmed by the last two columns of Table 3.B.9 where no statistically significant effect is detected.

Placebo cut-offs. If the identification strategy truly holds, there should be no detectable effect at other cut-offs. To test this implication, we run a series of placebo tests at other cut-offs. We report all tests run in the interval [-200;-5] and [5;200] days around the cut-off (the test are run at the five day interval) in order to avoid being selective in our reporting. The results from these tests are reported, in the form the bias-corrected estimated treatment effect with 95% non-robust confidence intervals, in Figure 3.A.11. For these tests, we drop values above or below the cut-off in order to not contaminate the estimator with the true cut-off. Apart from two estimates which are significant at the 5% level, there is no systematic effect detected[17].

[17] Finding two estimates to be statistically significant out of 78, is not particularly surprising given the problem induced by multiple

All in all, these tests provide strong evidence that the identification strategy implemented in this paper does not lead to any unexpected implications or require any unrealistic assumption.

3.4 Interpretation

3.4.1 Discussion

To better understand the estimation results of Section 3.3 and, by extension, the degree to which wages are depressed by the inability of players to change teams, it is important to clarify how the post-*Arbitration* wage relates to other tangible and economically relevant market structures. For example, Krautmann (1999) argued that the *Free Agent* market is competitive and that, for this reason, their wages must be close to their Marginal Revenue Product (MRP; i.e, the marginal contribution of a player to his team's revenue).

To the degree that players post-*Arbitration* wages are similar to those of *Free Agents*, one may be tempted to qualify the post-*Arbitration* wage as equal to the MRP. However, there are several limitations to this argument. As noted in Bradbury (2013), the MLB labor market is not perfectly competitive and those imperfections are likely to translate into *Free Agents* being paid below their MRP. He argues that some players are ready to trade-off income for long-term contracts, hedging the risk of injury. Other players have a preference for their home team and are ready to cut their salary expectations to play for this team. Similarly, Rottenberg (1956) notes that many MLB teams and players are, at best, *imperfectly* substituable. These frictions are likely to render *Free Agents'* salaries below those of their MRP.[18]

Rather, we suggests that it is more natural to interpret the estimated results in terms of the value provided by a semi-competitive market. That is, we consider the post-*Arbitration* wages to be similar to those of *Free Agents'* and that the latter, being the result of an *imperfectly competitive market*, represents a relevant and natural benchmark to assess the damage of restricting the mobility of players across firms. This benchmark is pertinent for several reasons. First, the (potentially imperfect) market outcome is considered the natural benchmark within the context of Antitrust policy. Second, the imperfectly competitive market is not a hypothetical market structure: it is both real and observed despite being vague in its definition. Lastly, the "market rate" represents the *natural* rate which can be obtained without introducing an additional regulatory body. For these reasons, if the eligibility for *Arbitration* raises income by 47% at the minimum (or 92% at the maximum), then the restriction to player mobility can be considered to have depressed wages by 32% at the minimum (and 48% at the maximum) *below their market rate.*[19].

3.4.2 Empirical Evidence

To show this interpretation to be valid, we now provide empirical evidence. In Section 3.4.3, we supply theoretical support in favor of this interpretation by re-visiting the model of Final Offer Bargaining of Brams

testing.

[18] Under the light of *partial identification*, that is, by considering the post-*Arbitration* wage to be below that of the MRP, one can argue that our results provide a *lower bound* estimate of the labor share of income. In other words, if the eligibility for *Arbitration* raises income by 47% at the minimum (or 92% at the maximum), then the pre-*Arbitration* wage must be at most 68% (or at least 52%) of the MRP. Mathematically speaking, if the initial wage is w, the *Arbitration* effect τ, and the Marginal Revenue Product y, then we have $w \times (1+\tau) = y$ which implies the labor share of income $\frac{w}{y} = \frac{1}{1+\tau}$.

[19] If the initial wage is w, the wage reducing effect of the *Reserve Clause* is θ, and the competitive market wage W, then we have $W \times (1-\theta) = w$ which implies that, for treatment effect τ, we obtain $\theta = \frac{\tau}{1+\tau}$.

and Merrill (1983).

We argue that if the wages set for players eligible for *Arbitration* do not significantly differ from the rate determined by players who become Free Agents (and therefore have their wage determined by market forces), then the former can be interpreted as being akin to the latter. This implies that we should not see unexpected changes in income growth when players gain the right to become a free agent. To show this, we run a wage-growth regression which identifies the average income growth rate across the player's life-cycle. We implement this regression by running an exponential regression model[20] (PPML) on income, whilst controlling for age, team, year, and player fixed-effects (Blackburn, 2007). We add a set of variables which measure the elasticity of income to an additional year of service time. The results are available in Table 3.B.10 along with different specifications allowing the reader to ascertain the stability of the reported coefficients to different specifications. The coefficients of column (4), which are based on all control variables and are therefore considered to be the most reliable, are plotted in Figure 3.A.18 for clarity along with their 95% confidence interval, using heteroskedasticity robust standard errors. The figure reveals that wages rise drastically, *ceteris parisbus*, with years of service time when players become eligible for *Arbitration*, at year three. However, at year six (green vertical line), the growth elasticity is not particularly strong and appears part of a downward growth trend. This suggests that the post-*Arbitration* wage is not markedly different from the *Free Agent* wage. For this reason, it is plausible to qualify the post-*Arbitration* wage as being similar to its market value.

3.4.3 Theoretical Evidence

Finally, we now provide a simple model which shows that the post-arbitration wage is commensurable with the market rate. This model builds on the approach to Final-Offer Arbitration proposed by Brams and Merrill (1983). We show that when the arbitrator bases her expectations on wages of players determined by a market, the compensation of players eligible to arbitration (but not the free market) is equal to the salary of players on the market. It follows that one should interpret the findings obtained in Section 3.3 as the effect of players having their wage being determined under a monopsonistic regime to reflecting market forces and rates.

Setup. In this model, player P and team T submit wage-offers to arbitrator A. We suppose the arbitrator has a reference wage w_A in mind for player P which reflects what she perceives to be her value. The arbitrator chooses the wage-offer which is closest in absolute value to this reference wage. However, from the point of view of the team and of the player, this reference wage w_A is random and follows a well-behaved distribution F_{w_A}, such that $w_A \sim F_{w_A}$.

Strategies. Player P and team T submit wages w_T and w_P. The offer proposed by the player must always be above or equal the wage offered by the team, or else there is a possibility of a pareto improving wage agreement. As such, we take it that

$$w_P > w_T \tag{3.4.1}$$

The arbitrator chooses the offer w_T if it its closest (in absolute value) to the target wage w_A determined by the arbitrator. As stated by Farber (1980), this implies

$$w_A - w_P < w_T - w_A \tag{3.4.2}$$

[20] This alternative to the usual log-log regression is robust to potential heteroskedasticity of the error term, as advocated by Bellégo et al. (2021). The coefficients can nonetheless be interpreted as elasticities.

Proof. Offer from player T is chosen if

$$||w_T - w_A|| < ||w_P - w_A|| \Longleftrightarrow \sqrt{[w_T - w_A]^2} < \sqrt{[w_P - w_A]^2} \quad (3.4.3)$$

which implies by virtue of the square-root function being always positive that

$$[w_T - w_A]^2 < [w_P - w_A]^2 \quad (3.4.4)$$

We now look at the different cases regarding the relationship between w_P, w_T and w_A.

- Case (a) : $w_A > w_P > w_T$. We know $0 > w_P - w_A > w_T - w_A$ which implies that $0 < [w_P - w_A]^2 < [w_T - w_A]^2$ in contradiction of the assumption of Equation 3.4.4.
- Case (b) : $w_P > w_A > w_T$. By assumption of equation 3.4.4 and knowing that $w_P > w_A$, we see that $w_P - w_A > ||w_T - w_A||$ such that we can write $w_A < w_P - ||w_T - w_A|| < w_P - [w_T + w_A]$ by the triangle inequality. We then obtain $w_P + w_T > w_P - w_T > 2w_A$. Given this, we can obtain equation 3.4.2 by direct re-arrangement.
- Case (c) : $w_P > w_T > w_A$. This implies that $w_T - w_P > w_A - w_P$ which can be further written as $w_T - w_A > w_T - w_P > w_A - w_P$ because $w_A < w_P$ as in Equation 3.4.2.

□

We can now consider the associated probability given by

$$Pr(\text{Arbitrator chooses offer } w_T) = Pr(w_A < \frac{w_P + w_T}{2}) = F_{w_A}(\frac{w_P + w_T}{2}) \quad (3.4.5)$$

Preferences. Assuming the absence of risk-aversion among parties and that the game is zero-sum, as in Brams and Merrill (1983), we model the player's expected utility by

$$\mathbb{E}_{F_{w_A}}(U_P) = \underbrace{\left[1 - F_{w_A}(\frac{w_P + w_T}{2})\right] \times w_P}_{\text{Util if } w_P \text{ is chosen}} + \underbrace{F_{w_A}(\frac{w_P + w_T}{2}) \times w_T}_{\text{Util if } w_T \text{ is chosen}} \quad (3.4.6)$$

such that she weights her utility according to the arbitrator's probability of selecting her offer. The counter-part profits for the firm, who obtains revenue Y_P from player P, are given by

$$\mathbb{E}_{F_{w_A}}(\Pi_T) = Y_P - \left[1 - F_{w_A}(\frac{w_P + w_T}{2})\right] \times w_P - F_{w_A}(\frac{w_P + w_T}{2}) \times w_T \quad (3.4.7)$$

Nash Equilibrium. We now focus on the pure-strategy Nash Equilibrium described by Brams and Merrill (1983) in their Theorem 1. They show that if there exist a pure-strategy, then it equals

$$\begin{cases} w_T = M_{F_{w_A}} - 2F'_{w_A}(M_{F_{w_A}})^{-1} & \text{for player P,} \\ w_P = M_{F_{w_A}} + 2F'_{w_A}(M_{F_{w_A}})^{-1} & \text{for team T,} \end{cases} \quad (3.4.8)$$

where $M_{F_{w_A}}$ is the median of distribution F_{w_A}, such that $F'_{w_A}(M_{F_{w_A}}) = 0.5$, and $F'_{w_A}(\cdot)$ is the associated probability density function. As shown in Brams and Merrill (1983), this equilibrium not only requires $F'_{w_A}(M_{F_{w_A}})$ to exist, it also requires it to be non-zero. We will show that in the sub-case considered in this paper, the additional requirements which are necessary for this equilibrium to exist and to be global are met.

Proof. For completeness and readability, we reproduce the proof for this equilibrium in Brams and Merrill (1983) whilst adapting it to avoid normalizing the median to zero. We solve for this Nash Equilibrium by maximizing

equations 4.5.2 and 3.4.7 with respect to w_P and w_T. We obtain as first-order conditions

$$\partial_{w_P}\mathbb{E}_{F_{w_A}}(U_P) = \left[1 - F_{w_A}(\frac{w_P + w_T}{2})\right] - 0.5F'_{w_A}(\frac{w_P + w_T}{2})(w_P - w_F), \tag{3.4.9}$$

$$\partial_{w_P}\mathbb{E}_{F_{w_A}}(\Pi_T) = F_{w_A}(\frac{w_P + w_T}{2}) - 0.5F'_{w_A}(\frac{w_P + w_T}{2})(w_P - w_F), \tag{3.4.10}$$

Setting both conditions to zero for an optimal value, we first difference both to obtain

$$F_{w_A}(\frac{w_P + w_T}{2}) = 0.5 \tag{3.4.11}$$

such that at equilibrium, the average wage offer $M_{F_{w_A}} = \frac{w_P+w_T}{2}$ equals the median assessment of the arbitrator. Summing equations 3.4.9 and 3.4.10 gives

$$F'_{w_A}(\frac{w_P + w_T}{2})(w_P - w_F) = 1 \tag{3.4.12}$$

and substituting equation 3.4.11 provides us with the equilibrium wage-offers. □

Belief Formations. We now posit a model of belief-formation among arbitrators. We suppose the arbitrator randomly and independently samples with replacement the wages of players who are of the same quality as player P. We suppose that the pool of players is composed of a fraction $\rho \in (0, 1)$ with wages determined by arbitration and $(1 - \rho)$ with wages determined by market forces.[21]. Arbitrator A draws N wages from this pool and takes the average of this sample as her reference point w_A.[22]

Proposition 1 (Asymptotic Belief Distribution.)**.** *For sufficiently large N, the distribution of w_A, $F_{w_A}(\cdot)$ follows a normal distribution centered around the average draw μ and with variance σ^2 equal to such that*

$$w_A \sim N(\mu, \frac{\sigma^2}{N}) \tag{3.4.13}$$

Proof. Let Z_i be a draw from the pool of players. It is a weighted convolution between two random variables. The first is a draw from the pool of players eligible for arbitration. It is a bernouilli random variable which takes value w_P with probability 0.5 and w_T with probability 0.5, with mean $\frac{w_P+w_T}{2}$ and variance $\frac{(w_P-w_T)^2}{4}$. The second is a draw from the pool of players with wages determined by the market. We suppose this draw w_C^i follows a distribution with mean $E(w_C^i)$ and variance $V(w_C^i) < \infty$. Then, the mean of Z_i is given by

$$\mu = \rho \times \frac{w_P + w_T}{2} + (1 - \rho) \times E(w_C^i) \tag{3.4.14}$$

and the variance of Z_i by

$$\sigma^2 = \rho^2 \times \frac{(w_P - w_T)^2}{4} + (1 - \rho)^2 \times V(w_C^i) \tag{3.4.15}$$

We suppose that sample average of Z_i, denoted by $\bar{Z}_n$ determines the arbitrator's reference wage w_A such that $w_A = \bar{Z}_n$. Then, by the Lindeberg–Lévy Central Limit Theorem, as $N \to \infty$

$$\sqrt{N}(w_A - \mu) \xrightarrow{d} N(0, \sigma^2) \tag{3.4.16}$$

□

This setup seems plausible for several reasons. (a) Wages are publicly known and this allows the arbitrator to provide an objective argument to justify her assessment. (b) Players eligible for arbitration are of a sufficiently

[21] In practice, one can imagine the arbitrator discounts the wage-value of older players' of the same quality as player P but who are on the market wages to account for differences in experience and tenureship.

[22] This sense of *fair* assessment is also motivated by the way arbitrators are selected. Both the MLBPA and the league can propose an arbitrator and have veto power over the proposition of the other party. This suggests that arbitrators do not have any particular leniency for one party over the other. Unfortunately, arbitrators do not publish opinions, making it difficult to offer further evidence (Monhait, 2010).

homogeneous quality such that an arbitrator can find a sufficiently large amount of comparable players for the central limit theorem to be applicable. (c) Assuming that the arbitrator's beliefs are normally distributed is plausible in the sense that the normal distribution is symmetric around both its mean and median. It would be difficult to imagine why the distribution of beliefs would be asymmetric. (d) it is a common method of assessment in Corporate Finance to look at comparable assets in order to quantify its value (Vernimmen et al., 2018). Finally, (e) the normal distribution implies that the equilibrium defined in Equation 3.4.8 is a global equilibrium.[23] This suggests that the normal distribution does not imply undesirable features in this model.

Equilibrium Beliefs and Wages. We require consistent equilibrium beliefs between the value of a player P addressing her case to arbitrator A and the value assessed by arbitrator A of players with wages already determined by arbitration. By this, we mean that the expected value of the drawn reference wage w_A from the pool of players who were already eligible for arbitration must be equal to the expected wage from the overall pool of workers from which the arbitrator draws (i.e, $\mu = E(w_A)$). When this is the case, the wage set by the arbitrator is on average equal to the average market value of the player $E(w_C^i)$.

Theorem 1 (Equilibrium Wages)**.** *If the arbitrator has consistent equilibrium beliefs (i.e, $\mu = E(w_A)$), her expected reference wage w_A will equal the expected market wage $E(w_A) = E(w_C^i)$.*

Proof. Let arbitrator A have equilibrium beliefs which are consistent in the sense that her assessment for the value of player P must equal the average assessment for players of the same quality who have already undergone arbitration. Because w_A is normally distribution, its mean and median co-incide. So, we also have $E(w_A) = \frac{w_P + w_T}{2}$. In turn

$$E(w_A) = \mu = \rho \times E(w_A) + (1 - \rho) \times E(w_C^i) \tag{3.4.17}$$

which implies in non-degenerate cases (i.e, when $\rho < 1$) that

$$E(w_A) = E(w_C^i) \tag{3.4.18}$$

□

One can interpret the underlying mechanism as follows. Players and teams account for the uncertainty surrounding the arbitrator's reference point by, respectively, over and under-bidding compared to the median assessment for the arbitrator. However, this median assessment is selected to be on average consistent with the wages of players already eligible for arbitration. Recursively, only players with wages *exogeneously* determined by the market can provide a reference point for the arbitrator. As such, the average and median reference point must reflect this market value such that the change in wages observed when players become eligible to arbitration is akin to those of gaining access to free agents' labor market.

Conclusion

This research made the link between labor markets and antitrust law explicit. It argued that antitrust law protects the mobility of players which, in turn, is the source of workers' bargaining power. Barring this right, players suffer from depressed wages. To see this, we evaluated the impact of players gaining access to *Arbitration*, by which a third-part can set wages. Using the "Super Two" cut-off as an instrument randomly selecting players into the

[23] See example 1.E in Brams and Merrill (1983) for proof of this statement.

former, we find that players eligible to arbitration obtain extra-ordinary wage growth. To make this argument, we relied on a regression discontinuity design and supported our argument with several placebo and robustness tests.

We then argued that the post-*Arbitration* wage is akin to the one which could have resulted from the imperfectly competitive market for *Free Agents*. Indeed, we showed the absence of any kink in the growth rate of players' wages at Free Agency. Then, by revisiting the model of Final-Offer arbitration of Brams and Merrill (1983), we showed that this result can be expected when arbitrators use the wages of free agents to guide their assessment of players eligible for arbitration. In consequence, we conclude that that when players are unable to make team owners compete for their services, wages can be expected to fall between 32% to 48% below their market value.

In this regard, a note of caution for policy makers. As shown in Figure 3.A.19, which displays the quantile treatment effects developed by Frandsen et al. (2012), we observe that becoming eligible for *Arbitration* results in a drastic increase in the spread of the income distribution. That is, *Arbitration* can be shown to have increased the income of the highest paid players as well as inter-player income inequality. This suggests that removing anti-competitive clauses can be the source of significant growth in income inequality and that should be done whilst simultaneously developing regulatory forces and re-distributive instruments.

3.A Figures

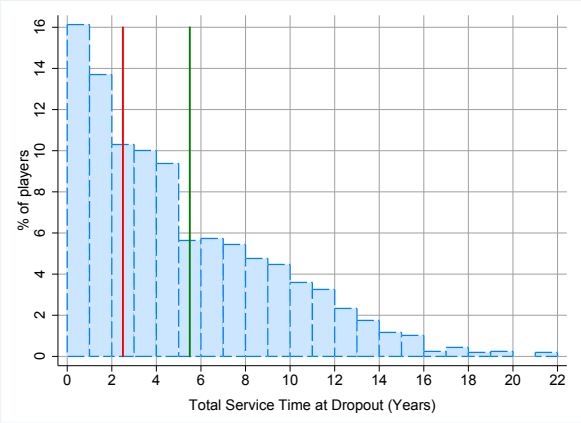


Figure 3.A.1: Selection in Major League Baseball Players (MLB)

This figure plots the event density. That is, the distribution of player drop-out from Major League Baseball (MLB) in relation to service time. Each observation is a single baseball player along with her final observed service time. This figure reveals two events of significant drop-out. The first in red, at three years of service time, players appear to be let go once *Arbitration* becomes mandatory. The second in green, at six years, when players can become *Free Agents*. Therefore, there appears to be significant non-random attrition among MLB players.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

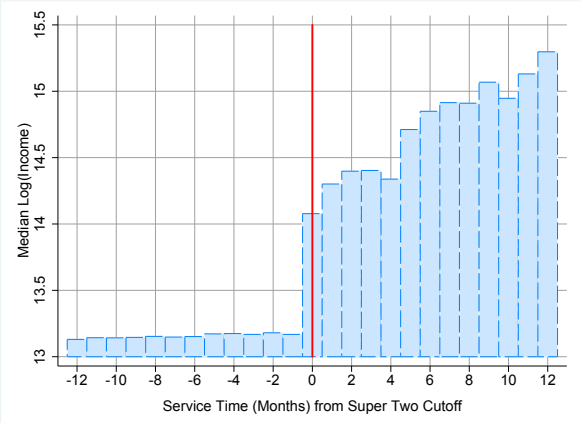


Figure 3.A.2: The *Super Two* cut-off and Major League Baseball Pay

This figure plots the relationship between Service Time (in months) and the median income of MLB players, around the *Super Two* cut-off. Each observation consists of a combination of a player and of a year, grouped in bins of 30 days (i.e., a month). Each bar represents the median income observed in one of these bins. This figure reveals a clear jump in the incomes of players above the *Super Two* cut-off compared to those lacking service time. Pay of players without *Arbitration* (lacking service time) is stable across service time but increasing in service time for player with *Arbitration*.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

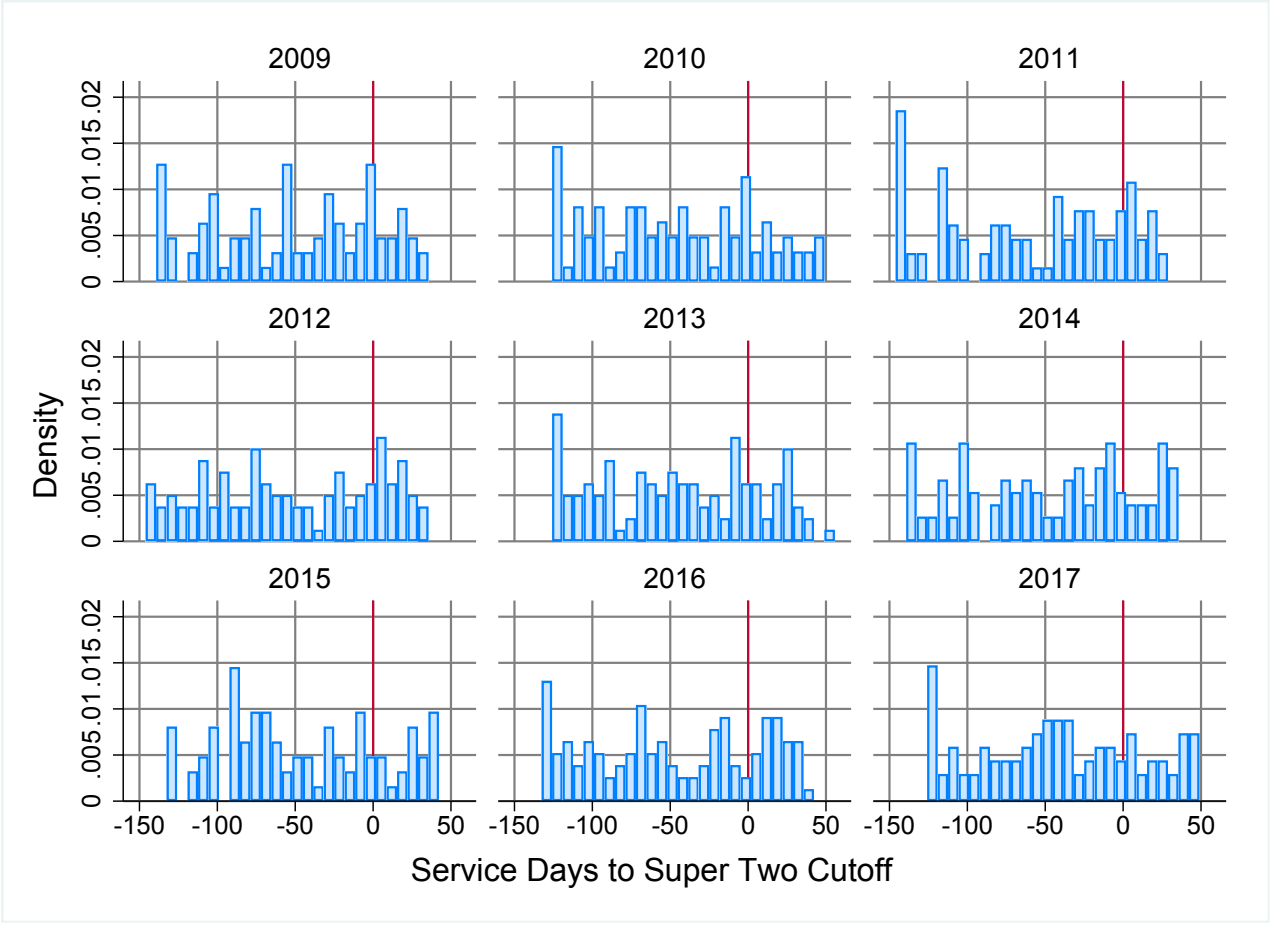


Figure 3.A.3: 2nd Year Distribution of Service-time around *Super Two* cut-off

This figure plots the distribution of players across Service Time (in months) for players with only two years of accrued service time, around the *Super Two* cut-off. Each subplot represents the distribution of players for different years of cut-offs.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

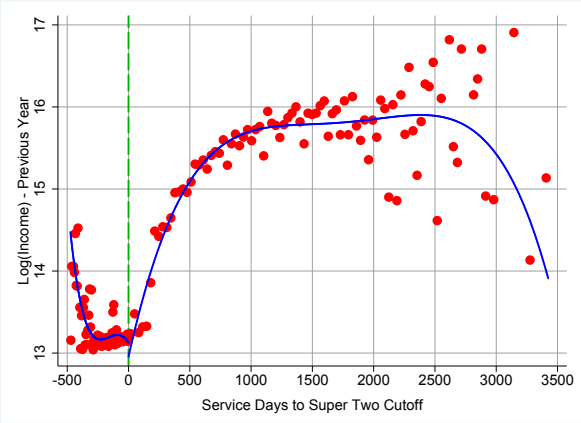


Figure 3.A.4: Falsification Test - Impact of *Super Two* cut-off on Previous Year Income

This figure plots a falsification test. The full sample (a panel of players across years) is used. It shows the relationship between the accrued service time and player's past income (in natural logarithms) around the *Super Two* cut-off point. As described in Calonico et al. (2015), each red dot (or bin) is the sample log-income at a given service time. The blue line approximates the population conditional expectation functions, above and below the cut-off, by a kernel-weighted polynomial function of order 4 (standard order) using the Epanechnikov kernel. This figure reveals no discontinuity, in previous year log-income at the cut-off.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

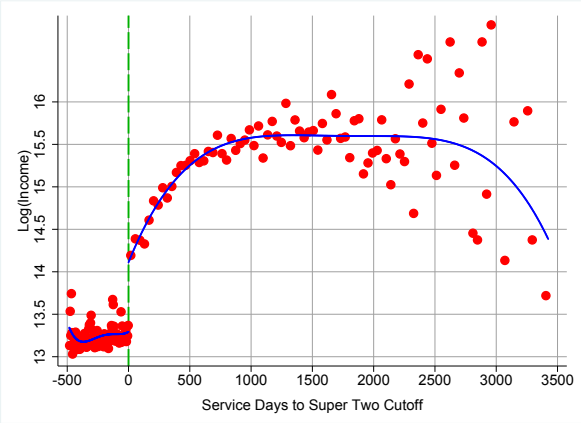


Figure 3.A.5: Main Specification of the Regression Discontinuity Design

This figure plots the main specification of this paper. The full sample (a panel of players across years) is used. It shows the relationship between the accrued service time and player's income (in natural logarithms) around the *Super Two* cut-off point. As described in Calonico et al. (2015), each red dot (or bin) is the sample log-income at a given service time. The blue line approximates the population conditional expectation functions, above and below the cut-off, by a polynomial function of order 4 (standard order). This figure reveals a large discontinuity, in log-income at the cut-off. For clarification, the extreme values around -500 days of service time are outliers but do not reflect an error within the data. These points reflect the existence of highly paid rookies starting in 2014. Their exclusion does not change the main point-estimate of this paper.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

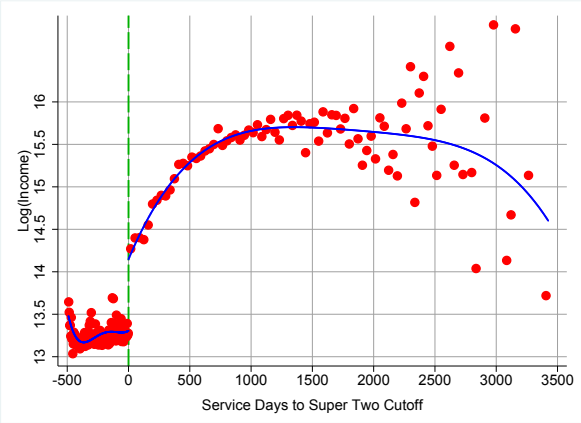


Figure 3.A.6: Main Specification using *Cot's Baseball Contracts* Database

This figure plots the main specification of this paper using an alternative dataset, the Cot's Contract Database. The full sample (a panel of players across years) is used. It shows the relationship between the accrued service time and player's income (in natural logarithms) around the *Super Two* cut-off point. As described in Calonico et al. (2015), each red dot (or bin) is the sample log-income at a given service time. The blue line approximates the population conditional expectation functions, above and below the cut-off, by a polynomial function of order 4 (standard order). This figure also reveals a large discontinuity, in log-income at the cut-off.
Source: Cot's Baseball Contracts, MLBTradeRumors.com, and author's calculations.

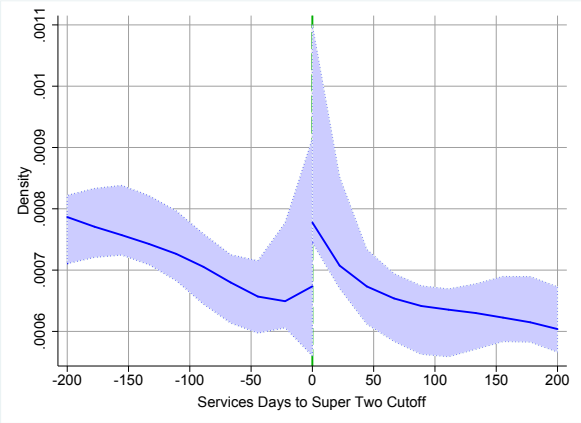


Figure 3.A.7: Manipulation Test - Continuity at cut-off

This figure plots the McCrary (2008) Manipulation test as implemented by Cattaneo et al. (2019). The full sample (a panel of players across years) is used. It shows the density in service time of players around the *Super Two* cut-off point. The blue line approximates the density function, above and below the cut-off, by a kernel-weighted polynomial function of order 2 using the Epanechnikov kernel. This figure reveals no discontinuity in the density and so, no sign of manipulation of players at the cut-off.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

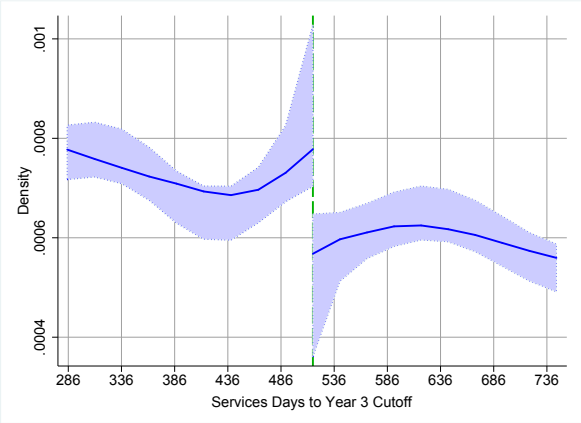


Figure 3.A.8: Manipulation Test - Continuity at Year 3 cut-off (516 days)

This figure plots the McCrary (2008) Manipulation test as implemented by Cattaneo et al. (2019). The full sample (a panel of players across years) is used. It shows the density in service time of players around the Year 3 cut-off point of 516 days. The blue line approximates the density function, above and below the cut-off, by a kernel-weighted polynomial function of order 2 using the Epanechnikov kernel. This figure reveals signs of discontinuity and so, of signs of manipulation from the team owners.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

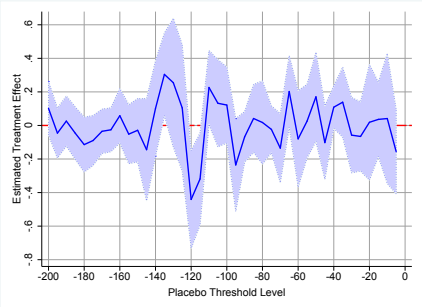


Figure 3.A.9: Below the cut-off

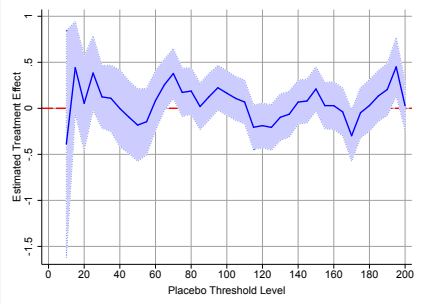


Figure 3.A.10: Above the cut-off

Figure 3.A.11: Placebo test - False cut-offs

This figure plots two sets of placebo tests based on false cut-offs. Figure 3.A.9 shows the point-estimate along with the 95% confidence intervals from running the main specification of this paper at the different service time days (with an interval of five days starting at -200 days) excluding values above zero. Figure 3.A.10 does the same, showing the (so called bias-robust) point-estimate along with their non-robust 95% confidence intervals from running the main specification of this paper at the different service time days (with an interval of five days starting at 20 days) , but excluding values below zero. More precisely, it shows the result from running the *Local Polynomial Regression Discontinuity Estimation with Robust Bias-Corrected Confidence Intervals* procedure developed in Calonico et al. (2014), Calonico et al. (2019) and Calonico et al. (2015) at the different cut-off days. Essentially, this method estimates a kernel-weighted (i.e, to limit the importance of values far away from the cut-off) polynomial (to have a flexible function form the conditional expectation) above and below the cut-off and tests for the existence of a statistically significant difference at this cut-off, using heteroskedastic-robust standard errors. The inherent bias of the conditional expectation polynomial of order (p) is corrected by estimating a polynomial of higher degree (q) and using its derivative. The Epanechnikov kernel is used along with the order two polynomial. These figures reveal that, taking into account that multiple testing will naturally lead to detecting some tests as indicating statistically significance, there is no evidence of discontinuous responses of income to service time other than at the *Super Two* cut-off.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

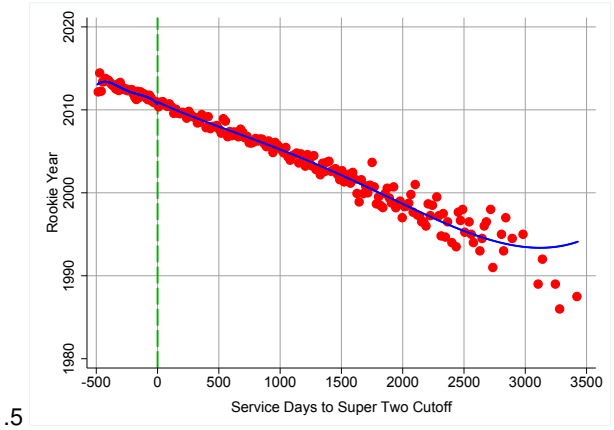


Figure 3.A.12: Rookie Year

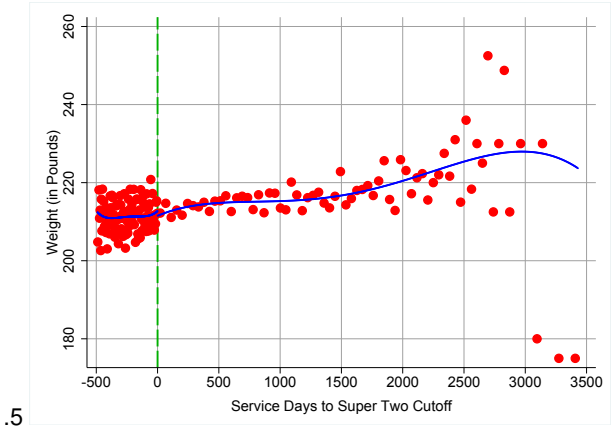


Figure 3.A.13: Weight

Figure 3.A.14: Robustness Check - Irrelevant Variables

These figures plot robustness tests. The full sample (a panel of players across years) is used. It shows the relationship between the accrued service time and player's rookie year (Figure 3.A.12), and with their weight (in Figure 3.A.13). As described in Calonico et al. (2015), each red dot (or bin) is the sample mean of the outcome at a given service time. The blue line approximates the population conditional expectation functions, above and below the cut-off, by a kernel-weighted polynomial function of order 4 (standard order) using the Epanechnikov kernel. Figure 3.A.12 and Figure 3.A.13 reveal no evidence of a discontinuity, suggesting an absence of unexpected result.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

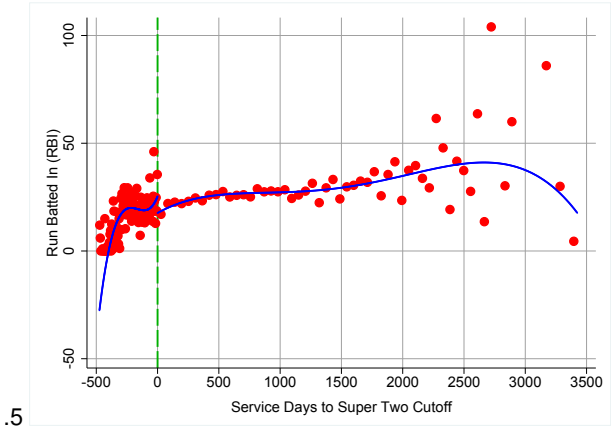


Figure 3.A.15: Run Batted In

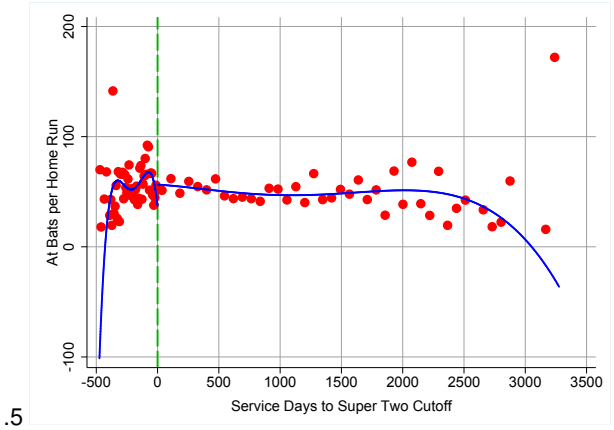


Figure 3.A.16: At Bats per Home Run

Figure 3.A.17: Robustness Check - Relevant Variables

These figures plot robustness tests. The full sample (a panel of players across years) is used. It shows the relationship between the accrued service time and player's Run Batted In (RBI, in Figure 3.A.15), and with the At Bats per Home Run (ABHR, in Figure 3.A.16). As described in Calonico et al. (2015), each red dot (or bin) is the sample mean of the sabermetric at a given service time. The blue line approximates the population conditional expectation functions, above and below the cut-off, by a kernel-weighted polynomial function of order 4 (standard order) using the Epanechnikov kernel. Figure 3.A.15 reveals a potential discontinuity at the *Super Two* cut-off whereas Figure 3.A.16 does not.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

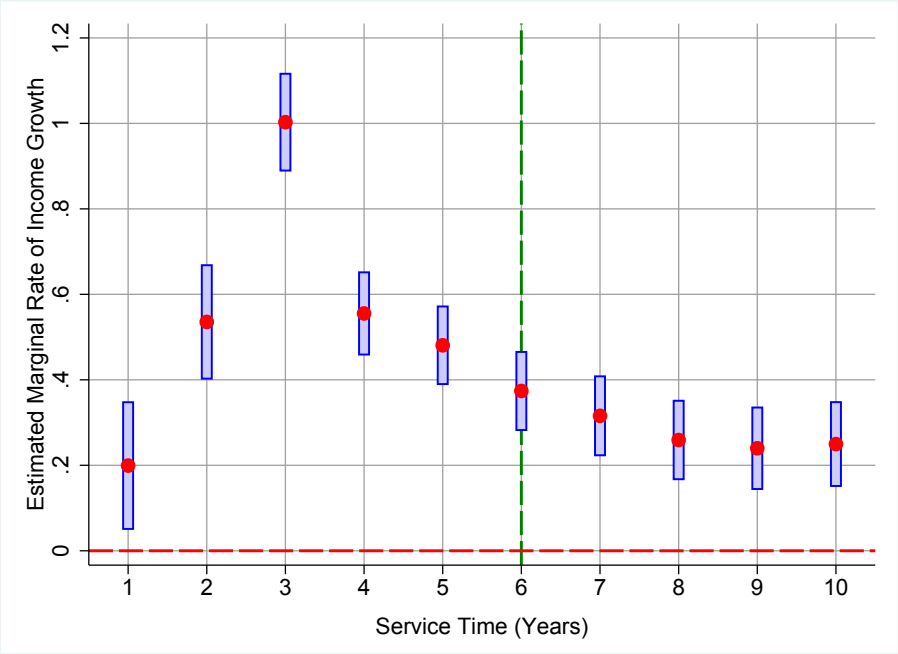


Figure 3.A.18: Marginal Years of Increase Income by Service Time

This figure shows the coefficients estimated in Table 3.B.10, column (4). This table reports the results from Exponential Regressions Wage-regression using the full sample and heteroskedasticity robust standard errors. A set of variables for each year of service time is estimated. Contrary to usual fixed-effects, the variable of year of service number x equals to one if the player has accrued at least x years of service. The interpretation of the estimates of year 6 in the specification of column 4 is that, keeping fixed the year, the age, the team and the player, an increase of service time from year 5 to year 6 increases income by 34%.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

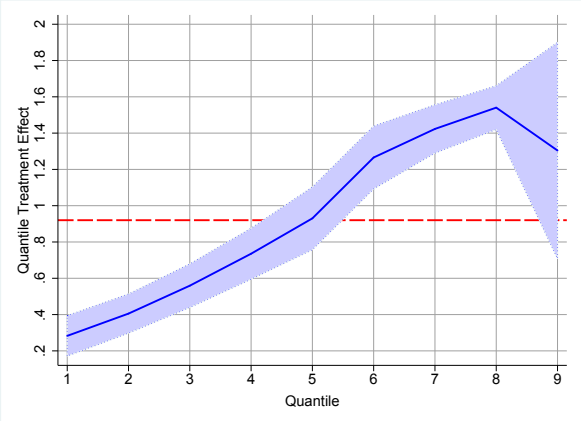


Figure 3.A.19: Quantile Treatment Effects

This figure shows the coefficients estimated in Table 3.B.11. This table reports the results from running the *Quantile Treatment Effects in Regression Discontinuity* model of Frandsen et al. (2012). This method is akin to the one developed by Calonico (2017) but focuses on quantiles of the dependant variable. It compares the θ-th quantile of the dependant variable above and below the cut-off. This quantile is estimated using a local-linear kernel weighted approximation (i.e, polynomial of degree one). This model was estimated on the full sample using the log-income as a dependant variable and the Epanechnikov kernel. The bandwidth is based on the optimal mean squared error procedure and was simply recovered from the degree one local kernel-weighted polynomial approximation provided in column (1) from 3.B.4. The Average Treatment Effect (ATE) from the latter is plotted in red, as a reference. This figure reveals that *Arbitration* has a positive effect over the whole distribution but with disproportionately large effects at the top of the income distribution.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

3.B Tables

Table 3.B.1: Descriptive Statistics

	All Players			***Pre-Salary Arbitration***			***Post-Salary Arbitration***		
	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs	Mean	Std.Dev.	Obs
Player History	rf.			rf.			rf.		
Rookie Year	2009.01	4.59	7603	2012.37	2.68	2955	2006.87	4.25	4648
Weight (in Pounds)	213.53	21.00	7407	211.37	20.85	2882	214.90	20.98	4525
Income (USD)	3980401.24	5340517.19	7582	697741.68	1186805.85	2953	6074523.88	5878704.89	4629
Batting Statistics	rf.			rf.			rf.		
At Bats per Home Run (ABHR)	51.67	54.55	3335	54.18	55.19	1190	50.28	54.16	2145
Run Batted In (RBI)	21.08	29.34	7015	17.65	26.04	2698	23.23	31.04	4317
Treatment and Outcome	rf.			rf.			rf.		
Days to Super Two Cutoff	319.34	636.54	7603	–252.78	135.33	2955	683.07	557.43	4648
Within +/- 10 days of Cutoff	0.02	0.13	7603	0.02	0.14	2955	0.02	0.12	4648

This table presents the main descriptive statistics of the data. The data is represented in two ways: all players (column 1) and according to treatment status (column 2 for players below the cut-off lacking *Arbitration* and column 3 for those above the cut-off and subject to *Arbitration*). The main variables of interest describing the players are the *Rookie Year* (broadly speaking, the first year of participation into in Major League Baseball), the player's *Weight*, and his *Income* measured in nominal American dollars (this measure does not include potential *ad hoc* bonuses). These players are characterized by their playing skill in terms of their At Bats per Home Run (ABHR), which is a common sabermetric measuring the number of times a player had the opportunity to hit (At Bat) divided by the number of Home Runs, and the Run Batted In (RBI), which measures the number of times a batter makes a play allowing a run to be scored. Two variables of interest characterise the quasi-experimental setup. The variable *Days to Super Two cut-off* counts the service time (in days) to the *Super Two* cut-off. A negative value signifies that the player lacks service time and a positive value indicates that the player has gone beyond the necessary service time. The zero value implies that the player has sufficient service time. The last line of the table reports the statistics for a dummy variable equal to one if the player is within 10 days (above or below) of the cut-off. This line reports that 2% of players of within this gap (i.e, 126 players). Among those below the threshold, 2% of players (i.e, 55 players) are within the ten day bandwidth. Among those above the threshold, 2% of players (i.e, 71 players) are within the ten day bandwidth.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.2: Yearly *Super Two* cut-off and Sample

Arbitration Year	*Super Two* Year	cut-off	Nb. Observations	%
2010	2009	2.139	833	11.0
2011	2010	2.122	809	10.6
2012	2011	2.146	851	11.2
2013	2012	2.140	871	11.5
2014	2013	2.122	887	11.7
2015	2014	2.133	869	11.4
2016	2015	2.130	853	11.2
2017	2016	2.131	851	11.2
2018	2017	2.123	779	10.2

The *Super Two* cut-off is provided in the standard form used within the literature. It is to be interpreted as YY.DDD where Y stands for Year, and D for Day. Players that were selected into *Arbitration* by the 2009 *Super Two* cut-off, at 2 years and 139 days of service time, received their new income in 2010. 11% of the sample are wage observations in the year 2010.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.3: Share of Sample in Year 2 above *Super Two* cut-off

Super Two Year	Share above cut-off (%)
2009	22.58
2010	24.18
2011	19.79
2012	25.42
2013	22.88
2014	23.42
2015	25.00
2016	27.19
2017	24.75

This table displays the share of players above the *Super Two* cutoff date, per year. We observe that broadly speaking, the share of players above the cut-off is, on average, in line with the expected value (22%). This suggests the sample correctly covers the underlying population.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.4: Main Specification

	(1) Log(Income)	(2) Log(Income)	(3) Log(Income)	(4) Log(Income)	(5) Log(Income)	(6) Log(Income)
Conventional	0.903***	0.834***	0.582***	0.501***	0.544***	0.553***
	[0.061]	[0.076]	[0.108]	[0.112]	[0.126]	[0.134]
Bias-corrected	0.921***	0.810***	0.577***	0.469***	0.572***	0.580***
	[0.061]	[0.076]	[0.108]	[0.112]	[0.126]	[0.134]
Robust	0.921***	0.810***	0.577***	0.469***	0.572***	0.580***
	[0.072]	[0.083]	[0.119]	[0.117]	[0.132]	[0.140]
Robust 95% CI	[.780 ; 1.061]	[.647 ; .972]	[.344 ; .809]	[.240 ; .699]	[.313 ; .831]	[.306 ; .854]
Kernel Type	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov
BW Type						
Observations	7582	7582	7582	7582	7582	7582
Conventional p-value	0.000	0.000	0.000	0.000	0.000	0.000
Robust p-value	0.000	0.000	0.000	0.000	0.000	0.000
Order Loc. Poly. (p)	1.000	2.000	3.000	4.000	5.000	6.000
Order Bias (q)	2.000	3.000	4.000	5.000	6.000	7.000

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

This table reports the main specification of this paper. It shows the result from running the *Local Polynomial Regression Discontinuity Estimation with Robust Bias-Corrected Confidence Intervals* procedure developed in Calonico et al. (2014), Calonico et al. (2019) and Calonico et al. (2015). Essentially, this method estimates a kernel-weighted (i.e, to limit the importance of values far away from the cut-off) polynomial (to have a flexible function form the conditional expectation) above and below the cut-off and tests for the existence of a statistically significant difference at this cut-off, using heteroskedastic-robust standard errors. The inherent bias of the conditional expectation polynomial of order (p) is corrected by estimating a polynomial of higher degree (q) and using its derivative. The *Conventional* estimator does not correct for the bias induced by the use of kernel functions, the *Bias-Corrected* does but relies on the same standard errors as the *Conventional* estimator, and the *Robust* estimator uses both bias-corrected estimates and robust standard errors. The Epanechnikov kernel is used along with the different polynomial orders (p=1 to p=6 in column 1 to column 6). This table reveals that receiving *Arbitration* increases income, on average, by 92% (column 1).
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.5: Main Specification using Cot's Contract Database

	(1) Log(Income)	(2) Log(Income)	(3) Log(Income)	(4) Log(Income)	(5) Log(Income)	(6) Log(Income)
Conventional	0.946***	0.754***	0.703***	0.623***	0.677***	0.704***
	[0.063]	[0.092]	[0.103]	[0.105]	[0.117]	[0.124]
Bias-corrected	0.974***	0.720***	0.692***	0.606***	0.705***	0.723***
	[0.063]	[0.092]	[0.103]	[0.105]	[0.117]	[0.124]
Robust	0.974***	0.720***	0.692***	0.606***	0.705***	0.723***
	[0.069]	[0.096]	[0.112]	[0.112]	[0.122]	[0.129]
Robust 95% CI	[.839 ; 1.109]	[.531 ; .909]	[.471 ; .912]	[.386 ; .827]	[.466 ; .944]	[.47 ; .976]
Kernel Type	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov
Observations	8690	8690	8690	8690	8690	8690
Conventional p-value	0.000	0.000	0.000	0.000	0.000	0.000
Robust p-value	0.000	0.000	0.000	0.000	0.000	0.000
Order Loc. Poly. (p)	1.000	2.000	3.000	4.000	5.000	6.000
Order Bias (q)	2.000	3.000	4.000	5.000	6.000	7.000

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

This table reports the main specification of this paper using the Cot's Contract Database as an alternative dataset. It shows the result from running the *Local Polynomial Regression Discontinuity Estimation with Robust Bias-Corrected Confidence Intervals* procedure developed in Calonico et al. (2014), Calonico et al. (2019) and Calonico et al. (2015). Essentially, this method estimates a kernel-weighted (i.e, to limit the importance of values far away from the cut-off) polynomial (to have a flexible function form the conditional expectation) above and below the cut-off and tests for the existence of a statistically significant difference at this cut-off, using heteroskedastic-robust standard errors. The inherent bias of the conditional expectation polynomial of order (p) is corrected by estimating a polynomial of higher degree (q) and using its derivative. The *Conventional* estimator does not correct for the bias induced by the use of kernel functions, the *Bias-Corrected* does but relies on the same standard errors as the *Conventional* estimator, and the *Robust* estimator uses both bias-corrected estimates and robust standard errors. The Epanechnikov kernel is used along with the different polynomial orders (p=1 to p=6 in column 1 to column 6). This table reveals that receiving *Arbitration* increases income, on average, by 97% (column 1).
Source: Cot's Contract Database, MLBTradeRumors.com, and author's calculations.

Table 3.B.6: Falsification Test - Impact of *Arbitration* on Previous Year Income

	(1) Previous Log(Income)	(2) Previous Log(Income)	(3) Previous Log(Income)	(4) Previous Log(Income)	(5) Previous Log(Income)	(6) Previous Log(Income)
Conventional	–0.079	–0.128	–0.132	–0.113	–0.013	–0.005
	[0.058]	[0.072]	[0.076]	[0.083]	[0.095]	[0.096]
Bias-corrected	–0.113	–0.152*	–0.124	–0.089	0.008	0.018
	[0.058]	[0.072]	[0.076]	[0.083]	[0.095]	[0.096]
Robust	–0.113	–0.152	–0.124	–0.089	0.008	0.018
	[0.064]	[0.079]	[0.084]	[0.087]	[0.101]	[0.099]
Robust 95% CI	[-.239 ; .012]	[-.306 ; .002]	[-.288 ; .04]	[-.259 ; .08]	[-.189 ; .205]	[-.177 ; .212]
Kernel Type	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov
Observations	5542	5542	5542	5542	5542	5542
Conventional p-value	0.177	0.075	0.081	0.175	0.893	0.957
Robust p-value	0.077	0.053	0.139	0.301	0.936	0.859
Order Loc. Poly. (p)	1.000	2.000	3.000	4.000	5.000	6.000
Order Bias (q)	2.000	3.000	4.000	5.000	6.000	7.000

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

This table reports a falsification test based on replacing the current income with the previous year's income..It shows the result from running the *Local Polynomial Regression Discontinuity Estimation with Robust Bias-Corrected Confidence Intervals* procedure developed in Calonico et al. (2014), Calonico et al. (2019) and Calonico et al. (2015). Essentially, this method estimates a kernel-weighted (i.e, to limit the importance of values far away from the cut-off) polynomial (to have a flexible function form the conditional expectation) above and below the cut-off and tests for the existence of a statistically significant difference at this cut-off, using heteroskedastic-robust standard errors. The inherent bias of the conditional expectation polynomial of order (p) is corrected by estimating a polynomial of higher degree (q) and using its derivative. The *Conventional* estimator does not correct for the bias induced by the use of kernel functions, the *Bias-Corrected* does but relies on the same standard errors as the *Conventional* estimator, and the *Robust* estimator uses both bias-corrected estimates and robust standard errors. The Epanechnikov kernel is used along with the standard order one polynomial. This table reveals that we can reject the hypothesis according to which receiving *Arbitration* increases the previous year's income given that in all specifications (p=1 to p=6 in column 1 to 6), there is only one test which comes out as statistically significant at 5% (which is normal in cases of multiple testing) and that this test relies on conventional standard errors.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.7: Manipulation Test - Continuity of the Service Time Density near cut-off

	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic
Conventional	0.34	1.27	1.15	2.02*	0.46	0.57
Robust	0.53	1.44	0.89	1.20	1.05	0.83
Kernel Type	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov
Observations	7603	7603	7603	7603	7603	7603
Conventional p-value	0.73	0.20	0.25	0.04	0.64	0.56
Robust p-value	0.60	0.30	0.15	0.37	0.22	0.41
Order Loc. Poly. (p)	1	2	3	4	5	6
Order Bias (q)	2	3	4	5	6	7

$^{*}\ p < 0.05, ^{**}\ p < 0.01, ^{***}\ p < 0.001$

This table reports the test results from the McCrary (2008) Manipulation test as implemented by Cattaneo et al. (2019) . The full sample (a panel of players across years) is used. This test evaluates the continuity in the density function of players in terms of service time around the *Super Two* cut-off. It is based on approximating the density function using kernel-weighted polynomials around the cut-off. The *robust* estimator is bias corrected and uses robust standard errors. The Epanechnikov kernel is used along with an array of different polynomials (p=1 to p=6, in column 1 to 6). This table reveals that we can reject the hypothesis according to which the density is discontinuous at the cut-off, suggesting the absence of manipulation from the team owners of players' service time.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.8: Manipulation Test - Continuity of the Service Time Density near Year 3

	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic	McCrary t-statistic
Conventional	−2.70**	−2.30*	−2.48*	−3.03**	−1.28	−3.06**
Robust	−1.65	−2.48**	−2.72**	−1.11	−0.68	−0.60
Kernel Type	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov
Observations	7603	7603	7603	7603	7603	7603
Conventional p-value	0.0070	0.0216	0.0133	0.0025	0.2011	0.0022
Robust p-value	0.0997	0.0131	0.0065	0.2698	0.4989	0.5973
Order Loc. Poly. (p)	1	2	3	4	5	6
Order Bias (q)	2	3	4	5	6	7

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This table reports the test results from the McCrary (2008) Manipulation test as implemented by Cattaneo et al. (2019) . The full sample (a panel of players across years) is used. This test evaluates the continuity in the density function of players in terms of service time around the Year 3 cut-off (516 days of service time). This is the point at which all players become eligible for *Arbitration*. It is based on approximating the density function using kernel-weighted polynomials around the cut-off. The *robust* estimator is bias corrected and uses robust standard errors. The Epanechnikov kernel is used along with an array of different polynomials (p=1 to p=6, in column 1 to 6). This table reveals that we can reject the hypothesis according to which the density is continuous at the cut-off, if the density function is sufficiently well approximated by local polynomials of degree three. This suggests manipulation, from the team owners, of players' service time.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.9: Robustness Test - Pre-determined and Irrelevant Variables

	(1) Rookie Year	(2) Weight (Pounds)	(3) Run Batted In (RBI)	(4) At Bats per Home Run (ABHR)
Conventional	–0.571	–1.951	–9.074	5.170
	[0.486]	[3.757]	[6.100]	[18.606]
Bias-corrected	–0.652	–2.395	–9.862	0.786
	[0.486]	[3.757]	[6.100]	[18.606]
Robust	–0.652	–2.395	–9.862	0.786
	[0.521]	[4.080]	[6.649]	[19.745]
Robust 95% CI	[-1.674 ; .37]	[-10.392 ; 5.603]	[-22.893 ; 3.169]	[-37.913 ; 39.486]
Kernel Type	Epanechnikov	Epanechnikov	Epanechnikov	Epanechnikov
BW Type				
Observations	7603	7407	5228	2625
Conventional p-value	0.239	0.604	0.137	0.781
Robust p-value	0.211	0.557	0.138	0.968
Order Loc. Poly. (p)	4.000	4.000	4.000	4.000
Order Bias (q)	5.000	5.000	5.000	5.000

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$, $^{***}\ p < 0.001$

This table reports robustness tests using irrelevant and pre-determined variables as outcomes of the main specification used in this paper. It shows the result from running the *Local Polynomial Regression Discontinuity Estimation with Robust Bias-Corrected Confidence Intervals* procedure developed in Calonico et al. (2014), Calonico et al. (2019) and Calonico et al. (2015). Essentially, this method estimates a kernel-weighted (i.e, to limit the importance of values far away from the cut-off) polynomial (to have a flexible function form the conditional expectation) above and below the cut-off and tests for the existence of a statistically significant difference at this cut-off, using heteroskedastic-robust standard errors. The inherent bias of the conditional expectation polynomial of order (p) is corrected by estimating a polynomial of higher degree (q) and using its derivative. In the four cases reported in this table, the Epanechnikov kernel is used along with the standard order one polynomial. Four different outcomes are tested : the player's (1) rookie year, (2) weight, (3) Runs Batted In, and (4) At Bats per Home Run (see main text for further explanations of these sabermetrics). This table reveals a certain robustness of the methodology because it does not detect an effect in variables which are either predetermined (column 1) or irrelevant (column 2). However, there is evidence of a sudden fall in the Run Batted In of players who cross the threshold (column 3). This is problematic if team owners are capable of avoiding *Arbitration* for players who are particularly good and which, presumably, would become much more expensive after the re-evaluation of their wage. This fear in partially mitigated by column (4) which reveals a lack of discontinuity in the At Bats per Home Run (ABHR) value of players selected into *Arbitration*.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.10: Exponential Wage Regression with Service Years Fixed Effects

		(1)	(2)	(3)	(4)
Years of Service (Cascading Fixed Effect)	Income	Income	Income	Income	
1	0.00626	0.0277	0.0360	0.199***	
	(0.0684)	(0.0721)	(0.0718)	(0.0756)	
2	0.428***	0.469***	0.477***	0.536***	
	(0.0682)	(0.0701)	(0.0699)	(0.0677)	
3	0.882***	0.939***	0.937***	1.003***	
	(0.0559)	(0.0574)	(0.0575)	(0.0578)	
4	0.461***	0.545***	0.530***	0.555***	
	(0.0398)	(0.0393)	(0.0397)	(0.0491)	
5	0.355***	0.448***	0.438***	0.481***	
	(0.0403)	(0.0388)	(0.0392)	(0.0463)	
6	0.215***	0.330***	0.309***	0.374***	
	(0.0461)	(0.0434)	(0.0422)	(0.0466)	
7	0.0980*	0.224***	0.220***	0.316***	
	(0.0520)	(0.0471)	(0.0443)	(0.0471)	
8	0.0977*	0.224***	0.211***	0.259***	
	(0.0548)	(0.0497)	(0.0473)	(0.0468)	
9	0.0607	0.187***	0.175***	0.240***	
	(0.0610)	(0.0558)	(0.0533)	(0.0486)	
10	0.0859	0.206***	0.196***	0.250***	
	(0.0679)	(0.0624)	(0.0587)	(0.0501)	
11	0.00438	0.0976	0.0894	0.189***	
	(0.0789)	(0.0718)	(0.0675)	(0.0523)	
12	0.0901	0.226***	0.200***	0.243***	
	(0.0894)	(0.0796)	(0.0766)	(0.0570)	
13	-0.102	0.0650	0.0298	0.281***	
	(0.108)	(0.0967)	(0.0958)	(0.0627)	
14	-0.0790	0.0296	0.0553	0.154**	
	(0.142)	(0.132)	(0.123)	(0.0759)	
15	-0.104	0.0905	0.0421	0.161*	
	(0.198)	(0.175)	(0.156)	(0.0940)	
16	0.408*	0.564***	0.449***	0.461***	
	(0.237)	(0.181)	(0.165)	(0.101)	
17	-0.171	0.0303	0.0574	0.214*	
	(0.260)	(0.184)	(0.177)	(0.127)	
18	0.0913	0.113	0.146	0.469***	
	(0.290)	(0.242)	(0.222)	(0.134)	
19	-0.757	-0.755	-0.824	-0.550	
	(0.543)	(0.531)	(0.549)	(0.599)	
20	0.848	0.967*	0.858	0.707	
	(0.724)	(0.550)	(0.560)	(0.608)	
21	-0.303	-0.132	-0.0806	0.553***	
	(0.661)	(0.550)	(0.452)	(0.141)	
22	-2.139***	-1.394**	-1.111*	-0.00416	
	(0.425)	(0.701)	(0.674)	(0.718)	
Year FE	Yes	Yes	Yes	Yes	
Age FE	No	Yes	Yes	Yes	
Team FE	No	No	Yes	Yes	
Player FE	No	No	No	Yes	
Observations	7582	7579	7579	7057	

Robust Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This table reports various specifications of an Exponential Regression Wage-regression using the full sample and heteroskedasticity robust standard errors, as advocated by Bellégo et al. (2021).. A set of variables for each year of service time is estimated. Contrary to usual fixed-effects, the variable of year of service number x equals to one if the player has accrued at least x years of service. In the first column, only year fixed effects are used. The second adds age fixed effects, whilst the third also has team fixed effects. The final column includes player fixed effects. The latter requires a connected set of players of moving across teams and the removal of singletons (player with a single observation), explaining the fall in the number of observations. The interpretation of the estimates of year 6 in the specification of column 4 is that, keeping fixed the year, the age, the team and the player, an increase of service time from year 5 to year 6 increases income by 34%. This figure reveals that income does not rise particularly significantly when players become *Free Agents* (year 6). To the contrary, income rises particularly strongly when players reach salary arbitration in year 3.

Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Table 3.B.11: Quantile Treatment Effects

	Log(Income)
Quantile 1	0.283***
	(0.0568)
Quantile 2	0.405***
	(0.0548)
Quantile 3	0.559***
	(0.0613)
Quantile 4	0.735***
	(0.0716)
Quantile 5	0.930***
	(0.0881)
Quantile 6	1.266***
	(0.0883)
Quantile 7	1.423***
	(0.0678)
Quantile 8	1.540***
	(0.0613)
Quantile 9	1.304***
	(0.304)
Observations	7603
Kernel	Epanechnikov
Bandwidth	222.8

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

This table reports the results from running the *Quantile Treatment Effects in Regression Discontinuity* developed by Frandsen et al. (2012). This method is akin to the one developed by Calonico (2017) but focuses on quantiles of the dependant variable. It compares the θ-th quantile of the dependant variable above and below the cut-off. This quantile is estimated using a local-linear kernel weighted approximation (i.e, polynomial of degree one). This model was estimated on the full sample using the log-income as a dependant variable and the Epanechnikov kernel. The bandwidth is based on the optimal mean squared error procedure and was simply recovered from the degree one local kernel-weighted polynomial approximation provided in column (1) from Table 3.B.4. The results are plotted in Figure 3.A.19. This table reveals that *Arbitration* has a positive effect over the whole distribution but with disproportionately large effects at the top of the income distribution.
Source: Lahman's Baseball Database, MLBTradeRumors.com, Baseball-Reference.com, and author's calculations.

Chapter 4

Price Discrimination and Big Data: Evidence from a Mobile Puzzle Game

This chapter is based on work with Christian Helmers, Alessandro Iaria, Stefan Wagner, and Julian Runge.

Abstract: We use data from a mobile puzzle game to investigate the welfare consequences of price discrimination. We rely on experimental variation to characterize player behavior and estimate a model of demand for game content. Our counterfactual simulations show that the game developer's observed pricing is far from optimal. Profit would increase by 340% if the game developer used optimal uniform pricing instead. What is more important, our results suggest that optimal uniform pricing results in almost the same increase in profits as first-degree price discrimination (347%). All pricing strategies considered—including optimal uniform pricing—induce a transfer of surplus from players to game developer without, however, generating sizeable dead-weight losses.

4.1 Introduction

Price discrimination is ubiquitous in offline markets (Varian, 1989). As is well known, charging different prices to different consumers according to their willingness-to-pay often enables companies to increase profits (Tirole, 1988).

Despite its promises, price discrimination has however not been implemented on a large scale in the monetization of digital products, where "examples remain fairly limited" (Council of Economic Advisors, 2015). This is surprising because price discrimination could be implemented at comparably lower costs for digital products. By the nature of the underlying technology, companies can easily collect large amounts of detailed data about consumer characteristics and behavior (Goldfarb and Tucker, 2019). The resulting (big) data could be used to determine personalized prices for consumers with different characteristics and search or consumption histories, that is, engage in price discrimination, to increase a company's profit. On the one hand, firms reported to fear consumer backlash and negative press coverage when engaging in price discrimination (Garbarino and Maxwell, 2010; Li and Jain, 2016; DellaVigna and Gentzkow, 2019). An example is Amazon's early attempt to price discriminate buyers of DVDs based on their individual purchase history, which met dramatic resistance and negative publicity (Rosencrance, 2019). Similarly, the Wall Street Journal revealed

in 2012 that Staples was charging consumers online different prices based on their location. On the other hand, there is a growing body of evidence to suggest that, in many important industries, because of path-dependence, imperfect information, learning, or conflicting incentives, sometimes for-profit firms do not maximize profit (Cho and Rust, 2010; DellaVigna and Gentzkow, 2019; Fioretti, 2020; Hortaçsu et al., 2021; Huang et al., 2020; Orbach and Einav, 2007) and, indeed, Dube and Misra (2019) estimate that, in the context of a digital recruiting firm, both profit *and* consumer surplus would often increase with personalized pricing.

In this paper, we contribute to this debate by investigating the welfare consequences of price discrimination for a mobile gaming app. Gaming accounts for over three quarters of total app revenue in the major app stores, including Apple's App Store and Google's Playstore (TechCrunch, June 11, 2019). Mobile games attracted over 150 million users worldwide and generated almost US$100 billion in revenue in 2020 (Statista, 2021). While game developers in this market collect around 90% of their revenues through paying customers on the basis of freemium models[1] (advertising accounts only for a small share of revenue), the efficiency of alternative pricing strategies in this industry has not yet received much scholarly attention. In our study, we focus on the popular category of "casual games" (games characterized by a sequence of levels that can be solved in a short amount of time) and analyze data from one of the most popular match-3 games of all times. The data consist of the full in-game behavior and purchase decisions of about 300,000 players around the world for a two-week period between the end of October and early November 2013.

In the game, players solve puzzles through levels, and while the first 40 levels are free, from level 40 players must unlock a "pay-gate" every 20 levels to proceed (to the premium levels). Players can unlock any of these pay-gates by purchasing a "key." In this paper, we study the welfare consequences—for both the game developer and players—of five alternative pricing strategies to unlock pay-gates: the game developer's observed pricing; optimal uniform pricing; two forms of personalized pricing (third-degree price discrimination), one based on a player's gaming ability as measured in the free levels prior to the first pay-gate, the other based on the GDP per capita of a player's country; and first-degree price discrimination.

We combine experimental variation in the data with a structural model to estimate demand for free and paid-for content (additional levels) and then simulate the above counterfactual pricing strategies. We rely on the experimental variation in the data in two ways. First, we use it to learn about player behavior and document that players are unsophisticated and myopic, which greatly simplifies the specification of the structural model and the subsequent counterfactual simulations. Second, we rely on the experimental variation to estimate the demand model, in particular to address the standard challenges of price endogeneity and endogenous sample selection that would otherwise complicate identification (Gandhi and Nevo, 2021). Another helpful feature of the game is that no advertisement was displayed to players around the time of data collection. This allows us to focus on in-app purchases as the only source of revenue. The co-existence of in-app purchases and advertisement would introduce dynamic interactions between pricing and advertising decisions which would be extremely hard to model, estimate, and ultimately simulate in counterfactual scenarios (Dubois et al., 2017).

Our counterfactual simulations suggest that observed pricing is far from optimal. By relying on optimal uniform pricing, the game developer could increase profit by 340%. Even more strikingly, while more flexible and discriminatory pricing strategies would lead to larger profit, the relative increases would be very limited

[1] Freemium refers to a hybrid pricing model combining free and paid features of a product—basic features of a product can be used for free perpetually while more advanced features or more intensive use requires the payment of a fee. Freemium is particularly common for mobile apps, where consumers strongly favor apps that are free and monetized through in-app purchases rather than advertisement (Ghose and Han, 2014a).

compared to simple uniform pricing: first-degree price discrimination would generate a mere 2% increase in profit over optimal uniform pricing. Our analysis suggests that this is the result of myopic player behavior, which limits the extra gains of more elaborate pricing strategies. All the alternative pricing strategies considered—including uniform pricing—would induce a transfer of surplus from players to game developer without generating, however, sizeable dead-weight losses on average.

Our findings are consistent with the aforementioned literature documenting that, sometimes, for-profit firms do not maximize profits (Cho and Rust, 2010; DellaVigna and Gentzkow, 2019; Dube and Misra, 2019; Fioretti, 2020; Hortaçsu et al., 2021; Huang et al., 2020; Orbach and Einav, 2007). The fact that uniform pricing results in profit close to more complex and discriminatory pricing strategies is in line with Chu et al. (2011). They show that in the context of a theater company, simple pricing rules can sometimes generate almost as much profit as complex ones that would however be difficult to implement. Our results are also consistent with Levitt et al. (2016), who document limited gains of second-degree price discrimination for a large online gaming firm, and more in general with the empirical literature on the trade-offs of price discrimination and personalized pricing in the era of big data (Rossi et al., 1996; Shiller and Waldfogel, 2011; Shiller, 2015; Waldfogel, 2015). Limited gains from price discrimination may partly explain why it is rarely observed in business practice, where—as already mentioned above—additional risks tied to consumer backlash also need to be considered (Council of Economic Advisors, 2015; DellaVigna and Gentzkow, 2019).

In contrast to our results, however, Dube and Misra (2019) document substantial returns of personalized pricing for a digital recruiting firm, highlighting the need for caution in drawing general conclusions: while we do not see any evidence for this in our analysis, in other digital contexts more complex pricing strategies may be more profitable. That said, both our results and Dube and Misra (2019) stress the large potential of "empirical" pricing rules. In the case of the game we study, by optimally choosing a uniform price on the basis of detailed data and appropriate empirical methods, the game developer could increase profit more than fourfold. Importantly, our results also highlight that, although these increases in profit would necessarily come at the expense of consumer surplus, the pricing strategies considered do not generate average losses in total welfare.

Our paper contributes to a recent and growing literature investigating various aspects of mobile apps. Due to data limitations, most researchers have either exclusively focused on the supply side or employed very aggregate measures of demand, such as aggregate rankings or number of downloads from app stores (Bresnahan et al., 2015; Carare, 2012; Ershov, 2018; Ghose and Han, 2014b; Yi et al., 2019; Yin et al., 2014; Yuan, 2020; Wen and Zhu, 2019). Our user-level panel data instead allow us to delve deeper into the in-app purchase behavior of about 300,000 users around the world and to investigate the efficiency of discriminatory pricing strategies in a mobile game.

Despite widespread interest amongst practitioners and scholars alike (Fudenberg and Villas-Boas, 2006, 2012; Varian, 1989), there is relatively limited empirical evidence on the returns of price discrimination in practical applications, and essentially none for mobile games.[2] In general, the extant empirical evidence is mixed, documenting limited returns in some cases (Rossi et al., 1996; Levitt et al., 2016; Shiller and Waldfogel, 2011; Shiller, 2015; Waldfogel, 2015) but larger in others (Adams and Williams, 2019; Cho and Rust, 2010;

[2]Even though we focus on first-degree and third-degree price discrimination, there is a small empirical literature investigating the returns of second-degree price discrimination (quantity discounts): in carbonated soft drinks (Iaria and Wang, 2021), in coffee shops (McManus, 2007), in cable television (Crawford and Shum, 2007), in the yellow pages (Aryal and Gabrielli, 2020), and for an online gaming company (Levitt et al., 2016).

DellaVigna and Gentzkow, 2019; Dube and Misra, 2019; Iaria and Wang, 2021; List, 2004). Our paper contributes to this debate by providing the first empirical investigation on price discrimination for a mobile game. Despite the focus on a specific game, our empirical analysis speaks to a broader audience than freemium game providers: pay-gates are important monetization mechanisms also for other types of digital content providers including newspapers, magazines, and streaming services (e.g., Amazon Prime and You Tube).[3]

From a methodological perspective, ours is one of few empirical papers that combine both structural methods and randomized experiments (Cohen et al., 2016; Dube and Misra, 2019; Einav and Levin, 2010; Levitt and List, 2009; Todd and Wolpin, 2020). Our structural demand model for game content is needed to simulate the likely welfare consequences of counterfactual pricing strategies not observed in the data, while the experimental variation allows us to mitigate some of the standard endogeneity issues that would otherwise cripple identification and estimation.

The paper continues as follows. Section 4.2 describes the game for which we have data and discusses the way prices are set. Section 4.3 describes the data and the available sources of exogenous variation. Section 4.4 describes how we model player behavior and Section 4.5 reports our estimation results. Section 4.6 discusses our simulations for a number of counterfactual pricing rules. Section 4.7 draws some conclusions.

4.2 Mobile Game

4.2.1 Game Description

We empirically investigate the efficiency of price discrimination in the context of a game app which was produced by a large mobile game developer ("firm") and launched in August 2013. Like other popular mobile games, such as Candy Crush Saga or Bejeweled, the game we study is a casual game characterized by a sequence of levels that can be cleared in a relatively short amount of time. It belongs to the mobile puzzle game genre and has been downloaded around 80 million times so far, making it one of the most popular match-3 games of all time. The goal for players is to clear levels by connecting lines of jellies of the same color in order to "splash" them and achieve varying objectives.

The initial allocation of jellies is random and a move consists of connecting at least three jellies of the same color; the longer the line (also called "snake") of connected jellies, the more points are awarded (see Figure 4.1). Connected jellies are removed and replaced by a random set of new jellies. Players must achieve different objectives to clear different levels, for example reach a minimum score, remove slime, move diamonds from top to bottom, and so forth, which are all achieved by connecting and removing jellies. The number of moves for each level is capped, and the maximum number of allowed moves varies by level. In contrast to traditional video games, the difficulty of each level does not increase as players advance. There are occasional spikes in difficulty in certain levels, although these do not occur at regular intervals (Debeauvais and Lopes, 2015). Levels distinguish themselves by their layout, objectives, or features, such as the presence of obstacles, and so on. To advance, players must clear every level. Once a level has been cleared, it can be replayed at any time.

Players are awarded a score for their performance, which largely depends on the length of the snakes

[3] Another related emerging literature is that on algorithmic pricing. The majority of economics papers in this literature have so far been theoretical, mainly about the potential for algorithmic pricing to facilitate collusion (Miklós-Thal and Tucker, 2019; Calvano et al., 2020; Brown and MacKay, 2021), even though a few studies in progress are investigating the topic empirically, such as Assad et al. (2020) for gas stations and Hortaçsu et al. (2021) for airline companies.

Figure 4.1: 3-match mobile game

formed as well as the total number of moves needed to clear a level. Upon clearing a level, players are awarded one, two, or three "stars" depending on their score for that level. Stars are cumulative and, as we explain below, play an important role in the monetization of the game, whereas the score is specific to each level and plays no role other than to determine the number of stars.

4.2.2 In-App Purchases and Monetization

The game is a freemium product. A certain number of levels can be played for free (with a few restrictions), but premium content, such as additional levels or features, need to be unlocked via in-app purchases. Importantly for our empirical analysis, during the period of our data collection, no in-app advertisement was displayed to players. This allows us to focus on in-app purchases as the only source of revenue for the firm and, in turn, to estimate a tractable choice model useful for the simulation of counterfactual pricing strategies. The co-existence of in-app purchases and advertisement would introduce dynamic interactions between pricing and advertising decisions which would be extremely hard to model, estimate, and ultimately simulate in counterfactual scenarios (Dubois et al., 2017).

In-app purchases must be paid in "virtual coins" and each player receives an initial endowment of 70 of these. This endowment of virtual coins corresponds to approximately $1 at the time our data collection. Once players have spent their endowment, they must purchase additional virtual coins to buy any of the following features.[4] First, players can purchase additional "moves" if they run out of these before having successfully cleared a given level. Second, players are initially endowed with five "lives." A life is lost every time a player attempts to but does not successfully clear a level. Lives replenish automatically, a life being added every 30 minutes. If a player loses all five lives, they either wait for 30 minutes before they can continue to play, or purchase a bundle of five lives. Alternatively, a player can gain lives by inviting friends to download the game via Facebook.

[4] Note that, since our data were collected, some of the game's features have changed and additional opportunities for in-app purchases were introduced.

Figure 4.2: Players' decision tree

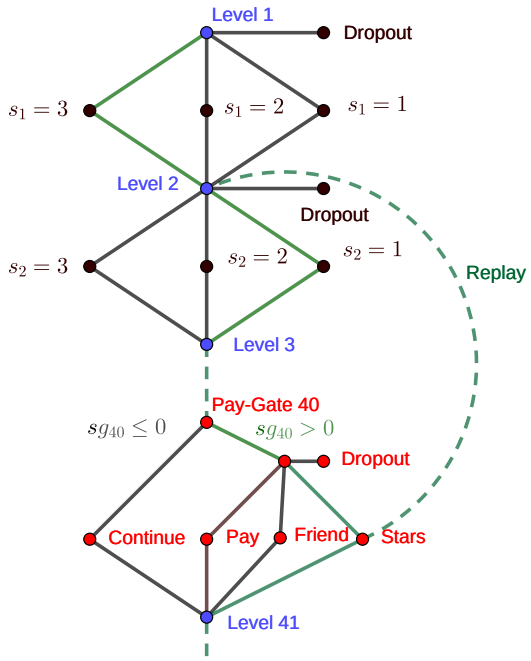


Figure 4.2 illustrates player i's decision tree. They start at level one and then either stop playing immediately or clear this level. By clearing any level, i can obtain three, two, or one star depending on the score obtained in that level. Starting from level 40, every 20 levels player i meets a "pay-gate," which we denote by $t = 40, 60, ..., T$ (i.e., pay-gates appear at level 40, 60, etc.). These pay-gates separate free levels from premium levels, and must be unlocked for i to proceed in the game. Players have three options to unlock pay-gates (see Figure 4.3): (a) purchase a "key" using 70 virtual coins, (b) invite friends on Facebook to download the game, or (c) accumulate a sufficient number of stars.

Regarding option (c), each pay-gate has a threshold number of stars that is pay-gate specific and rising as the player progresses through the game. If a player gets to the pay-gate with a number of stars equal or greater than this threshold, the pay-gate unlocks. For brevity, we refer to the difference between i's number of accumulated stars through their play up to pay-gate t and the number of stars needed to unlock pay-gate t as "star gap" and we denote it by $sg_{i,t}$. Only players with a positive star gap ($sg_{i,t} > 0$) must unlock pay-gate t. To do so, they can use either option (a) or (b), or alternatively can go back and re-play previous levels to gain additional stars where they obtained less than three, a behavior called "grinding."

At the moment of the data collection, the firm was relying on a simple uniform pricing strategy of 70 virtual coins (approximately $1) across all pay-gates and players for the purchase of a key to unlock a pay-gate.[5] As illustrated in Figure 4.4, the purchase of keys to unlock the first three pay-gates (levels $40, 60, 80$) corresponded to the largest share of in-app purchases (43%). Because of this and to maintain the econometric model and simulations practically viable, we focus on the firm's choice of which prices to charge for the keys to unlock

[5]Some form of price discrimination was still implemented by offering features in bundles as well as by offering quantity discounts on larger amounts of virtual coins.

Figure 4.3: Example of a pay-gate.

these three pay-gates.[6]

[6] As discussed below, in Appendix 4.A.2 we also provide supporting evidence that purchases of keys do not appear to crowd out other in-app purchases, suggesting that the two can be studied separately without excessive loss of generality.

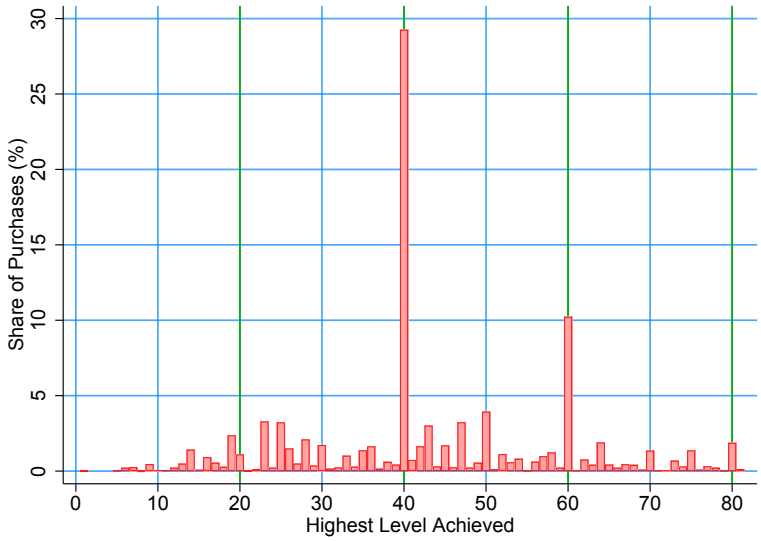


Figure 4.4: Spending patterns across life cycle

Notes: This figure displays the relative frequency of purchases observed in the game according to the highest level achieved. For example, a player who reaches pay-gate 40 with a positive star gap has as highest level achieved $\ell = 40$. If keys are purchased at this pay-gate, we count a purchase for $\ell = 40$. The sample includes all players (i.e, Group 20, Group 40, and Group No Star) and concerns only purchases at or before reaching level $\ell = 80$.

4.3 Data

4.3.1 Data and Variables

We have tracking data for all users around the world that installed the game between October 30th and November 4th 2013 on Apple devices (iPhones and iPads) and that played at least one round of the game. We have a sample of 292,179 players, and for each we observe the full history of play at an extraordinary level of detail for the 15 days following the installation of the game, including any purchase of virtual coins.[7] In particular, we rely on the following information to describe players' behavior and characteristics.

Level attempted and completed: For each player, we observe the level played in any given round of playing.[8] This allows us to track the sequence in which different levels are played and re-played. We also observe whether a level was cleared or not at a given attempt. Finally, we assume that a player drops out after the last attempt to clear a level.

Score and stars awarded: We observe the score each player was awarded for clearing a level. As discussed above, the score reflects how well a player performs in a given level. The number of stars awarded in a level is then determined as a function of the score obtained (stars are only awarded if a level is successfully cleared and star thresholds vary across levels).

Player's ability: A unique feature of online games, as opposed to more traditional offline games, is the possibility of measuring, almost in real time, a player's gaming ability. We measure a player's ability as the average snake-length over the first 20 rounds played, where the longer the average snake of connected jellies, the larger the

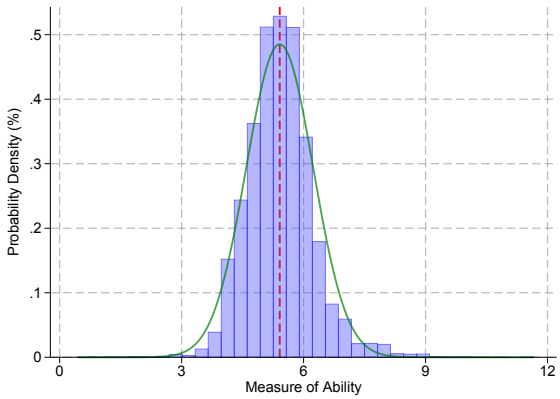
[7] Parts of these data were also used in Wagner and Runge (2018).

[8] Rounds denominate the cumulative number of levels played irrespective of whether a level was played multiple times or not.

score obtained by the player. The larger our measure of ability, the greater the player's skill in identifying the patterns required to succeed in the game. As shown in Figure 4.5, this measure of player's ability is almost normally distributed with some additional concentration around the mean. It can also be interpreted in relative terms: the top player (with a measure of 11.6) can be considered "twice as able as" the average player (with average measure of 5.4).

Controlling for ability is important both economically and econometrically. Economically, willingness to pay may vary with players' ability, affecting the optimal pricing strategy of the firm. Analogously, not controlling for players' ability may give rise to complex forms of endogeneity hard to address econometrically.

Figure 4.5: Distribution of players' ability



Notes: This figure displays the distribution of our measure of ability across players. The mean is presented through the vertical red line and the plot is overlaid with a normal distribution in green. The description of this variable is provided in Section 4.3.1. The sample is a cross-section of all players (Group 20, Group 40, and Group No Grind).

Star gap: As discussed above, each pay-gate is unlocked (and will not appear again) if a player reaches it with a sufficient number of stars, what we call a non-positive star gap. When a player reaches a locked pay-gate (i.e., with a positive star gap), they cannot proceed in the game until they unlock it. When reaching a locked pay-gate, players can grind in an attempt to decrease their star gap enough so to unlock the pay-gate. To simplify the analysis, we consider a measure of star gap inclusive (or gross) of grinding, rather than considering grinding as a separate decision: we measure $sg_{i,t}$ to be i's star gap at pay-gate t *after* all the grinding—when they either unlock t or drop out of the game. If star gaps were measured *net* of grinding, i's decision at pay-gate t with $sg_{i,t} > 0$ would be whether to unlock t, to grind in order to lower $sg_{i,t}$, or to stop playing. Differently, by measuring star gaps *gross* of grinding, we simplify i's decision at pay-gate t with $sg_{i,t} > 0$ to be only between unlocking t or stopping to play, given that $sg_{i,t} > 0$ is already inclusive of all of i's grinding at t.[9]

Pay-gate locked and unlock mechanism: We observe whether a player reaches a locked pay-gate and how they unlock it (or whether they drop out of the game). As mentioned above, when approaching pay-gate t with fewer stars than those necessary to unlock it, $sg_{i,t} > 0$, player i's options to unlock t are: (a) paying 70 virtual coins to purchase a key, (b) inviting a friend on Facebook to download the game, or (c) going back to previous levels to collect more stars (i.e., grinding). Importantly for our econometric analysis, as discussed in detail below,

[9]This will become clearer after having formally specified the choice model in Section 4.5, see in particular footnote 15 and the surrounding discussion.

players are not notified about the appearance of pay-gates every 20 levels when they start playing.

Price to unlock a gate: Any player i reaching pay-gate t with a positive star gap $sg_{i,t} > 0$ cannot proceed in the game without unlocking it. As discussed above, a way to unlock a pay-gate is to purchase a key. In the period of our data, the price of a key was set by the firm to $p_{i,t} = 70$ virtual coins, approximately \$1, uniformly for any i and t. The appropriate choice of $p_{i,t}$ (in terms of virtual coins) by the firm, potentially discriminating across i's and t's, is the main object of our empirical analysis. Because keys are priced in virtual coins and players can rely on their endowment of virtual coins to buy keys (thus spending potentially less in terms of real money), we consider $p_{i,t}$ as the effective or residual price of purchasing a key: e.g., the full price of the key minus i's endowment of virtual coins when reaching t (e.g., if the full price of a key is 70 virtual coins and i owns 30 virtual coins when reaching pay-gate t with $sg_{i,t} > 0$, then $p_{i,t} = 70 - 30 = 40$).

Player demographics: We observe a number of player-specific characteristics measured when a player downloads the game. These variables are collected in a vector we call X_i throughout the paper. Most of these characteristics relate to the device used to play the game. We know whether the game was downloaded to a mobile phone or a tablet (iPad). We also observe whether a player has updated their device to the latest version of the relevant operating system (iOS7) and whether the device was "jailbroken" by its owner (Jailbroken).[10] Finally, we observe the country of a player (as indicated by the national app store used to download the app) and relate it to its 2013 GDP per capita measured in purchasing power parity. We assign these countries to fourteen groups which we refer to as "regions." This assignment is detailed in Appendix 4.G and the share of players in each region is displayed in Figure 4.G.1.

Table 4.1 reports summary statistics for X_i among the 292,179 players in the data.

Table 4.1: Descriptive statistics of players' characteristics X_i

	mean	sd	min	max
Maximum Level Reached	21.30	17.16	0.00	179.00
Player's Ability	5.41	0.82	0.44	11.67
Log(GDP per Capita (PPP, 2013))	10.55	0.48	6.54	11.85
Jailbroken Dummy	0.01	0.11	0.00	1.00
iOS7 Dummy	0.78	0.42	0.00	1.00
iPad Dummy	0.31	0.46	0.00	1.00
Num. of Players	292,179			

Notes: This table provides descriptive statistics for the demographic variables. The definitions of the variables are detailed in Section 4.3.1. The sample includes all players (Group 20, Group 40, and Group No Grind) and the statistics are computed across this cross-section of players.

4.3.2 Exogenous Variation

In addition to the extremely detailed player-specific information described above, our data are also unique in providing various sources of exogeneous variation helpful to characterize players' behavior and to identify our econometric model. The first source of exogeneous variation is represented by controlled experiments conducted by the firm during the period of our data collection. The second is represented by a form of randomness in the degree of difficulty faced by different players when playing any level.

[10] Jailbreaking means removing all restrictions imposed on the device in order to allow the installation of software not supported by Apple.

4.3.2.1 Controlled Experiments

During the period of our data collection, the firm conducted a controlled experiment that randomly allocated players to three different designs of the pay-gates separating free from premium levels. Figure 4.3 illustrates what the firm considered the default design of the pay-gates in the game: 40 free levels before the first pay-gate appears, with a new pay-gate appearing every 20 levels thereafter. The default design allows for three options to unlock a pay-gate: (a) paying 70 virtual coins to purchase a key, (b) inviting friends to download the game via Facebook, or (c) having a non-positive star gap $sg_{i,t} \leq 0$. About 16% of all players were allocated to this default design (called *Group 40*).

The experimental variation introduced by the firm consists of two variations relative to the default design. In the first variation (called *Group 20*), a subset of nearly 16% of players was exposed to an earlier first pay-gate already after clearing level 20. This setting allows for the same three options to unlock pay-gates as the default design. In the second variation (called *No Stars*), the first pay-gate appears after clearing level 40 as in the default design, however option (c) to unlock pay-gates with non-positive star gaps is not available. When i from this group reaches any pay-gate t, independently of $sg_{i,t}$, they must choose either option (a) or (b) to unlock it and proceed in the game. The No Stars group represents around 68% of players in our sample. Table 4.2 summarizes the main features of these three groups.

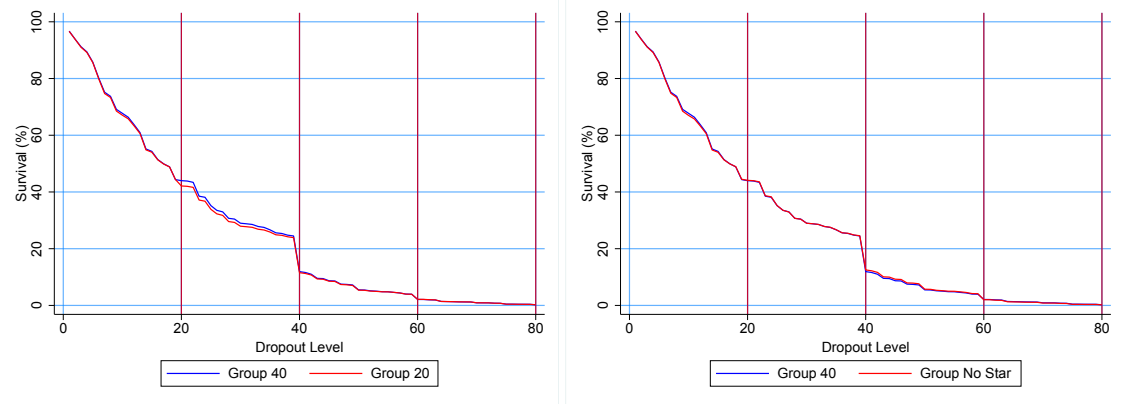
Table 4.2: Experimental groups

Group	Level 1st pay-gate	Unlock with sg≤0	Share in sample
Group 40	40	Yes	0.16
Group 20	20	Yes	0.16
No Stars	40	No	0.68

Table 4.G.2 in the Appendix provides descriptive statistics for these three groups, confirming that player-specific characteristics are well-balanced.

Player attrition. Player attrition is a common feature in this type of game. Figure 4.6 shows the share of active players at the start of each level (i.e., players who have not yet dropped out), comparing Group 20 to Group 40 (left panel) and No Stars to Group 40 (right panel). The graph shows that attrition is high in all three groups as players progress in the game. Overall, only about 24% of all players clear level 40 and reach pay-gate 40. Not surprisingly, the left panel of Figure 4.6 reveals a slight difference in attrition rates at level 20 between Group 20 and Group 40, as only players in Group 20 face a pay-gate at level 20. However, between levels 21 and 40, the gap between the two groups closes again and approximately 23.9% of Group 20 reaches level 40 compared to 24.5% of Group 40. The right panel of Figure 4.6 shows similar attrition rates for Group 40 and the No Stars group at level 40, suggesting that both groups are strongly affected by the appearance of the first pay-gate. Overall, this evidence is in line with Debeauvais and Lopes (2015), who document—for a different cohort of players—that attrition is larger in levels with pay-gates than in regular levels. In our analysis, we discard observations beyond level 80. This is essentially without loss of generality, in that 99.95% of all players drop out before level 80.

Figure 4.6: Kaplan-Meier survival function (by group)



Notes: This figure displays the share of players according to the highest level ℓ they have reached, by treatment group. The left panel compares Group 20 and 40. The right panel compares Group 40 and Group No Star. These treatment groups are defined in Section 4.3.2.1. The sample includes all players.

Choices at pay-gates. In Table 4.3, we summarize players' choices at the pay-gates of levels 20 and 40. In Group 20 about 83.09% of all players reached the first pay-gate at level 20 with sufficient stars to unlock it ($sg \leq 0$). The vast majority of players without sufficient stars ($sg > 0$) used the initial endowment of virtual coins to purchase a key (13.93%), only 1.41% used real money, while 1.56% invited friends on Facebook to download the game.

Table 4.3: Comparison of unlock mechanisms between treatment and control groups

		Group 20		**Group 40**		**No Stars**	
		1^{st} pay-gate at level 20		1^{st} pay-gate at level 40			
		Number	%	Number	%	Number	%
Players reach pay-gate 20		19,183	44.38	19,367	44.35	91,394	44.49
Players unlock pay-gate 20		18,203	94.89				
Unlock option							
$sg \leq 0$		15,126	83.09				
$sg > 0$	Buy key: Real money	257	1.41				
	Buy key: Endowment	2,536	13.93				
	Facebook	284	1.56				
Players reach pay-gate 40		10,329	23.90	10,694	24.49	50,469	24.57
Players unlock pay-gate 40		4,989	48.30	5,203	48.65	25,701	50.92
Unlock option							
$sg \leq 0$		1,135	22.75	1,094	21.02		
$sg > 0$	Buy key: Real money	853	17.10	826	15.87	5,350	20.81
	Buy key: Endowment	1,791	35.89	2,168	41.66	14,121	54.94
	Facebook	1,210	24.25	1,115	21.42	6,230	24.24

This indicates that the pay-gate at level 20 is a relatively soft monetization trigger, as most players were able to unlock it either with $sg \leq 0$ or using their initial endowment of virtual coins. At level 40, the share

of players with $sg \leq 0$ is significantly lower compared to the pay-gate at level 20, yet similar across Group 20 (22.75%) and Group 40 (21.05%). The share of players purchasing a key using real money increases to 17.10% for Group 20 and to 15.87% for Group 40. As expected, players in the No Stars group, which cannot unlock pay-gate 40 using their accumulated stars, were significantly more likely than the others to purchase a key, either using real money (20.81%) or their endowment of virtual coins (54%).

As discussed in Sections 4.4 and 4.5, we use this experimental variation to test competing hypotheses of players' behaviour and specify a more appropriate choice model, such as the degree of forward-looking behavior with respect to upcoming pay-gates (comparing Group 20 and Group 40), and to overcome identification concerns about endogenous selection on positive star gaps in estimation (relying on the No Stars group).

4.3.2.2 Randomness in the Difficulty of Levels

The structure of the game offers another useful source of exogenous variation: the difficulty of levels. Every time a new round of the game is played, there is a random draw of jellies which may incidentally deliver an easier or harder problem for the player to solve. A "good" draw may lead the player to succeed at a given level, while a "bad" one may be enough to induce the same player to fail. When a player gets closer to failing a level, they may face more of an incentive to purchase and spend virtual coins to obtain additional lives or moves, so to get the final boost needed to clear the level. Controlling for a player's ability, worse random draws of jellies will result in stronger incentives to purchase and spend virtual coins for reasons other than a key.

Following this line of reasoning, as detailed in Section 4.5, we exploit i's "bad luck in the random draws of jellies" as a source of exogenous variation—an instrument—for i's effective price of a key $p_{i,t}$ (defined as 70 virtual coins minus i's endowment) in the estimation of i's probability to purchase a key.

4.4 Characterizing Player Behavior Using Exogenous Variation

4.4.1 The Firm's Revenue Function

Our goal is to estimate a realistic but parsimonious model of player behavior useful to simulate alternative pricing strategies and their returns both to the firm and to players. We focus on the firm's revenue from purchases of keys and ignore other in-app purchases (e.g., additional lives or moves). This approach is motivated by the evidence shown in Figure 4.4 (most in-app purchases are concentrated at pay-gates) and the absence of any in-app advertising at the time of our data collection. We also show in Appendix 4.A.2 that purchases of keys do not crowd out other in-app purchases, suggesting that the two can be studied separately without excessive loss of generality.

We aggregate player i's decisions between any two pay-gates into a single choice, abstracting from the intermediate choices made at each level. Our choice model has two components. First, the probability that i reaches pay-gate t (denoted by $i \rightarrow t$) with a positive star gap given that she already unlocked pay-gate $t-20$:

$$Pr_{i,t}(i \rightarrow t, sg_{i,t} > 0 | t+20, ..., T). \tag{4.4.1}$$

Second, conditional on pay-gate t being locked to i (denoted by the indicator $\text{lock}_{i,t} = 1$) and effective

price $p_{i,t}$, the probability of purchasing a key at t (denoted by $\text{buy}_{i,t} = 1$, we describe this categorical variable in Section 4.5 in more detail):

$$Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t}, t+20, ..., T). \tag{4.4.2}$$

Relying on (4.4.1) and (4.4.2), we then specify the firm's expected revenue from player i at pay-gate t, $R_{i,t}$. Finally, our simulation exercises entail the maximization of $R_{i,t}$ with respect to $p_{i,t}$ across all players and pay-gates—under various constraints on the flexibility of prices (from uniform price to first degree price discrimination). Because every additional player does not generate any increase in the firm's costs (at least within the range observed in our sample), throughout the paper we assume that the firm's marginal costs are zero, and that expected revenue equals expected profit.

In general, probabilities (4.4.1) and (4.4.2) could be complex functions of i's expectations about future realizations of any variable (e.g., $sg_{i,t+20}$ and $p_{i,t+40}$). To keep the empirical model manageable, especially in view of our extensive simulation exercises, we propose (and then verify empirically) the following simplifying assumption.

Assumption 1 (Myopia)**.** *Players' decisions in t are conditionally independent of expectations about future decisions and variables to be realized in $t+20, t+40, ..., T$.*

Assumption 1 implies that player behavior can be represented by a (i, t)-specific static choice model, so that (4.4.1) and (4.4.2) simplify to $Pr_{i,t}(i \to t, sg_{i,t} > 0)$ and $Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t})$, respectively. Importantly, the conditional independence implied by this assumption should be intended with respected to the observable characteristics we can control for in the data, such as i's demographics and ability, and various fixed effects. We discuss the details of the empirical specification of our model in Section 4.5.

Denote i's effective prices from pay-gate $t+20$ until T by $p_{i,>t} = (p_{i,t+20}, p_{i,t+40}, ..., p_{i,T})$. Then, given choice models (4.4.1) and (4.4.2) and Assumption 1, the firm's expected revenue from player i at pay-gate t of charging effective prices $p_{i,\geq t}$, given that i already unlocked pay-gate $t-20$, can be expressed as:

$$\begin{aligned}
&R_{i,t}(p_{i,t} | p_{i,>t}) \\
&= Pr_{i,t}(i \to t, sg_{i,t} > 0) \times Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t}) \times p_{i,t} \\
&+ Pr_{i,t}(i \to t, sg_{i,t} > 0) \times \big(1 - Pr_{i,t}(\text{buy}_{i,t} = 0 | \text{lock}_{i,t} = 1, p_{i,t})\big) \times R_{i,t+20}(p_{i,t+20} | p_{i,>t+20}) \\
&+ Pr_{i,t}(i \to t, sg_{i,t} \leq 0) \times R_{i,t+20}(p_{i,t+20} | p_{i,>t+20}).
\end{aligned} \tag{4.4.3}$$

As this expression indicates, despite the simplifying assumptions, the firm's expected revenue from i at pay-gate t is an intricate recursive function. Conditional on i having unlocked pay-gate $t-20$, it depends on i's probability of reaching pay-gate t with $sg_{i,t} > 0$ (so that pay-gate t is locked, $\text{lock}_{i,t} = 1$), i's probability of purchasing a key ($\text{buy}_{i,t} = 1$) given $\text{lock}_{i,t} = 1$, i's probability of unlocking the current pay-gate ($\text{buy}_{i,t} \neq 0$) given $\text{lock}_{i,t} = 1$, price $p_{i,t}$, and finally—should i unlock the current pay-gate—the expected revenue stemming

from potential purchases of keys to unlock future pay-gates.[11] The first line of model (4.4.3) denotes the firm's expected revenue from i's *current* purchase of a key to unlock pay-gate t, the second and third lines instead denote the firm's expected revenue from i's *future* purchases of keys to unlock pay-gates $t+20$, $t+40$, ..., T. Before getting to the details of how we specify the components of model (4.4.3), we rely on the exogeneous variation described above to test the consistency of Assumption 1 with observed player behavior.

4.4.2 Testing Assumption 1

The validity of the expected revenue function in (4.4.3) crucially depends on the validity of Assumption 1 for the game we study. Here we rely on the exogenous variation available in the data to provide empirical evidence in support of this simplifying assumption.

Assumptions 1 requires that players' current choices are not influenced by their expectations regarding future events. The experimental variation in our data allows us to test for the absence of forward-looking behavior in various ways. In particular, we exploit the exogenous information shock to players in Group 20, who become aware of the existence of pay-gates twenty levels before the other players.

If players were forward-looking, those in Group 20 could show different attrition rates compared to other players. Having passed the the first pay-gate at level 20, their expected utility from continuing the game could be lower due to the anticipation of additional pay-gates (i.e., entailing costs with positive probability) at future levels. As a consequence, attrition rates could be higher. In particular, we test for differences in the total number of rounds played between players in Group 20 that were exposed to the pay-gate at level 20 but had $sg_{i,t} \leq 0$ and similar players in Group 40 who were not exposed to the pay-gate at level 20. These "similar" players in Group 40 are those who had a sufficient number of stars to immediately unlock the pay-gate at level 20 had they been allocated to treatment Group 20. We do not find significant differences (Table 4.4, row 1).

Table 4.4: Experimental evidence for myopia: Group 20 vs Group 40

	Group 40	Group 20	Diff.	Std. Err.	Obs.
Number of rounds before drop out	121.026	122.860	-1.833	1.633	21,899
Rounds played between 21 and 40 (or drop out)	51.400	51.582	-0.181	0.509	21,899
Stars collected between 21 and 40 (or drop out)	24.014	23.951	0.063	0.149	21,899
Re-played levels between 21 and 40 (or drop out)	26.375	27.354	-0.978	0.761	21,899

Notes: This table presents evidence regarding players' forward-looking behavior. The description of the variables is provided in Section 4.3.1. The sample of players includes all players in Group 20 and 40 who have crossed level $\ell = 20$ with a non-positive star gap. Columns "Group 40" and "Group 20" report the mean for Group 40 and 20 players, respectively. Column "Diff." provides the (mean) difference between the two former columns. Column "Std. Err." presents the standard errors associated with the mean of column "Diff.".

Awareness of the existence of pay-gates should affect the propensity to grind of forward-looking players (i.e., re-play past levels to collect additional stars). A player that is aware of the existence of future pay-gates, and the possibility of unlocking them with a sufficient number of stars, should grind more than unaware players, in order to increase the chance of reaching the next pay-gate with a non-positive star gap. However, we find no significant difference in the number of rounds played between levels 21 and 40 for the same groups of players

[11] As mentioned in Section 4.2.2, players can unlock pay-gates not only by purchasing keys but also by asking friends on Facebook to download the game. In this sense, as detailed in Section 4.5, the categorical variable $\text{buy}_{i,t}$ can take more values than only 0 (stop playing) and 1 (purchase a key), and $(1 - Pr_{i,t}(\text{buy}_{i,t} = 0|\text{lock}_{i,t} = 1, p_{i,t})) \geq Pr_{i,t}(\text{buy}_{i,t} = 1|\text{lock}_{i,t} = 1, p_{i,t})$.

used in the previous test (Table 4.4, row 2). We also find no significant differences in the number of stars collected (Table 4.4, row 3) and in the number of re-played levels between levels 21 and 40 (Table 4.4, row 4).[12]

We look for evidence of forward-looking behavior in two additional ways. First, we test for evidence of forward-looking behavior by checking whether the number of additional stars collected after having cleared a level for the first time affects the player's probability to re-play that level. For example, if player i cleared level 23 for the first time with $n \in \{1, 2, 3\}$ stars and is able to move on to level 24, we check if their probability to re-play level 23 depends on n. Obtaining only one or two stars leaves open the possibility to collect an additional two or one stars, respectively, by re-playing the level, hence increasing the chance of reaching the next pay-gate with a non-positive star gap. Table 4.5 shows estimation results for a multinomial logit model of the probability to re-play any level between 21 and 40 that was cleared for the first time with n stars (with the case of $n = 3$ stars as the excluded category, i.e. no incentives to re-play those levels). In line with economic intuition, the estimated intercepts suggest that the probability to re-play any level in 21-40 is increasing in the number of stars a player can still collect by re-playing it. The Group 20 indicator, however, is not significantly different from 0, providing no statistical evidence in support of forward-looking behavior. This is also shown graphically in a more disaggregate way, level by level starting from level 1, in Appendix Figures 4.A.1, 4.A.2, and 4.A.3. These graphs confirm that players in Groups 20 and 40 have virtually identical probabilities of re-playing any specific level initially cleared with a given number of stars, both before and—importantly—after players in Group 20 become aware of the existence of the pay-gate at level 20.

Table 4.5: Experimental evidence for myopia: Group 20 vs Group 40

	Multinomial Logit
Pr(Re-play level in 21-40, Stars $= 1$)	
Group 20 (relative to Group 40)	0.014
	(0.057)
Constant	0.602
	(0.039)
Pr(Re-play level in 21-40, Stars $= 2$)	
Group 20 (relative to Group 40)	0.013
	(0.053)
Constant	0.083
	(0.037)
Observations	20,997

Notes: This table presents the results from multinomial logit regressions where we estimate the propensity to replay levels depending on the number of stars collected when first clearing the level. Each observation is a case where a player re-played a level of the game which was previously cleared. The sample includes all players in Group 40 and 20 which cleared level $\ell = 20$ with a non-positive star gap. The base category is (Re-play level in 21-40, Stars $= 3$). Standard Errors are clustered at the player-level.

Second, we inspect the distribution of effective prices players face when they reach the pay-gate at level 40. Remember that $p_{i,t}$ is defined as 70 virtual coins (the price of a key to unlock any pay-gate) minus i's residual endowment of virtual coins by the time they reach t. Since the initial endowment provided to every player is 70 virtual coins, by not using any of it until level 40, i would face $p_{i,40} = 0$ and be able to purchase a key to unlock pay-gate $t = 40$ without spending any real money. However, players can spend their endowment

[12] In additional tests (not reported, but available on request), we also check the propensity to re-play specific levels for the same group of players and find no significant differences.

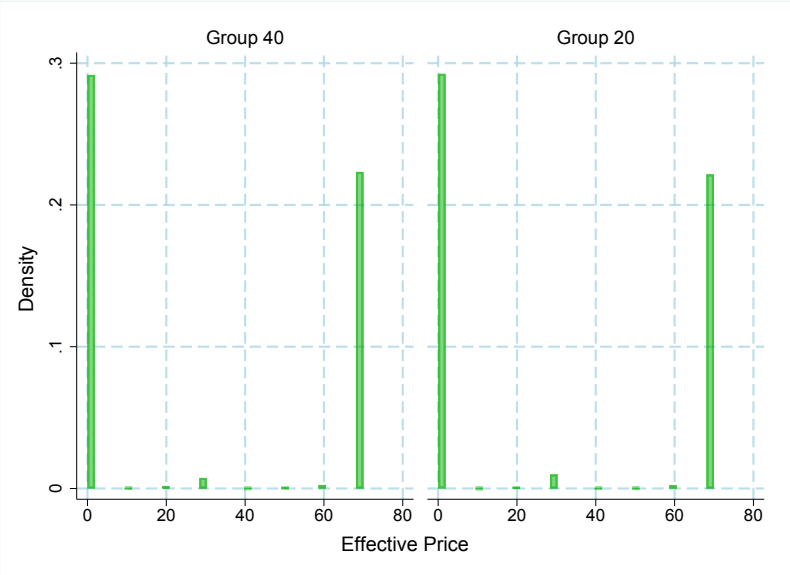


Figure 4.7: Distribution of effective prices $p_{i,t}$ at pay-gate $t = 40$

Notes: This histogram displays the effective price distribution at pay-gate 40 for Group 20 and Group 40. Each observation is a player who faces pay-gate 40 with a positive star gap. The sample includes all players of Group 40 and 20 who satisfy this condition.

on a number of items before reaching pay-gate $t = 40$ to enhance their play experience, such as purchasing boosters, additional lives or moves, etc. Awareness of the existence of pay-gates is expected to induce forward-looking players to save up on their endowment of virtual coins, so to ensure a lower effective price at the next pay-gate. We then ask whether players in Group 20 with $sg_{i,20} \leq 0$, who are aware of the existence of pay-gates from level 20, spend their endowment between levels 21 and 40 differently than players in Group 40, who are unaware of the existence of pay-gates until level 40. Figure 4.7 illustrates the distribution of effective prices faced by players in Group 20 and Group 40 when reaching the pay-gate at level 40. The distributions of effective prices faced by the two groups is very similar and indicates that awareness of the existence of pay-gates did not lead players in Group 20 to save more of their endowment of virtual coins in anticipation of the next pay-gate. To corroborate these results, Appendix Figures 4.A.4 and 4.A.5 repeat the same exercise for pay-gates $t = 60$ and $t = 80$.

4.5 Choice Model: Specification and Estimation

Our model abstracts from players' disaggregate level-specific choices and focuses on whether they reach each pay-gate t (discrete choice model (4.4.1)) and, conditional on reaching it with $sg_{i,t} > 0$, whether they choose to unlock it and how (discrete choice model (4.4.2)). Here we specify the empirical counterparts of these discrete choice models and estimate them relying on the exogenous variation described in Section 4.3.2.

4.5.1 Discrete Choice Model (4.4.1): Reaching Pay-Gates with a Positive Star Gap

We specify discrete choice model (4.4.1) as the product of two binary choice models:

$$Pr_{i,t}(i \to t, sg_{i,t} > 0) = Pr_t(i \to t|X_i) \times Pr_t(sg_{i,t} > 0|i \to t, X_i), \tag{4.5.1}$$

where X_i is a vector of observable i-specific characteristics such as i's demographics and player ability (see Section 4.3.1 for a description of these variables). After having unlocked pay-gate $t-20$, player i can either clear all levels between $t-19$ and t and reach the next pay-gate t, or stop playing before reaching it. This is the first binary choice model in (4.5.1), $Pr_t(i \to t|X_i)$. Upon reaching pay-gate t, we then distinguish between $sg_{i,t} > 0$ and $sg_{i,t} \leq 0$ to determine if i faces the next choice, discrete choice model (4.4.2), of whether and how to unlock pay-gate t. This is the second binary choice model in (4.5.1), $Pr_t(sg_{i,t} > 0|i \to t, X_i)$.

Neither of the binary choice models in our empirical specification (4.5.1) depends on the effective price $p_{i,t}$ of purchasing a key to unlock pay-gate t. While this is true by construction for $Pr_t(i \to t|X_i)$, in that $p_{i,t}$ can only be determined when i reaches pay-gate t, it may not be true for $Pr_t(sg_{i,t} > 0|i \to t, X_i)$. Given our definition of star gap as inclusive of grinding (see Section 4.3.1), this exclusion restriction may be violated for example if players were more likely to grind (and so to lower their star gaps) when facing higher effective prices, since unlocking pay-gates by non-positive star gaps would become relatively cheaper than by purchasing keys. Table 4.6 suggests this is not the case and reports supportive empirical evidence in favor of this exclusion restriction using the sample of players in Group 40,[13] a linear probability model for $Pr_t(sg_{i,t} > 0|i \to t, X_i)$ has an estimated coefficient on $p_{i,t}$ which is very close to zero, especially once we control for the observable i-specific characteristics X_i.

	(1) $Pr_t(sg_{i,t} > 0\|i \to t)$	(2) $Pr_t(sg_{i,t} > 0\|i \to t)$	(3) $\text{Pr}_t(sg_{i,t} > 0\|i \to t, X_i)$
Effective price, $p_{i,t}$	0.0000359	0.0000515*	0.0000258
	(0.0000192)	(0.0000206)	(0.0000188)
Pay-gate fixed effects	No	Yes	Yes
Player-specific characteristics (X_i)	No	No	Yes
Observations	12,600	12,600	12,600
Num. of players	10,692	10,692	10,692

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.6: Linear Probability Model of Effective Price on Positive Star Gap

Notes: This table shows estimation results for the linear probability model of the effective price $p_{i,t}$ on player i facing a positive star gap at pay-gate t. The dependent variable is a dummy variable equal to one when a player has a positive star gap, as defined by equation (4.4.1), and zero otherwise. The explanatory variable is the effective price, defined in Section 4.3. The sample includes all players from Group 40 who have reached pay-gate 40, 60, or 80. Each observation is a player/pay-gate combination. In the first column, we include no controls. In the second column, we add pay-gate fixed effects. In the third column, we include the demographics X_i defined in Section 4.3.1. Standard errors are clustered at the player level.

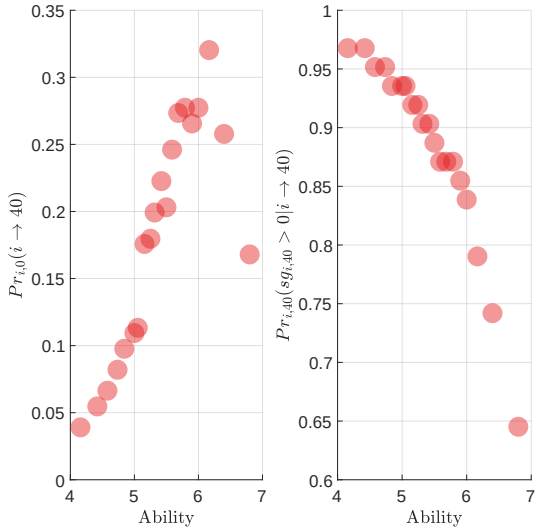
The exclusion of $p_{i,t}$ from the empirical specification of model (4.5.1) has practical implications for our analysis. Because of their conditional independence of $p_{i,t}$, the binary choice models in (4.5.1) will only act as "constant weights" in any of the maximizations of the firm's expected revenue (4.4.3) to be performed in our simulations. We therefore do not make any further assumption and estimate them as non-parametric functions of X_i for each pay-gate t.

We separately estimate each of the two binary choice models in (4.5.1) by a standard K-Nearest Neighbors

[13] As explained below, we focus on the players in this group to estimate model (4.5.1).

(kNN) estimator from the sample of players in Group 40 (the same sample used in Table 4.6). We restrict the estimation of model (4.5.1) to the players in Group 40, since this is the default design of the game.[14] In Appendix 4.B, we describe the kNN estimator of (4.5.1) and present its estimates, while in Figure 4.8 we plot the two estimated binary choice models as functions of player's ability.

Figure 4.8: kNN Estimates of Model (4.5.1) and Player's Ability



Notes: These figures display binned scatter plots of the relationship between the estimated probabilities of model (4.5.1) and ability. The estimates are produced using the kNN procedure described in Appendix 4.B. Ability is defined in Section 4.3.1. The left panel displays the probability of reaching pay-gate 40 given that the player is at level 0, $Pr_{i,0}(i \to 40)$. The right panel displays the probability of a positive star gap given that the player has reached pay-gate 40 $Pr_{i,40}(sg_{i,40} > 0 | i \to j)$. The sample used includes all players of Group 40.

We see that, in line with intuition, more able players are less likely to drop out of the game (except for the top 5% of players) (left panel) and are more likely to reach pay-gates with a non-positive star gap (thus unlocking them without the need to purchase keys) (right panel).

4.5.2 Discrete Choice Model (4.4.2): Purchasing a Key to Unlock a Pay-Gate

The estimation of discrete choice model (4.4.2), $Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t})$, presents at least two challenges: sample selection on $sg_{i,t} > 0$ and endogenity of $p_{i,t}$.

First, given the structure of the game, only players with $sg_{i,t} > 0$ face $\text{lock}_{i,t} = 1$ and can be observed to purchase a key at pay-gate t. In the absence of experimental variation, we would then have to estimate (4.4.2) exclusively on the sample of players observed to reach pay-gate t with $sg_{i,t} > 0$, $Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t}, sg_{i,t} > 0)$. It is however possible that the willingness to purchase a key at t differs systematically between the players observed with $sg_{i,t} > 0$ and those observed with $sg_{i,t} \leq 0$, so that $Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t}, sg_{i,t} > 0) \neq Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t})$. Fortunately, the experimental variation described in Section 4.3.2 allows us to overcome this by restricting estimation of model (4.4.2) to the sample of players in the No Stars group. Players in the No Stars group cannot use their accumulated stars to unlock

[14] Players in Group 20 face an additional pay-gate at level 20 which increases their attrition, while those in No Stars cannot use their accumulated stars to unlock pay-gates; hence they have fewer incentives to obtain non-positive star gaps.

pay-gates and, independently of their observed star gap, always face $\text{lock}_{i,t} = 1$.

Second, the effective price $p_{i,t}$, computed as 70 virtual coins minus i's residual endowment of virtual coins at pay-gate t, depends on i's decision of whether to purchase and spend virtual coins before reaching pay-gate t. This decision, in turn, may correlate to i- and t-specific unobservable characteristics that also drive i's willingness to purchase a key at pay-gate t. We address this potential endogeneity in two ways. On the one hand, our empirical specification of model (4.4.2) controls for i's ability in the game, which would otherwise be the most worrying omitted variable. On the other, as mentioned in Section 4.3.2 and discussed in more detail below, we also exploit randomness in the difficulty of each level across players as an instrument for $p_{i,t}$: at the beginning of each level, different players get a random draw of jellies which determines the level's difficulty, and in turn the players' incentives to purchase and spend virtual coins for reasons other than a key (e.g., additional lives or moves).

Model Specification. When i reaches locked pay-gate t, $\text{lock}_{i,t} = 1$, for given effective price $p_{i,t}$, i faces three options, denoted by $\text{buy}_{i,t} \in \{0, 1, 2\}$:[15]

$\text{buy}_{i,t} = 0$: Do not unlock pay-gate t and stop playing.

$\text{buy}_{i,t} = 1$: Purchase a key to unlock pay-gate t at a price of $p_{i,t}$ virtual coins, where 70 virtual coins cost around \$1 in terms of real money.

$\text{buy}_{i,t} = 2$: Ask a friend on Facebook to download the game.

In order for our simulations to be practically viable but still capture rich forms of observed and unobserved player-specific heterogeneity, we specify discrete choice model (4.4.2) parametrically as a mixed logit (McFadden and Train, 2000). Player i's conditional indirect utility from choosing $\text{buy}_{i,t}$ when facing locked pay-gate t is:

$$U_{\text{buy},i,t}(\eta_i) = \begin{cases} \epsilon_{0,i,t} & \text{if } \text{buy}_{i,t} = 0 \\ \delta_1 + \delta_{1,t} + \delta_{1,i} + X_i\beta_1 - (\alpha + \alpha_t + \alpha_i + X_i\pi)p_{i,t} + \epsilon_{1,i,t} & \text{if } \text{buy}_{i,t} = 1 \\ \delta_2 + \delta_{2,t} + \delta_{2,i} + X_i\beta_2 + \epsilon_{2,i,t} & \text{if } \text{buy}_{i,t} = 2, \end{cases} \tag{4.5.2}$$

where $(\delta_{\text{buy}} + \delta_{\text{buy},t} + \delta_{\text{buy},i})$ is an intercept given by the sum of a common δ_{buy} component among players, a pay-gate specific shift $\delta_{\text{buy},t}$ equal to zero at $t = 40$, and an unobserved player-specific random coefficient $\delta_{\text{buy},i}$, X_i is a vector of observable i-specific characteristics (see Section 4.3.1), $(\alpha + \alpha_t + \alpha_i + X_i\pi)$ denotes i's sensitivity to the effective price at pay-gate t, which is both a function of unobserved random coefficient α_i and observed heterogeneity $(\alpha + \alpha_t + X_i\pi)$, while $\epsilon_{\text{buy},i,t}$ is a residual error term we describe below. We allow for a particularly flexible specification of price sensitivity, itself a function of the observable characteristics X_i, to investigate the potential of pricing strategies that take advantage of the detailed player-specific information routinely collected by the firm. We gather the three random coefficients $(\delta_{1,i}, \delta_{2,i}, \alpha_i)$ into the random vector η_i and assume that they are jointly normal:

[15] Because our measure of star gaps is inclusive of grinding (see Section 4.3.1), any i with $sg_{i,t} > 0$ (and consequently with $\text{lock}_{i,t} = 1$) cannot—by definition—be observed to unlock pay-gate t by further grinding: all of i's grinding for pay-gate t is already included in $sg_{i,t}$. We therefore do not consider the option to grind as a further alternative to unlock pay-gates in model (4.5.7): all the grinding is captured by the second component of model (4.5.1), $Pr_t(sg_{i,t} > 0 | i \to t, X_i)$.

$$\eta_i = \begin{pmatrix} \delta_{1,i} \\ \delta_{2,i} \\ \alpha_i \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \rho_{12} & \rho_{1\alpha} \\ - & \sigma_2^2 & \rho_{2\alpha} \\ - & - & \sigma_\alpha^2 \end{pmatrix} \right]. \tag{4.5.3}$$

As is standard, we normalize the systematic component of the indirect utility of $\text{buy}_{i,t} = 0$ (the outside option of dropping out) to zero, $U_{0,i,t} - \epsilon_{0,i,t} = 0$, and include the effective price $p_{i,t}$ of purchasing a key only in the indirect utility of purchasing a key, $\text{buy}_{i,t} = 1$.

Price Endogeneity and Control Function. As mentioned above, the effective price $p_{i,t}$ could correlate with the residual error term $\epsilon_{1,i,t}$ and be endogenous. Given the player-level nature of the data, we cannot rely on the standard instrumental variable techniques to address price endogeneity typically used in demand estimation (Berry et al., 1995; Nevo, 2001b). However, we mitigate price endogeneity in two other ways. First, we include i's ability among the observed regressors X_i—thus removing from the unobservable $\epsilon_{1,i,t}$ the most problematic omitted variable. Second, we estimate the parameters of model (4.5.2) on the basis of a control function approach (Blundell and Powell, 2004; Blundell et al., 2013).

Our control function relies on an instrument $Z_{i,t}$ for price $p_{i,t}$ obtained from the randomness in the difficulty of each level across players (see Section 4.3.2). Controlling for i's ability, random variation in a level's difficulty prior to reaching pay-gate t will induce random variation in i's incentives to purchase and spend virtual coins on items other than a key (e.g., more lives or moves to clear the level), and consequently in $p_{i,t}$. In practice, we define $Z_{i,t}$ as the number of times we observe i being close to failing any of the levels between pay-gates $t-20$ and t, where we consider i "being close to failing" level ℓ as i's score in ℓ within a 5% interval below the ℓ-specific score threshold necessary to clear level ℓ:[16]

$$Z_{i,t} = \sum_{\ell=1}^{t} 1\left(0.95 \times \text{Necessary Score}_\ell < \text{Score}_{i,\ell} < \text{Necessary Score}_\ell\right). \tag{4.5.4}$$

We follow Petrin and Train (2010) and implement the control function approach to estimate model (4.5.2) as follows. Given the instrument $Z_{i,t}$ in (4.5.4), we assume that $p_{i,t}$ is given by:

$$p_{i,t} = \zeta_t + X_i\gamma + \lambda Z_{i,t} + \mu_{i,t}, \tag{4.5.5}$$

where ζ_t is an intercept and $\mu_{i,t}$ is an unobserved component of effective price potentially correlated with $\epsilon_{1,i,t}$ (causing price endogeneity) but independent of $(\epsilon_{0,i,t}, \epsilon_{2,i,t})$. We also assume that the expectation of $\epsilon_{1,i,t}$ conditional on $\mu_{i,t}$ is linear:[17]

$$\epsilon_{1,i,t} = \theta\mu_{i,t} + \tilde{\epsilon}_{1,i,t}, \tag{4.5.6}$$

where $\theta\mu_{i,t}$ is our control function. Finally, by substituting (4.5.6) back into (4.5.2), defining $(\tilde{\epsilon}_{0,i,t}, \tilde{\epsilon}_{2,i,t}) = (\epsilon_{0,i,t}, \epsilon_{2,i,t})$, $V_{\text{buy},i,t}(\eta_i) = U_{\text{buy},i,t}(\eta_i) - \tilde{\epsilon}_{\text{buy},i,t}$ for each $\text{buy}_{i,t} \in \{0,1,2\}$, and assuming that $(\tilde{\epsilon}_{0,i,t}, \tilde{\epsilon}_{1,i,t}, \tilde{\epsilon}_{2,i,t})$ are i.i.d. Gumbel (McFadden, 1974), we obtain our mixed logit specification of discrete choice model (4.4.2):

$$Pr_{i,t}(\text{buy}_{i,t} = 1 | \text{lock}_{i,t} = 1, p_{i,t}) = \int \frac{\exp(V_{1,i,t}(\eta_i))}{1 + \exp(V_{1,i,t}(\eta_i)) + \exp(V_{2,i,t}(\eta_i))} \phi(\eta_i|\Sigma) d\eta_i, \tag{4.5.7}$$

[16] In Appendix 4.C.1, we discuss this instrument in more detail and report the first step estimates of equation (4.5.5) along with alternative specifications of the instrument. Overall, estimation results are robust to alternative specifications of $Z_{i,t}$.

[17] We attempted the estimation of model (4.5.2) on the basis of various—more elaborate—specifications of both (4.5.5) and (4.5.6), but found no substantial differences. As a consequence, we decided to stick to these simpler and similarly effective linear specifications.

where $\phi(\cdot|\Sigma)$ is the normal density of η_i in (4.5.3) with Σ denoting its variance-covariance matrix and $V_{1,i,t}(\eta_i) = \delta_1 + \delta_{1,t} + \delta_{1,i} + X_i\beta_1 - (\alpha + \alpha_t + \alpha_i + X_i\pi)p_{i,t} + \theta\mu_{i,t}$ includes control function $\theta\mu_{i,t}$, based on (4.5.4), (4.5.5), and (4.5.6), to account for the potential endogeneity of $p_{i,t}$.

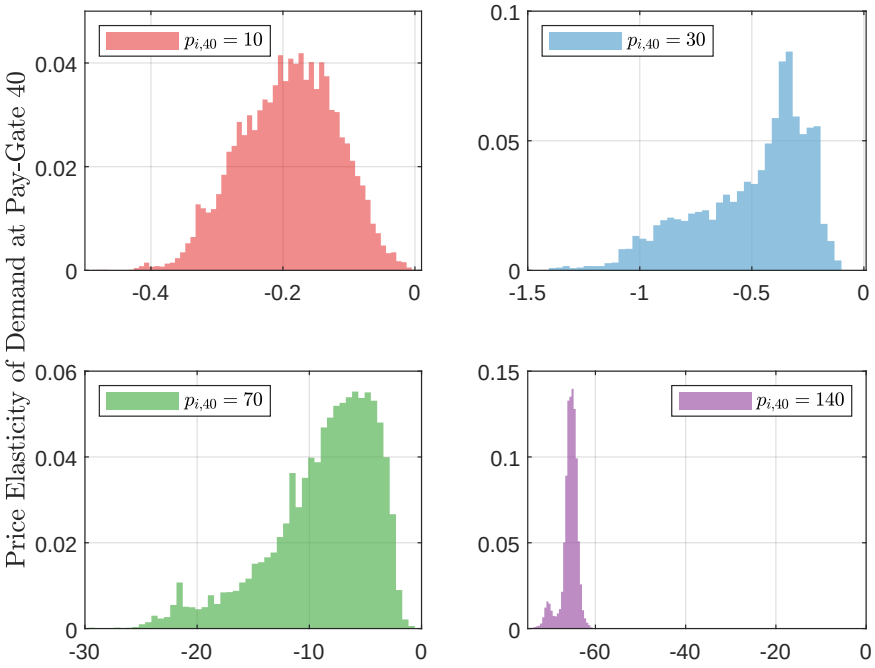
Estimation Results. In the absence of experimental variation, only players with $sg_{i,t} > 0$ face a locked pay-gate t and so the choice of whether to unlock it. This raises the concern that $Pr_{i,t}(\text{buy}_{i,t} = 1|\text{lock}_{i,t} = 1, p_{i,t}, sg_{i,t} > 0) \neq Pr_{i,t}(\text{buy}_{i,t} = 1|\text{lock}_{i,t} = 1, p_{i,t})$, which complicates the identification of mixed logit model (4.5.7). As mentioned above, however, the experimental variation described in Section 4.3.2 allows us to overcome this form of endogenous selection by estimating the model on the sample of players in the No Stars group, who cannot use their accumulated stars to unlock pay-gates and always face $\text{lock}_{i,t} = 1$.

Using the sample of players in the No Stars group, we address the additional concern of price endogeneity by estimating (4.5.7) on the basis of the control function approach proposed by Petrin and Train (2010). We first estimate (4.5.5) by OLS, compute each $\hat{\mu}_{i,t}$ as the fitted residual of that regression, then plug $\hat{\mu}_{i,t}$ in $V_{1,i,t}(\eta_i)$, and finally estimate mixed logit model (4.5.7) by Simulated Maximum Likelihood using 100 random Halton sequences per player (Bhat, 2003). We compute the variance-covariance matrix of the estimator as in Karaca-Mandic and Train (2003) to account for the two-step nature of the control function procedure. We report the results for both estimation steps in Appendix 4.C, while here we visually summarize the implied estimated price elasticities.

Figure 4.9 plots the distribution of the estimated price elasticities at pay-gate 40 evaluated at effective prices $p_{i,40} = 10, 30, 70, 140$ virtual coins (see Appendix 4.D.1 for the computational details). Each panel plots the distribution of price elasticities across players when everyone faces the same effective price $p_{i,40}$.[18] Two intuitive findings emerge from Figure 4.9: first, for any given effective price, there is heterogeneity across players facing the same pay-gate; second, as $p_{i,40}$ increases, a player's demand quickly becomes extremely elastic.

[18] Importantly, while we relied on the players in the No Stars for the estimation of model (4.5.7) to avoid sample selection complications, here we only plot the implied price elasticities for the players in Group 40, who face the default design of the game. See also footnote 19.

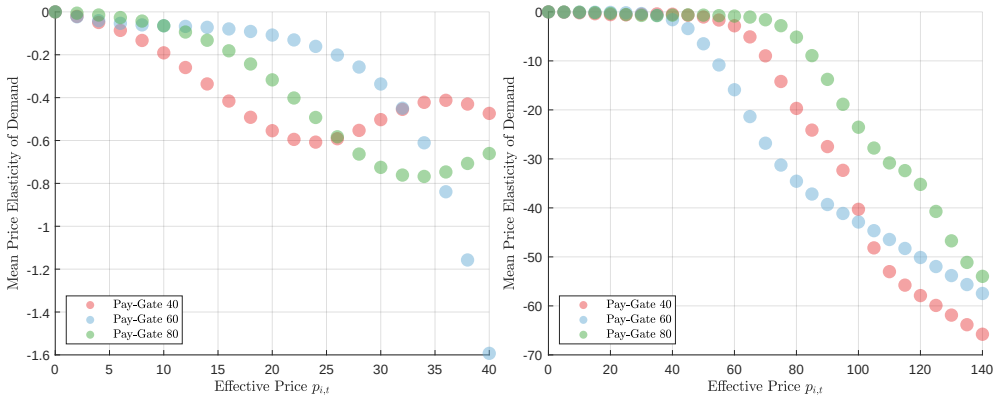
Figure 4.9: Price Elasticity of Demand at Pay-Gate 40



Notes: These histograms display the distribution of the price elasticity of demand at pay-gate 40 evaluated at different effective prices $p_{i,40} = 10, 30, 70, 140$. Price elasticity of demand is defined in Appendix 4.D.1. In each panel, the distribution of the price elasticity is evaluated at the same effective price $p_{i,40}$ for all players. The sample used to compute these price elasticities includes all players of Group 40 (see footnote 19).

Figure 4.10 compares average price elasticities of demand across pay-gates when evaluated at a given effective price. The left panel plots results for effective prices ranging from 0 to 40 virtual coins, while the right panel plots results for effective prices ranging from 0 to 140 virtual coins (note the much larger scale on the y-axis). Figure 4.10 makes clear that price elasticity is heterogeneous not only across players at a given pay-gate, but also across pay-gates. This is driven by heterogeneity across players dropping out of the game and thus "surviving" at each pay-gate. Moreover, the price elasticity seems to substantially increase (i.e., become negative) for all pay-gates once the effective price goes beyond 50 virtual coins (right panel).

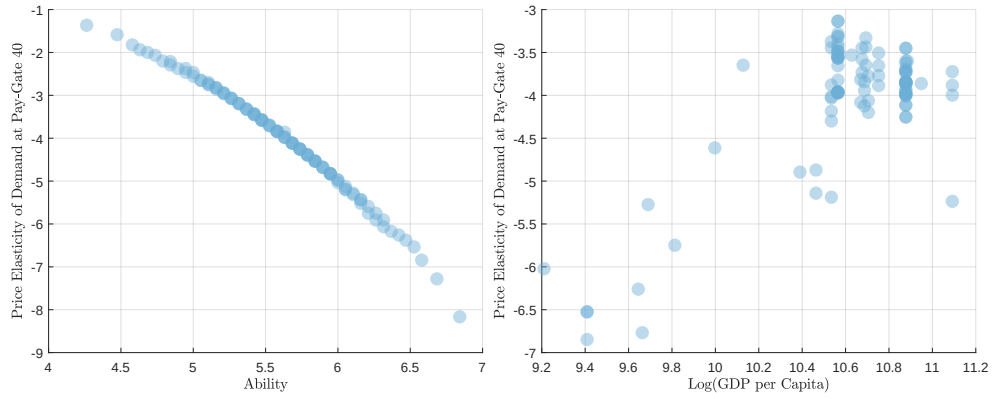
Figure 4.10: Average Price Elasticity of Demand across Pay-Gates



Notes: This figure displays the average price elasticity of demand across different pay-gates for given effective price. The left panel plots results for effective prices ranging from 0 to 40 virtual coins while the right panel for effective prices ranging from 0 to 140 virtual coins. For each given value of effective price p on the x-axis, we plot the average price elasticity of demand of each pay-gate so that $p_{i,40} = p_{i,60} = p_{i,80} = p$. The price elasticity of demand is defined in Appendix 4.D.1. The sample includes all players of Group 40 (see footnote 19).

Finally, Figure 4.11 displays two binned scatter plots of the price elasticity of demand at pay-gate 40 with respect to a player's ability (left panel) and the log(GDP per capita) of their country (right panel). The left panel shows a striking relationship between price elasticity and player's ability. Less able players are more inelastic with respect to the effective price $p_{i,t}$. For a given increase in $p_{i,t}$, they are more likely to purchase keys using real money rather than to stop playing or ask a friend to download the game. The right panel of Figure 4.11 then confirms an intuitive positive relationship between price elasticity and GDP per capita of a player's country: for given increase in $p_{i,t}$, players from wealthier countries are far less price elastic at pay-gate 40 than players from poorer countries.

Figure 4.11: Price Elasticity of Demand at Pay-Gate 40 by Ability and Log(GDP per Capita)



Notes: This figure shows two binned scatter plots of the price elasticity of demand at pay-gate 40 with respect to a player's ability (left panel) and the log(GDP per capita) of their country (right panel). Price elasticity of demand is defined in Appendix 4.D.1. Players' ability and log(GDP per capita) are defined in Section 4.3.1. In each panel, we segment the x-axis in 100 equally sized groups. For each of these groups, we then plot the average price elasticity of demand on the y-axis. The sample includes all players of Group 40 who reached pay-gate 40 with a positive star gap (see footnote 19). The effective prices used to calculate the price elasticity of demand are based on the empirical distribution of effective prices as described in Appendix 4.D.2.

Overall, these results suggest that, choosing effective prices on the basis of routinely collected data may

be profitable for the firm. For example, Figure 4.10 suggests that the firm may gain by setting effective prices somewhere around 40-50 virtual coins, higher than the observed average effective prices of around 35 virtual coins. In addition, Figure 4.11 confirms the importance of observing and controlling for player's ability in studying the pricing strategies of the firm, something that was not possible until a few years ago with standard offline games.

4.5.3 Model Validation

In Appendix 4.E, we conduct some model validation analysis and illustrate the estimated model's ability to predict player behavior under counterfactual pricing strategies.

4.6 Simulation of Alternative Pricing Strategies

In this last Section, we rely on our estimated model to evaluate the returns of alternative pricing strategies for the firm. To provide intuition, we first highlight some of the most salient trade-offs faced by the firm when choosing effective prices. These trade-offs uncover the complex nature of the optimization problem and highlight the value of the empirical methods we employ. Second, we simulate alternative pricing strategies characterized by increasing discrimination and compare their implied expected revenues to those observed to be earned by the firm. While we relied on the players in Group 40 for the estimation of model (4.5.1) and on those in No Stars for the estimation of model (4.5.7), we perform all counterfactual simulations only with respect to the players in Group 40. Players in Group 40 face the default design of the game, which corresponds to our discrete choice model in equation (4.5.1).[19]

4.6.1 Understanding the Firm's Optimization Problem

Here we use our estimates of models (4.5.1) (probability of reaching the next gate) and (4.5.7) (choice of how to unlock a gate) and the simulation procedures detailed in Appendices 4.D.1 and 4.D.2 to investigate whether the per-player expected revenue of the firm is affected by dynamic considerations (i.e., if prices at different pay-gates should be set jointly or can instead be chosen independently) and by player heterogeneity (i.e., if the firm should condition prices on observed ability and/or GDP per capita). In all simulations, per-player expected revenue is averaged across the 43,660 players in Group 40 during the 15 days of our sample in 2013.

Dynamics across Pay-Gates. Although we show in Section 4.4.2 that players behave myopically, it is still possible for the firm's optimal pricing to involve dynamic considerations across pay-gates. As we saw in Figure 4.10, because different players drop out of the game at different levels, the price responsiveness of the "surviving" population changes at different pay-gates. Potentially, the firm could then increase expected revenue by influencing this selection mechanism with an appropriate choice of effective prices at different pay-gates. For instance, Figure 4.12 shows how per-player expected revenue (in $) from pay-gates 40, 60, and 80 changes as the firm sets different combinations of effective prices across pay-gates 40 and 60 (keeping $p_{i,80} = 70$). Each line represents the perimeter of an iso-revenue area, gathering all combinations of $(p_{i,40}, p_{i,60})$ that deliver an

[19] Counterfactual simulations based on players from the other experimental groups would not be very informative, in that both players in Group 20 and in No Stars face game rules incompatible with those of the default design of the game embodied in model (4.5.1).

identical per-player expected revenue. Darker shades of blue are associated with lower per-player expected revenue.

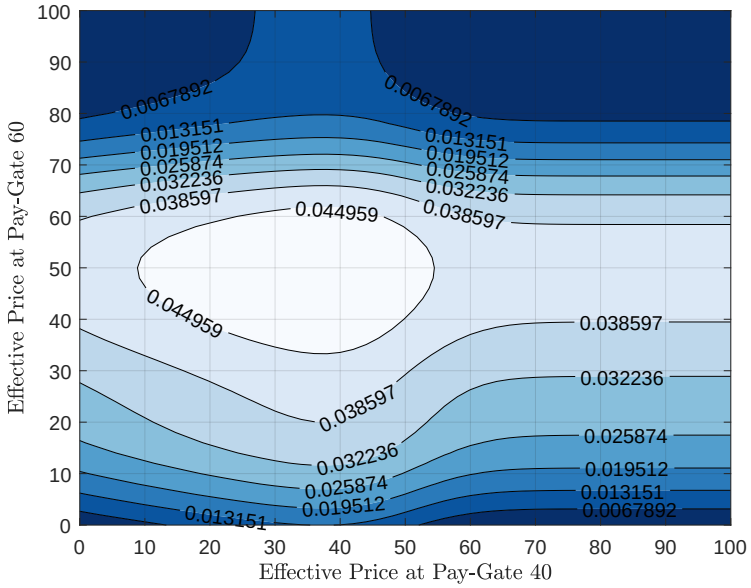


Figure 4.12: Dynamic Pricing and Iso-Revenue Lines

Notes: This figure compares the per-player expected revenue (in $) from pay-gates 40, 60, and 80 for different combinations of effective prices at different pay-gates. Each line represents the perimeter of an iso-revenue area, gathering all combinations of $(p_{i,40}, p_{i,60})$ (while keeping fixed $p_{i,80} = 70$) that deliver an identical per-player expected revenue. Darker shades of blue are associated to lower per-player expected revenue. The simulation of per-player expected revenue is based on our estimates of models (4.5.1) and (4.5.7) and the procedure detailed in Appendices 4.D.1 and 4.D.2. Per-player expected revenue is averaged across the 43,660 players in Group 40 during the 15 days of our sample in 2013.

The per-player expected revenue iso-quants depicted in Figure 4.12 reveal two findings. First, there is some dynamic connection between the effective prices across pay-gates. For example, for fixed $p_{i,60} = 50$ virtual coins, the firm can achieve the largest per-player expected revenue by setting $10 \leq p_{i,40} \leq 55$ virtual coins. For $10 > p_{i,40} > 55$, the firm would decrease per-player expected revenue because keys to unlock pay-gate 40 would be either too cheap or too expensive, so that too many players would not unlock pay-gate 40 and drop out of the game. Second, per-player expected revenue appears to be more responsive to changes in $p_{i,60}$ than to changes in $p_{i,40}$, stressing that not all pay-gates carry the same weight in terms of per-player expected revenue for the firm. Despite all this, it is however important to highlight that Figure 4.12 does not exclude the possibility that optimal effective prices across pay-gates may coincide, so that $p_{i,40} = p_{i,60}$.

The potential dynamics of the problem faced by the firm can also be directly seen by looking at the per-player expected revenue function presented in equation (4.4.3). The effective price $p_{i,t}$ plays two roles: (i) it affects i's expected revenue from pay-gate t, $Pr_{i,t}(\text{buy}_{i,t} = 1|\text{lock}_{i,t} = 1, p_{i,t}) \times p_{i,t}$ and (ii) it also affects i's expected revenue from future pay-gates by changing the probability of a player dropping out, $(1 - Pr_{i,t}(\text{buy}_{i,t} = 0|\text{lock}_{i,t} = 1, p_{i,t})) \times R_{i,t+20}(p_{i,t+20}|p_{i,>t+20})$. In this sense, the effective price $p_{i,t}$ must be chosen by the firm to balance the per-player expected revenue from current pay-gate t and that from future pay-gates $t' > t$. In Figure 4.13, we separately illustrate these by plotting the current (i) and future (ii) components of per-player expected revenue (in $) from pay-gate 40 as a function of $p_{i,40}$, holding fixed $(p_{i,60}, p_{i,80}) = (70, 70)$.

On the one hand, by increasing $p_{i,40}$, per-player expected revenue from *future* pay-gates 60 and 80 decreases due to a decrease in the probability of purchasing a key which is not compensated by an increased

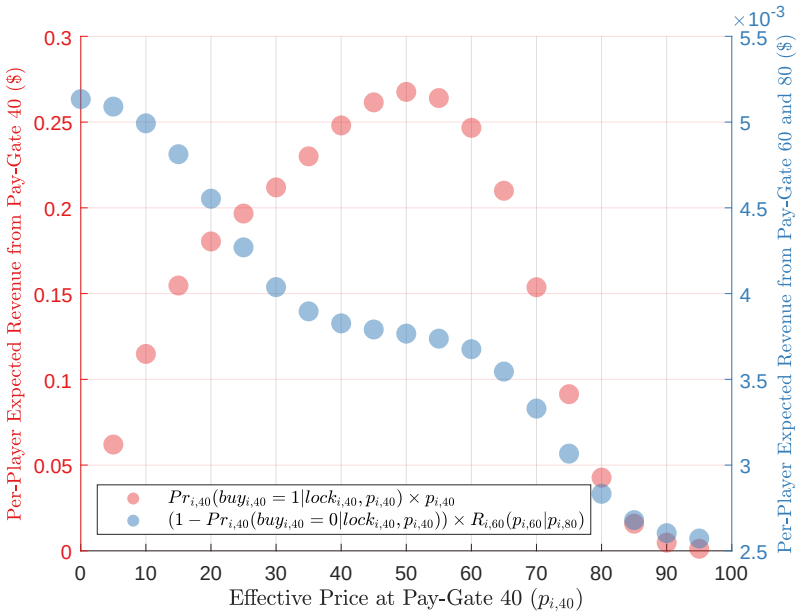


Figure 4.13: Current and Future Components of Per-Player Expected Revenue

Notes: This figure displays the current and future components of per-player expected revenue (in $) from pay-gate 40 as a function of $p_{i,40}$, as described in equation (4.4.3). On the x-axis we report different values of the effective price at pay-gate 40 ($p_{i,40}$) and on the y-axis we show: on the left, the per-player expected revenue from pay-gate 40 denoted by $Pr_{i,40}(\text{buys}_{i,t} = 1|\text{lock}_{i,40} = 1, p_{i,40}) \times p_{i,40})$; on the right, the per-player expected revenue from future pay-gates 60 and 80 evaluated at $(p_{i,60}, p_{i,80}) = (70, 70)$ and denoted by $(1 - Pr_{i,40}(\text{buys}_{i,t} = 0|\text{lock}_{i,40} = 1, p_{i,40})) \times R_{i,60}(p_{i,60} = 70|p_{i,80} = 70)$. The simulation of per-player expected revenue is based on our estimates of models (4.5.1) and (4.5.7) and the procedure detailed in Appendices 4.D.1 and 4.D.2. Per-player expected revenue is averaged across the 43,660 players in Group 40 during the 15 days of our sample in 2013.

probability to ask a friend to download the game to unlock pay-gate 40, $(1 - Pr_{i,40}(\text{buy}_{i,40} = 0|\text{lock}_{i,40} = 1, p_{i,40}))$. On the other, by increasing $p_{i,40}$, per-player expected revenue from *current* pay-gate 40 increases up to $p_{i,40} = 50$ and quickly falls afterward. Importantly though, note that the scale of the y-axis on the right-hand side of Figure 4.13 is of an order of magnitude smaller than that on the left-hand side. This suggests that the per-player expected revenue from current pay-gates could be what really matters when choosing effective prices, and that ignoring this inter-pay-gates trade-off may not be very costly for the firm.

Player Heterogeneity. Any given price change will not impact homogeneously the expected revenue from different players. As pointed out in Figure 4.11, this is due to player heterogeneity in terms of ability and GDP per capita, which translates into heterogeneous price sensitivities. We explore this in Figure 4.14 by comparing the per-player expected revenue from pay-gates 40, 60, and 80 for various effective prices across players with different ability (left panel) and from countries with different GDP per capita (right panel). In each panel, we set effective prices across all pay-gates and players to be uniform and equal to, in turn, 10, 30, and 70. Each point represents the average simulated per-player expected revenue for the 5% of players closest to the value of ability or log(GDP per capita) on the x-axes.

By looking at any binned scatter plot of the same color, we note that any uniform effective price leads to a different per-player expected revenue depending on a player's ability or log(GDP per capita). In addition, by comparing the vertical distances among plots of different colors, we also observe that different uniform effective prices will differently impact players with heterogeneous levels of ability or log(GDP per capita). For example, while the three uniform prices considered lead to similar per-player expected revenues for players with lower

ability (up to 5), more able players (with ability larger than 5) generate up to double the amount of per-player expected revenue when the uniform price is 30 as opposed to 10 or 70.

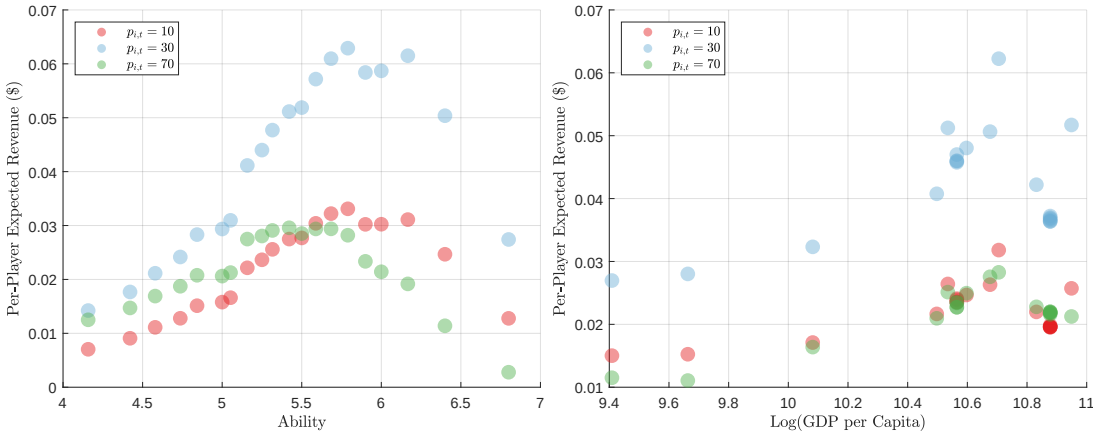


Figure 4.14: Per-Player Expected Revenue by Ability and GDP per Capita for Alternative Prices

Notes: These binned scatter plots compare the per-player expected revenue from pay-gates 40, 60, and 80 for various effective prices across players with different ability (left panel) and from countries with different GDP per capita (right panel). In each panel, we set effective prices across all pay-gates and players to be uniform and equal to, in turn: 10, 30, and 70. Each point represents the average simulated per-player expected revenue for the 5% of players closest to the value of ability or log(GDP per capita) on the x-axis. Ability and log(GDP per capita) are described in Section 4.3.1. The simulation of per-player expected revenue is based on our estimates of models (4.5.1) and (4.5.7) and the procedure detailed in Appendices 4.D.1 and 4.D.2. Per-player expected revenue is averaged across the 43,660 players in Group 40 during the 15 days of our sample in 2013.

4.6.2 Simulation Results

Next, we combine our estimates of models (4.5.1) (probability of reaching the next pay-gate) and (4.5.7) (choice of how to unlock a pay-gate), as well as the simulation procedures detailed in Appendices 4.D.1 and 4.D.2 to investigate the welfare effects of alternative pricing strategies, each requiring different levels of sophistication and amounts of data. All simulations are based on the 43,660 players in Group 40 during the 15 days of our sample in 2013.

We simulate alternative pricing strategies in which the firm directly chooses $p^*_{i,t}$ for all players and pay-gates $t = 40, 60, 80$ by maximizing per-player expected revenue from the perspective of level zero (just before players start the game) under various constraints on the flexibility of prices across players and/or pay-gates.[20] We compare the relative performance of the following pricing strategies, ordered in terms of increasing discrimination from uniform to first-degree price discrimination (see Appendix 4.D.2 for the details):

- ☐ Observed. The observed pricing chosen by the firm, where each $p^*_{i,t}$ equals 70 virtual coins minus i's remaining endowment when facing pay-gate t.
- ☐ Uniform (70). All players face the same effective price $p^*_{i,t} = 70$. This amounts to setting everybody's endowment of virtual coins to zero from the beginning of the game.

[20] The fact that in our counterfactuals the firm *directly* chooses effective prices corresponds to restricting players' freedom to use the initial endowment of 70 virtual coins. Independently of i's endowment at pay-gate t, to progress in the game, i must use real money to purchase a key at the effective price of $p_{i,t}$ virtual coins. This greatly simplifies our model and simulations because we can proceed without specifying and estimating a further choice model for the allocation of the initial endowment of virtual coins. We believe this assumption is without loss of generality: the firm could always change any feature of the initial endowment, such as restricting the way players are allowed to use it, changing its magnitude (allowing players to have more or less than 70 virtual coins), or even removing it altogether (every player gets an endowment of zero virtual coins).

Table 4.7: Counterfactual Pricing Strategies, Effective Prices and Expected Revenues

Pricing Strategy	Static Pricing					Dynamic Pricing				
	Effective Price		Per-Player Revenue ($)			Effective Price		Per-Player Revenue ($)		
	mean	s.d.	mean	s.d.	%	mean	s.d.	mean	s.d.	%
Observed	35.566	34.529	0.011	0.108	-	-	-	-	-	-
Uniform (70)	70.000	-	0.022	0.012	93.8%	-	-	-	-	-
Uniform (Optimal)	45.000	-	0.049	0.024	340.0%	10.000	2.449	0.051	0.026	358.9%
GDP per Capita	44.500	1.500	0.049	0.025	340.7%	51.167	11.950	0.051	0.026	359.7%
Ability	44.500	3.841	0.049	0.025	343.7%	52.833	12.429	0.051	0.026	362.8%
Individual Level	45.166	5.065	0.050	0.025	346.9%	52.617	12.756	0.052	0.026	368.1%

Notes: This table summarizes our counterfactual simulation results in terms of effective prices and per-player expected revenues. Each row refers to a pricing strategy and summarizes the simulated effective prices chosen by the firm (in virtual coins, where $\$1 \approx 70$ virtual coins) and the corresponding per-player expected revenues (in $). The columns denoted by "%" report the percentage increase in per-player expected revenue implied by the row pricing strategy with respect to the observed pricing chosen by the firm (i.e., 0% means same average as the observed pricing). All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. The left panel summarizes results for the case in which effective prices do not change among pay-gates (static pricing). The right panel instead summarizes results for the case in which effective prices are allowed to change also among pay-gates (dynamic pricing). All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Table 4.8: Counterfactual Pricing Strategies, Consumer Surplus and Total Surplus

Pricing Strategy	Static Pricing				Dynamic Pricing			
	Δ Consumer Surplus ($)		Δ Total Surplus ($)		Δ Consumer Surplus ($)		Δ Total Surplus ($)	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
Uniform (70)	-0.0187	0.0137	-0.0083	0.0121	-	-	-	-
Uniform (Optimal)	-0.0369	0.0211	0.0008	0.0079	-0.0346	0.0203	0.0052	0.0097
GDP per Capita	-0.0376	0.0223	0.0002	0.0090	-0.0344	0.0211	0.0055	0.0105
Ability	-0.0383	0.0227	-0.0002	0.0069	-0.0334	0.0204	0.0068	0.0086
Individual Level	-0.0380	0.0241	0.0005	0.0089	-0.0336	0.0224	0.0072	0.0108

Notes: This table summarizes our counterfactual simulation results in terms of per-player consumer surplus and per-player total surplus, computed as the sum between changes in per-player expected revenue and in per-player consumer surplus. Each row refers to a pricing strategy and summarizes the simulated change in per-player consumer surplus and in per-player total surplus (both in $) with respect to the observed pricing. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. The left panel summarizes results for the case in which effective prices do not change among pay-gates (static pricing). The right panel instead summarizes results for the case in which effective prices are allowed to change also among pay-gates (dynamic pricing). All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

- Uniform (Optimal). The firm optimally chooses either only one effective price p^* for all players and pay-gates (static) or an effective price p_t^* common across players but specific to each pay-gate t (dynamic).

- GDP per Capita. Third-degree price discrimination based on the observed GDP per capita of a player's country. As for "Uniform (Optimal)," we consider both a static version in which the effective prices are identical across pay-gates and a dynamic version in which the effective prices are also allowed to change across pay-gates.

- Ability. Third-degree price discrimination based on the observed gaming ability of each player.[21] We again consider both a static (identical effective prices across pay-gates) and a dynamic version (potentially different effective prices across pay-gates).

- Individual Level. First-degree price discrimination where the firm is free to choose a different effective price for each player. We again consider both a static (identical effective prices across pay-gates) and a dynamic version (potentially different effective prices across pay-gates).

Tables 4.7 and 4.8 summarize our counterfactual simulation results. Table 4.7 reports results of counterfactual effective prices and expected revenues. Table 4.8 reports changes in consumer surplus and total surplus, computed as the sum between changes in expected revenues and in consumer surplus. Each row of Table 4.7

[21] As described in Section 4.3.1, we compute ability from each player's performance during the first 20 rounds of the game, something observed by the firm by the time the player reaches the first pay-gate at level 40.

refers to a pricing strategy and reports mean and standard deviation of the simulated effective prices chosen by the firm (in virtual coins) and of the corresponding per-player expected revenues (in $). The columns denoted by "%" report the percentage increase in per-player expected revenue implied by the row pricing strategy with respect to the observed pricing chosen by the firm (0% means same average as the observed pricing). Analogously, each row of Table 4.8 refers to a pricing strategy and summarizes the simulated change in per-player consumer surplus and in per-player total surplus (both in $) with respect to the observed pricing. Figures 4.15–4.18 visualize these results by plotting the simulated distributions of effective prices, per-player expected revenue, changes in per-player consumer surplus, and changes in per-player total surplus for the "static" pricing strategies considered in the left panels of Tables 4.7 and 4.8. Appendix Figures 4.F.1, 4.F.2, 4.F.6, and 4.F.11 plot analogous simulated distributions for the "dynamic" pricing strategies considered in the right panels of Tables 4.7 and 4.8.

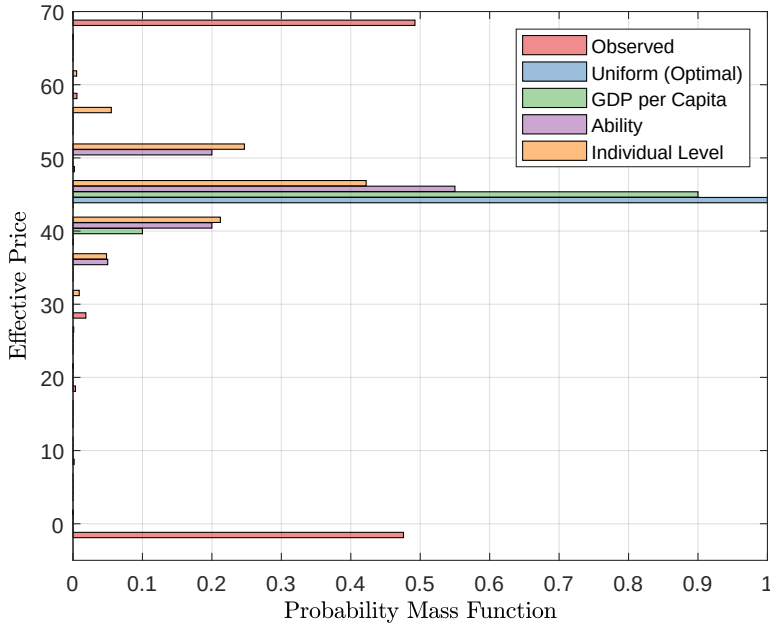


Figure 4.15: Distribution of Effective Prices in Static Pricing Strategies

Notes: This figure shows the simulated distribution of effective prices (in virtual coins, where $\$1 \approx 70$ virtual coins) across players and pay-gates for the "static" pricing strategies considered in the left panel of Table 4.7. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Our counterfactual simulation results accord with intuition: more flexible and discriminatory pricing strategies lead to larger per-player expected revenue at the expense of lower per-player consumer surplus. Uniform pricing is associated with lower per-player expected revenue than discriminatory pricing strategies. Similarly, dynamic pricing strategies that allow effective prices to vary across pay-gates lead to higher per-player expected revenue than their more restrictive static counterparts. While it is well known that price discrimination will in general enable a monopolist to seize larger portions of consumer surplus and thus increase profit at the expense of consumer surplus (Varian, 1989), Tables 4.7 and 4.8 provide three striking and perhaps less obvious insights.

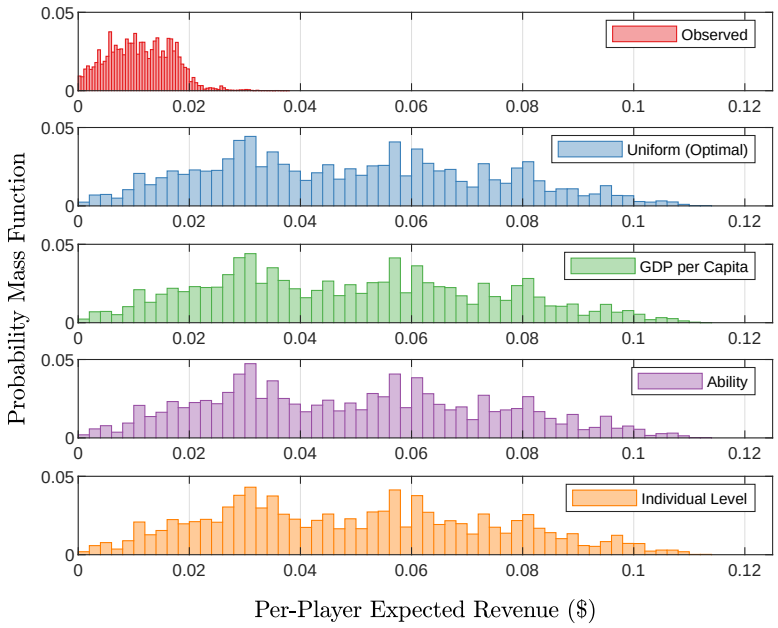
First, by comparing the first three rows of Table 4.7, from the firm's perspective the free endowment of 70 virtual coins given to players when they begin playing is too large. The resulting average observed effective

price of 35.6 virtual coins is too low. If the firm were to remove this free endowment and charge a uniform effective price of 70 virtual coins (as in "Uniform (70)," second row of Table 4.7), per-player expected revenue would almost double (93.8%). In addition, if it were also to optimally adjust the uniform effective price to 45 virtual coins (as in "Uniform (Optimal)," third row of Table 4.7), per-player expected revenue would more than quadruple (340%). Observed pricing seems far from maximizing profit and the firm could be substantially better-off by limiting itself to the use of uniform pricing. The fact that the firm may behave sub-optimally does not affect the validity of our counterfactual simulations. Indeed, because the marginal cost of each additional player is zero, we only rely on the estimation of demand for game content (which does not require any assumption about profit-maximization by the firm).[22]

Second, while more flexible and discriminatory pricing strategies lead to larger per-player expected revenue, the relative gains from their implementation are limited when compared to a simple uniform pricing strategy. This can be seen by comparing the third with the last three rows (fourth to sixth) of Table 4.7. Despite the different distributions of effective prices implied by each pricing strategy (Figure 4.15), the corresponding distributions of per-player expected revenues are remarkably similar (Figure 4.16), with the relative gains of static "Individual Level" ($45,2$ of average effective price) with respect to static "Uniform (Optimal)" (a unique effective price) being essentially negligible ($\$0.050 - \$0.049 = \$0.001$). Appendix Figure 4.F.3 further stresses this point: we construct 20 groups of players based on players' ability (left panel) and GDP per capita (right panel) and plot the average group-specific difference in per-player expected revenue. Remarkably, "Uniform (Optimal)" performs almost as well as any discriminatory pricing strategy not only on average (left panel of Table 4.7) but also conditional on player's ability and GDP per capita (Appendix Figure 4.F.3).

The right panel of Table 4.7 and Appendix Figures 4.F.1, 4.F.2, and 4.F.4 tell a similar story also for dynamic versions of these pricing strategies, underlying that even when the effective prices can change across pay-gates, "Uniform (Optimal)" (three prices $p^*_{40}, p^*_{60}, p^*_{80}$) still seizes most of the potential revenue of "Individual Level" (three prices $p^*_{i,40}, p^*_{i,60}, p^*_{i,80}$ for each $i = 1, \ldots, 43660$). Appendix Figure 4.F.5 then compares the relative gains of implementing a dynamic versus a static version of each pricing strategy by ability group and GDP per capita. Consistent with the findings that players behave myopically and that inter-pay-gates trade-offs may not be very relevant for the firm (Section 4.4.2 and Figures 4.12 and 4.13), Appendix Figure 4.F.5 shows that while dynamic pricing strategies slightly outperform their static counterparts, the implied relative gains are in practice very limited (note the smaller order of magnitude of the y-axis with respect to Appendix Figures 4.F.3 and 4.F.4).

[22] Deviations from profit-maximization instead represent a problem when the simulation of counterfactuals also requires the estimation of marginal cost functions, which typically hinges on the correct specification of the optimization problem solved by the firm (Berry et al., 1995; Nevo, 2001a).



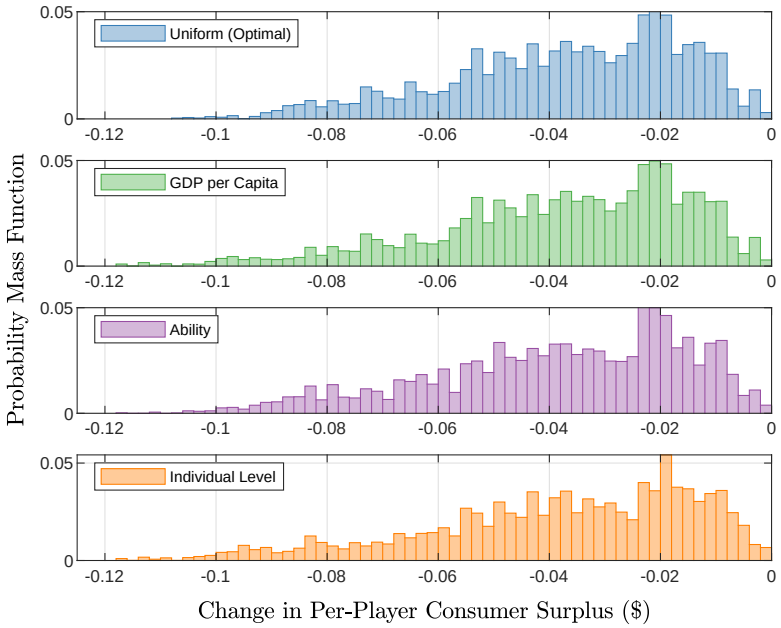
Notes: This figure shows the simulated distribution of per-player expected revenue (in $) across players for the "static" pricing strategies considered in the left panel of Table 4.7. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.16: Distribution of Per-Player Expected Revenue, Static Pricing Strategies

Mirroring these findings, Table 4.8 illustrates that each of the alternative pricing strategies would lead to a loss in per-player consumer surplus. This can be clearly seen in Figure 4.17, which shows that the distribution of changes in per-player consumer surplus associated to each alternative pricing strategy would always have negative support. This makes intuitive sense, in that each alternative pricing strategy would imply a higher average effective price than the observed one (of around 10 virtual coins, see left panel of Table 4.7), enabling the firm to extract more of the players' surplus. Importantly, mixed logit model (4.5.7) allows players to drop out of the game at any pay-gate t (by choosing $\text{buy}_{i,t} = 0$) if, for example, effective prices were "too high." In other words, this simulated extraction of consumer surplus is not conditional on the players being held "captive" in the game, but it is rather based on a more effective exploitation of their preferences. Appendix Figures 4.F.6–4.F.10 visualize additional dimensions of heterogeneity. Although also dynamic counterfactual pricing strategies induce losses in consumer surplus (Appendix Figure 4.F.6), they usually generate smaller decreases than their static counterparts (Appendix Figure 4.F.7).

Third, by summing the increases in expected revenue and the decreases in consumer surplus, Table 4.8 shows that, on average, the per-player total surplus implied by most of the alternative pricing strategies would be non-negative (with the exception of "Uniform (70)" and of the static version of "Ability"). On average, these counterfactual pricing strategies would generate enough additional expected revenue to compensate the corresponding loss in consumer surplus. As the right panel of Table 4.8 illustrates, the dynamic pricing strategies would perform slightly better than their static counterparts as a result of a slightly larger increase in expected revenue (Appendix Figure 4.F.5) and a slightly smaller decrease in consumer surplus (Appendix Figure 4.F.7). The fact that average per-player total surplus is non-negative suggests that these pricing strategies

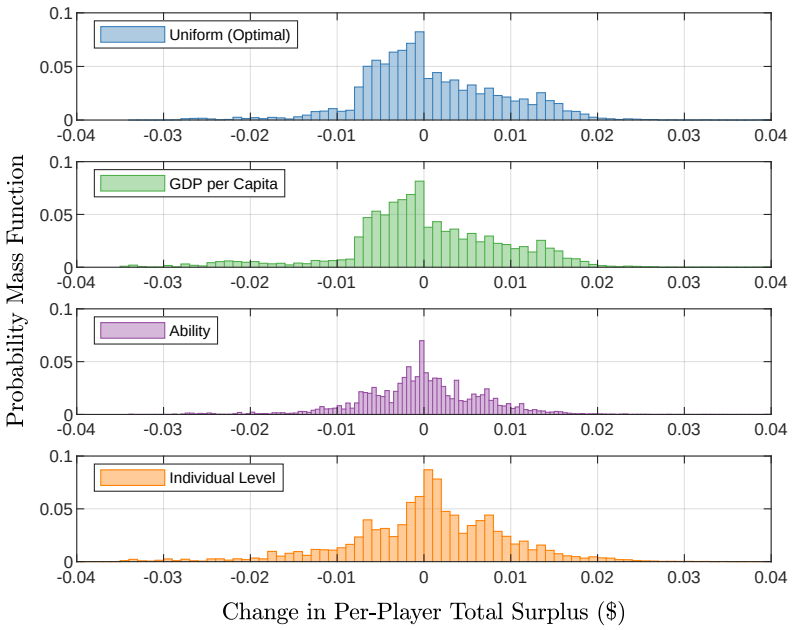
would not only enable the firm to extract more of the players' surplus, but that they would also not lead to sizeable dead-weight losses—despite the increase in average effective price of around 10 virtual coins. Figure 4.18 and Appendix Figure 4.F.11 highlight the distributional content of this result, stressing that—despite the non-negative average—, there would always be groups of players associated to negative changes in total surplus.



Notes: This figure shows the simulated distribution of changes in per-player consumer surplus (in $) across players for the "static" pricing strategies considered in the left panel of Table 4.7 as opposed to the observed pricing. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.17: Distribution of Δ Per-Player Consumer Surplus, Static versus Observed Pricing

To summarize, our counterfactual simulation results suggest that: (i) observed pricing is far from profit maximizing and leaves a lot of surplus in the hands of players, (ii) optimal uniform pricing would enable the firm to appropriate most of the returns associated to more complex pricing strategies, and (iii) each of the pricing strategies considered—including optimal uniform pricing—would induce a transfer of surplus from the players to the firm without, however, generating any sizeable dead-weight loss on average.



Notes: This figure shows the simulated distribution of changes in per-player total surplus (in $) across players for the "static" pricing strategies considered in the left panel of Table 4.8 as opposed to the observed pricing. Changes in per-player total surplus are computed as the sum between changes in per-player expected revenues and in per-player consumer surplus. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.18: Distribution of Δ Per-Player Total Surplus, Static versus Observed Pricing

Findings consistent with (i) have been documented in other industries: because of path-dependence, imperfect information, learning, or conflicting incentives, sometimes for-profit firms do not maximize profit (Cho and Rust, 2010; DellaVigna and Gentzkow, 2019; Dube and Misra, 2019; Fioretti, 2020; Hortaçsu et al., 2021; Huang et al., 2020; Orbach and Einav, 2007). Finding (ii) is in line with Chu et al. (2011), who show in the context of a theater company that simple pricing rules can sometimes generate almost as much profit as complex ones that would however be hard to implement. Finding (ii) is also close in spirit to Levitt et al. (2016), who document limited gains of second-degree price discrimination for a large online gaming firm, and more in general to the empirical literature on the trade-offs of price discrimination and personalized pricing in the era of big data (Rossi et al., 1996; Shiller and Waldfogel, 2011; Shiller, 2015; Waldfogel, 2015). Limited gains from price discrimination may partly explain why it is rarely observed in business practice, where additional risks tied to consumer backlash and regulatory scrutiny also need to be considered (Council of Economic Advisors, 2015; DellaVigna and Gentzkow, 2019).

In contrast to our results, however, Dube and Misra (2019) document substantial returns of personalized pricing for a digital recruiting firm, highlighting the need for caution in drawing general conclusions. While we do not find any such evidence, in other digital contexts more complex pricing strategies may be much more profitable. That being said, both our results and Dube and Misra (2019) stress the large potential of "empirical" pricing rules. In our context, the firm could increase per-player expected revenues more than fourfold by optimally choosing a uniform effective price on the basis of detailed data and appropriate empirical methods. Importantly, finding (iii) stresses that, although these increases in profit would necessarily come at the expense of decreases in consumer surplus, the pricing strategies considered would not generate average losses in total

welfare.

4.6.3 Robustness Checks

In Appendix 4.F.1, we repeat all counterfactual simulations accounting for the predictive biases of our estimated model as documented in Appendix 4.E. We do this by limiting our counterfactual simulations to the sub-sample of players in Group 40 for which the estimated model has the best predictive power. These checks show no qualitative difference in our results and suggest that theses predictive biases do not play a crucial role for our simulations exercises.

4.7 Conclusion

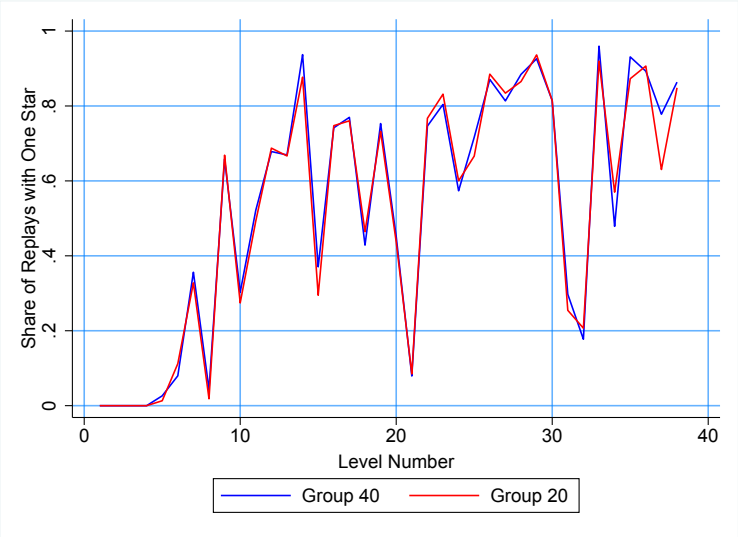
Our results indicate that the game developer can substantially increase profit by setting prices on the basis of readily available information on player characteristics and in-game behavior. As expected, the increase in profit would largely result from a transfer of surplus from players to game developer (Varian, 1989). However, most of the pricing strategies considered would not decrease total surplus on average. Our results also show that a simple uniform pricing strategy may already guarantee most of the profit implied by elaborate forms of price discrimination (Chu et al., 2011), which might help explain why price discrimination has been rarely used in online markets.

Our study analyzes price setting in a popular mobile game that, during the period of data collection, had a number of specific features. While these features and some of our modelling choices facilitate our empirical analysis, they may also limit the generality of our findings. First, no advertisement was shown in the game during data collection. With advertisement, the problem of the firm would differ, in that the firm could decide to trade-off revenue from in-app purchases in favor of consumption of game content, possibly by reducing the prices for in-app purchases. Second, our data show no evidence of a trade-off for the firm between revenue from pay-gates and other in-app purchases. Similar to advertisement, in other freemium apps this trade-off may be more prominent and lead to a more complex maximization problem for the firm. Third, similar to Dube and Misra (2019), our counterfactual simulations treat the firm as a monopolist and this may cause an overestimation of its market power when choosing prices for premium content. We believe this assumption to be appropriate in freemium games such as the one we study, where competition among mobile games occurs mostly before players download the game (for free) and then substantially softens after a player has downloaded and started to play (when the prices for premium content are incurred). However, this may be less applicable to non-freemium contexts in which competing firms charge positive prices already for the download of their apps.

4.A Assumptions

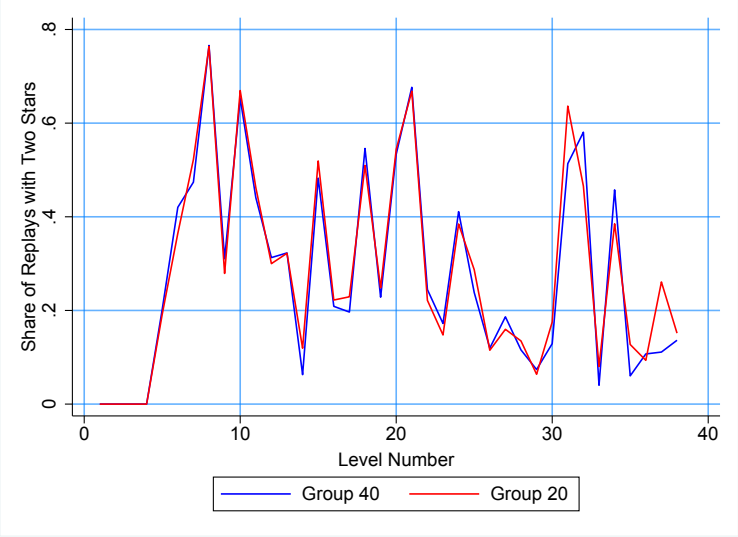
4.A.1 Further Evidence in Support of Assumption 1

Figure 4.A.1: Group 20 vs Group 40: Prob. to re-play level initially cleared with one star



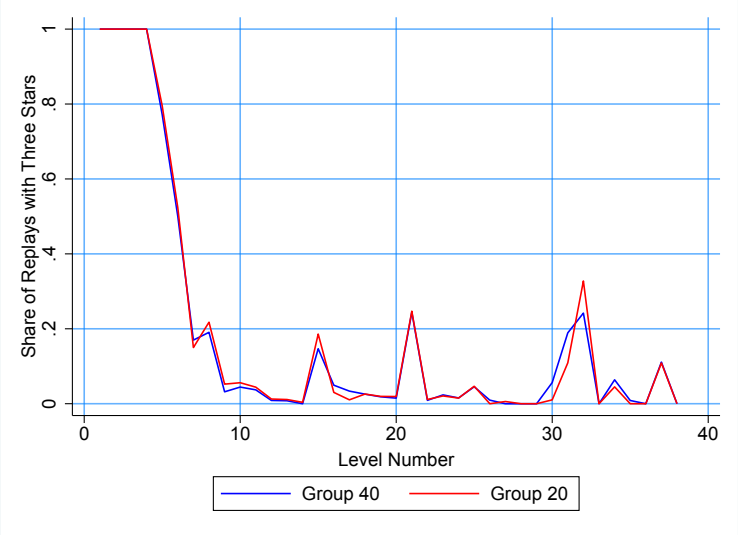
Notes: This figure compares the propensity to replay a level according to the number of available stars and according to being in Group 40 or 20. The y-axis variable is, for a given level ℓ, the share of replays which were done when only a single star had previously been collected at this level (i.e, there are two remaining stars). The sample includes players who have crossed pay-gate 20 but have no gone beyond pay-gate 40. Among these, we keep only players of Group 40 and 20 who hit pay-gate 20 with a non-positive star gap. The definition for stars is provided in Section 4.3.1.

Figure 4.A.2: Group 20 vs Group 40: Prob. to re-play level initially cleared with two stars



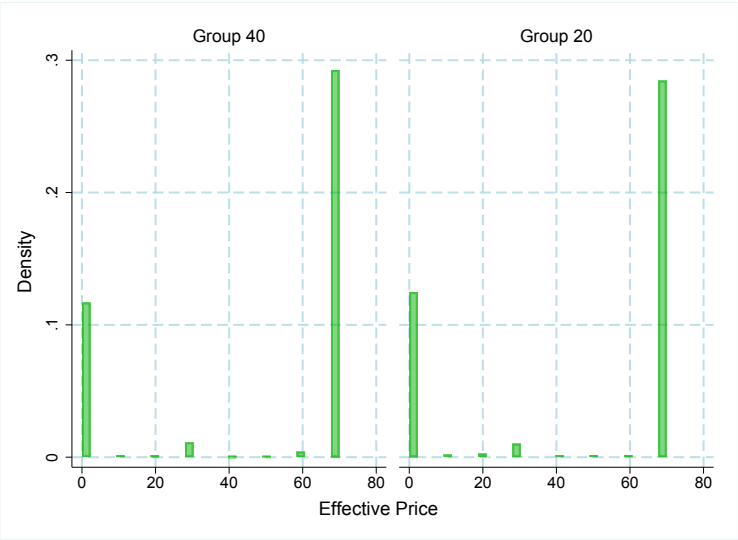
Notes: This figure compares the propensity to replay a level according to the number of available stars and according to being in Group 40 or 20. The y-axis variable is, for a given level ℓ, the share of replays which were done when two stars had previously been collected at this level (i.e, there are one remaining stars). The sample includes players who have crossed pay-gate 20 but have no gone beyond pay-gate 40. Among these, we keep only players of Group 40 and 20 who hit pay-gate 20 with a non-positive star gap. The definition for stars is provided in Section 4.3.1.

Figure 4.A.3: Group 20 vs Group 40: Prob. to re-play level initially cleared with three stars



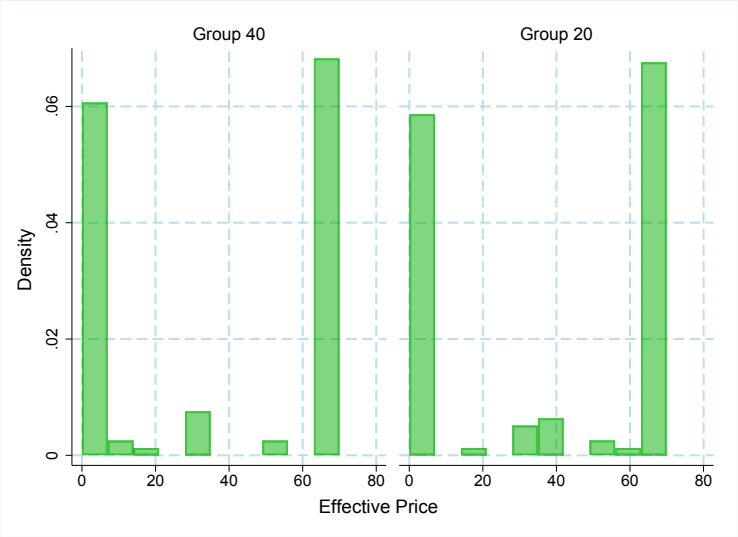
Notes: This figure compares the propensity to replay a level according to the number of available stars and according to being in Group 40 or 20. The y-axis variable is, for a given level ℓ, the share of replays which were done when three stars had previously been collected at this level (i.e, there are zero remaining stars). The sample includes players who have crossed pay-gate 20 but have no gone beyond pay-gate 40. Among these, we keep only players of Group 40 and 20 who hit pay-gate 20 with a non-positive star gap. The definition for stars is provided in Section 4.3.1.

Figure 4.A.4: Distribution of effective prices $p_{i,t}$ at pay-gate $t = 60$



Notes: This histogram displays the effective price distribution at pay-gate 60 for Group 20 and Group 40. Each observation is a player who faces pay-gate 60 with a positive star gap. The sample includes all players of Group 40 and 20 who satisfy this condition.

Figure 4.A.5: Distribution of effective prices $p_{i,t}$ at pay-gate $t = 80$



Notes: This histogram displays the effective price distribution at pay-gate 80 for Group 20 and Group 40. Each observation is a player who faces pay-gate 80 with a positive star gap. The sample includes all players of Group 40 and 20 who satisfy this condition.

4.A.2 Relationship between Pay-Gate Purchases and Non-Pay-Gate Purchases

This Appendix provides evidence that purchases outside of pay-gates (i.e., non-pay-gate purchases within the game) are not affected by purchases at pay-gates (i.e., purchases of keys to unlock pay-gates). This is important to justify our focus on the firm's revenue from purchases at pay-gates (see Section 4.4.1). The evidence presented in this appendix suggests that purchases at pay-gates do not crowd out non-pay-gate purchases. This allows us to analyze revenue from purchases at pay-gates separately from other in-game purchases.

We do this by relying on the experimental design described in Section 4.3.2.1 and used to test Assumption 1 in Section 4.4.2. In particular, we compare non-pay-gate purchases by players in Group 20 who faced an additional pay-gate at level 20 with those of players in the other experimental groups who did not. We consider players in Group 20 with a positive star gap at pay-gate $t = 20$ to be "treated" with an additional pay-gate. Players of Group 40 and No Star with a positive star gap at pay-gate $t = 20$ are instead considered as the control group, because they did not face any pay-gate $t = 20$. We implement this test using two alternative measures of payment outside of pay-gates. We define two alternative variables measuring the non-pay-gate purchases by players: "Accumulated Purchases" as the total number of purchases made by a player between level $\ell = 21$ and level $\ell = 39$ and "Indicator of Purchase" as a dummy equal to one if "Accumulated Purchases" is greater than zero.

We start by comparing non-pay-gate purchasing behaviors on the basis of t-tests. Table 4.A.1 compares the purchasing behavior of players in Group 20 versus players in No Star, while Table 4.A.2 compares Group 20 with Group 40. In both cases and for both variables, we find no statistically significant differences.

Table 4.A.1: T-test on Non-Pay-Gate Purchases: Group 20 VS No Star

	(1) No Star		(2) Group 20		(3) No Star−Group 20	
	mean	SE	mean	SE	diff.	t-test
Accumulated Purchases	0.061	0.671	0.053	0.506	-0.008	-1.207
Indicator of Purchase	0.023	0.151	0.023	0.149	-0.001	-0.352
Observations	39102		7409		46511	

Notes: This table presents evidence on whether facing an additional pay-gate affects the future non-pay-gate purchasing behavior of Group 20 players relative to No Star players. The variable "Accumulated Purchases" is the number of times a player made a purchase between levels $\ell = 21$ and $\ell = 39$. The variable "Indicator of Purchase" is instead a dummy equal to one if "Accumulated Purchases" is positive. The sample includes all 7,409 players in Group 20 and 39,102 in No Star who reached level 20 with a positive star gap, case in which players in Group 20 faced a pay-gate but those in No Star did not. Columns "No Star" and "Group 20" report the means and standard errors, respectively, for the players in No Star and Group 20. Column "No Star−Group 20" reports the difference in mean and associated t-test between the two previous columns. The t-test is calculated assuming unequal variances.

Table 4.A.2: T-test on Non-Pay-Gate Purchases: Group 20 VS Group 40

	(1) Group 40		(2) Group 20		(3) Group 40−Group 20	
	mean	SE	mean	SE	diff.	t-test
Accumulated Purchases	0.056	0.521	0.053	0.506	0.003	0.330
Indicator of Purchase	0.021	0.143	0.023	0.149	-0.002	-0.840
Observations	8252		7409		15661	

Notes: This table presents evidence on whether facing an additional pay-gate affects the future non-pay-gate purchasing behavior of Group 20 players relative to No Star players. The variable "Accumulated Purchases" is the number of times a player made a purchase between levels $\ell = 21$ and $\ell = 39$. The variable "Indicator of Purchase" is instead a dummy equal to one if "Accumulated Purchases" is positive. The sample includes all 7,409 players in Group 20 and 8,252 in Group 40 who reached level 20 with a positive star gap, case in which players in Group 20 faced a pay-gate but those in Group 40 did not. Columns "Group 40" and "Group 20" report the means and standard errors, respectively, for the players in Group 40 and Group 20. Column "No Star−Group 20" reports the difference in mean and associated t-test between the two previous columns. The t-test is calculated assuming unequal variances.

Next, we compare non-pay-gate purchasing behaviors on the basis of a non-parametric Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test detects whether the distribution of a variable differs between two samples. On the basis of the same variables and samples as above, Tables 4.A.3 and 4.A.2 report the Kolmogorov-Smirnov test results. For each variable, the first row assesses whether the players in No Star (or in Group 40) have smaller values than the players in Group 20. The second row instead performs the opposite comparison, assessing whether the players in Group 20 have smaller values than the players in No Star (or in Group 40). The third row "Combined K-S" is the overall test, which is the maximum between the previous two rows. In all cases and for both variables, we observe very high p-values suggesting the absence of any significant difference between the players in No Star (or in Group 40) and those in Group 20. Overall, based on our t-tests and Kolmogorov-Smirnov tests, we do not find evidence that facing the additional pay-gate at $t = 20$ affects players' non-pay-gate purchasing behavior between levels $\ell = 21$ and $\ell = 39$.

Table 4.A.3: Kolmogorov-Smirnov test on Non-Pay-Gate Purchases: Group 20 VS No Star

	Largest diff.	p-value
Accumulated Purchases		
No Star	0	1
Group 20	0,001121	0,984466
Combined K-S	0,001121	1
Indicator of Purchases		
No Star	0	1
Group 20	0,000667	0,994474
Combined K-S	0,000667	1

Notes: This table presents evidence on whether facing an additional pay-gate affects the future non-pay-gate purchasing behavior of Group 20 players relative to No Star players. The variable "Accumulated Purchases" is the number of times a player made a purchase between levels $\ell = 21$ and $\ell = 39$. The variable "Indicator of Purchase" is instead a dummy equal to one if "Accumulated Purchases" is positive. The sample includes all 7,409 players in Group 20 and 39,102 in No Star who reached level 20 with a positive star gap, case in which players in Group 20 faced a pay-gate but those in No Star did not. Columns "Largest diff." and "p-value" report, respectively, the largest difference and associated p-value for each row on the basis of the Smirnov-Kolmogorov test (where zero means "no difference"). The first row considers the largest difference between the players in No Star and in Group 20. The second row considers the largest difference between the players in Group 20 and in No Star. The last row considers the largest overall difference.

Table 4.A.4: Kolmogorov-Smirnov test on Non-Pay-Gate Purchases: Group 20 VS Group 40

	Largest Diff.	p-value
Accumulated Purchases		
Group 40	0,001967	0,970253
Group 20	-0,00125	0,987925
Combined K-S	0,001967	1
Indicator of Purchases		
Group 40	0,001967	0,970253
Group 20	0	1
Combined K-S	0,001967	1

Notes: This table presents evidence on whether facing an additional pay-gate affects the future non-pay-gate purchasing behavior of Group 20 players relative to Group 40 players. The variable "Accumulated Purchases" is the number of times a player made a purchase between levels $\ell = 21$ and $\ell = 39$. The variable "Indicator of Purchase" is instead a dummy equal to one if "Accumulated Purchases" is positive. The sample includes all 7,409 players in Group 20 and 8,252 in Group 40 who reached level 20 with a positive star gap, case in which players in Group 20 faced a pay-gate but those in Group 40 did not. Columns "Largest diff." and "p-value" report, respectively, the largest difference and associated p-value for each row on the basis of the Smirnov-Kolmogorov test (where zero means "no difference"). The first row considers the largest difference between the players in Group 40 and in Group 20. The second row considers the largest difference between the players in Group 20 and in Group 40. The last row considers the largest overall difference.

As a final piece of evidence, we estimate the correlation between pay-gate purchases and non-pay-gate purchases for players in Group 40 (i.e., those who play the standard version of the game). In Table 4.A.5, we regress measures of non-pay-gate purchases on a dummy variable equal to one if the player unlocked pay-gate t by purchasing a key. In the first column, we use as dependent variable the number of non-pay-gate purchases made by the player in the nineteen levels between the two subsequent pay-gates t and $t + 20$. In the second column, we instead use a dummy equal to one if the first dependent variable is greater than zero (i.e., the player makes at least one non-pay-gate purchase in these nineteen levels between pay-gate t and $t + 20$). Both regressions include player fixed effects and dropout-level fixed effects (these control for the specific levels at which players are observed to drop out of the game). Because of the player fixed effects, the sample includes only the players observed to reach at least two pay-gates with a positive star gap. In both regressions, the estimated coefficient is slightly positive but largely non-significant. Both regressions suggest

that once we control for a player's propensity to make in-app purchases (i.e., player fixed effects), pay-gate t purchases do not affect subsequent non-pay-gate purchases between levels $\ell = t + 1$ and $\ell = t + 19$.

	Nb. purchases in $t < \ell < t + 19$	Any purchase in $t < \ell < t + 19$
Purchase key at t	0.123	0.0256
	(0.178)	(0.0252)
Player FE	Yes	Yes
Dropout-level FE	Yes	Yes
Observations	2,438	2,438
Nb. Players	1,170	1,170

*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table 4.A.5: Correlation Between Non-Pay-Gate Purchases and Pay-Gate Purchases

Notes: This table presents evidence regarding the relationship between pay-gate purchases (i.e., purchases of keys to unlock pay-gates) and non-pay-gate purchases (i.e., other in-app purchases) for the players in Group 40 (i.e., those who play the standard version of the game). Each column presents OLS estimates of a measure of non-pay-gate purchases on a dummy variable equal to one if the player unlocked pay-gate t by purchasing a key. The first column reports estimation results for the first measure of non-pay-gate purchases: the number of non-pay-gate purchases made by the player in the nineteen levels between the two subsequent pay-gates t and $t+20$. The second column instead reports estimation results for the second measure: a dummy equal to one if the first measure is greater than zero (i.e., the player makes at least one non-pay-gate purchase in these nineteen levels between pay-gate t and and $t + 20$). Both regressions include player fixed effects and dropout-level fixed effects (these control for the specific levels at which players drop out of the game). Because of the player fixed effects, the sample includes the players in Group 40 observed to reach at least two pay-gates with a positive star gap. Standard errors are clustered at the player level.

4.B kNN Estimator of Model (4.5.1)

4.B.1 Theory

This description of the kNN procedure we use to estimate each of the two binary choice models in (4.5.1) is based on Altman (1992).

We observe an i.i.d. sample of data $\{X_i, Y_i\}$ for $i = 1, ..., N$, where X_i is a vector of explanatory variables and Y_i is a binary dependent variable taking values in $\{0, 1\}$. The objective is to estimate the probabilities $Pr(Y_i = 1|X_i)$ and $Pr(Y_i = 0|X_i) = 1 - Pr(Y_i = 1|X_i)$ associated with explanatory variables X_i without making parametric assumptions. To this end, we select a neighborhood $N(X_i)$ of points, with cardinality $k = |N(X_i)|$ around X_i and estimate the sample counter-parts of these probabilities as

$$\widehat{Pr}(Y_i = 1|X_i) = \frac{\sum_{k \in N(X_i)} Y_k}{k}. \tag{4.B.1}$$

The neighborhood of points $N(X_i)$ for each observed value of X_i depends on two features. First, the size of the neighborhood denoted generically by the integer $k \in [1, N]$. Second, the distance between any two points X_j and X_s, denoted by $d_{j,s}$, is calculated using a metric. Examples of such metrics include the Euclidean distance ($d_{j,s} = (X_j - X_s)(X_j - X_s)'$), the Mahalanobis distance ($d_{j,s} = (X_j - X_s)V^{-1}(X_j - X_s)'$ where V is the covariance matrix of the matrix X which stacks the vectors of differences), or more generally the Minkowski distance ($d_{j,s} = \left[\sum_{s=1}^{N} |X_j - X_s|^p\right]^{1/p}$ for $p \in \mathbb{N}$).[23] To make variables comparable, we standardize (by subtracting the mean and dividing by the standard deviation) each explanatory variable in X_i. This makes our analysis robust to scale and location distortions.

4.B.2 Implementation

To select a sufficient number of neighbors k and an appropriate metric d, we search through various possible combinations. In particular, we follow the approach taken in Mitchell (1997) and discussed in Mullin and Sukthankar (2000) by selecting the combination that provides the smallest 5-fold cross validation loss based on the Mean Squared Error $MSE(d, k) = \sum_{j=1}^{N} \left(Y_i - \hat{Y}_j\right)$, where $\hat{Y}_j$ is the model's predicted outcome. The final distance (d), number of neighbors (k), and 10-fold cross validation error rate for equation (4.5.1) are reported in Tables 4.B.1 and 4.B.2. We observe a 10-fold cross validation error rate between 15% and 30% which suggests the models are predicting relatively well the underlying probabilities.

Table 4.B.1: kNN Estimation : $Pr_{i,t}(sg_{i,t} > 0|i \to t)$ for Group 40

	Observations	k	Distance	Error Rate
Pay-gate 40	7,812	62	Standardized Euclidean	14%
Pay-gate 60	1,433	55	Standardized Euclidean	21%
Pay-gate 80	129	10	Hamming	23%

[23] In our implementation, other metrics include the correlation distance, the hamming distance, the cosine distance, the Chebychev, the Jaccard distance, and the Spearman distance.

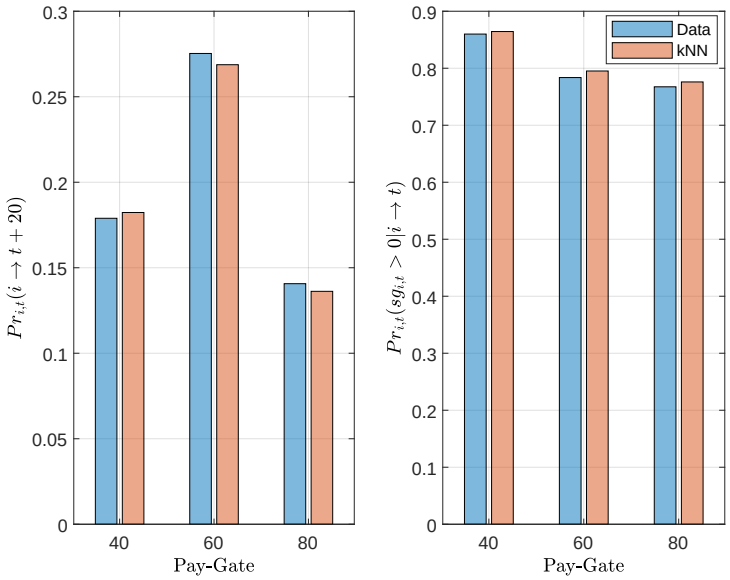
Table 4.B.2: kNN Estimation : $Pr_{i,t}(i \to t+20)$ for Group 40

	Observations	k	Distance	Error Rate
Pay-gate 40	43,660	256	Minkowski	18%
Pay-gate 60	5,205	1,996	Cityblock	28%
Pay-gate 80	917	374	Correlation	14%

4.B.3 Validation

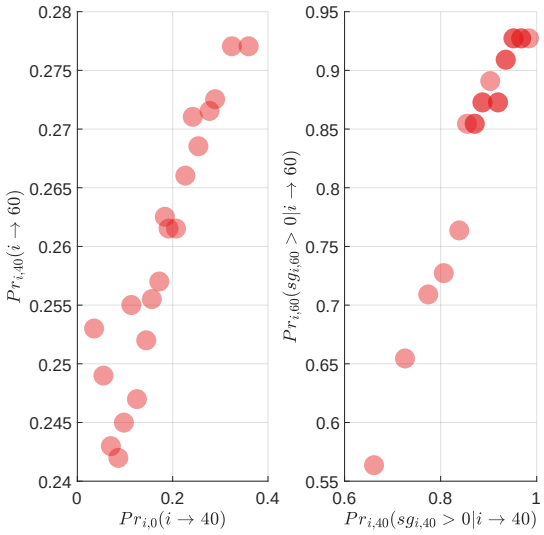
To validate our kNN estimates, we first compare their accuracy with respect to the underlying data. In Figure 4.B.1, we show that the kNN estimates match, on average, the empirical transition probabilities of equation (4.5.1). Moreover, as a sanity check, we display in Figure 4.B.2 the relationship across pay-gates between the estimated probabilities at pay-gate 40 and at pay-gate 60. As expected, players who are more likely to reach pay-gate 40 are also more likely to reach pay-gate 60 (conditional on unlocking pay-gate 40) and those who are more likely to have a positive star gap at pay-gate 40 (conditional on reaching pay-gate 40) are also more likely to have a positive star gap at pay-gate 60 (conditional on reaching pay-gate 60).

Figure 4.B.1: Comparing Observed and kNN Estimates of Probabilities in Model (4.5.1)



Notes: This figure displays the average probability of reaching the next pay-gate (left panel) and of having a positive star gap conditional on having reached a pay-gate (right panel). These probabilities are displayed based on the data (in blue) and based on the kNN estimates (in red) (described in the main text around equation (4.5.1) and above in this Appendix). The sample used is made of all players in Group 40 who have cleared the previous pay-gates.

Figure 4.B.2: Binned scatter plot of kNN Estimates of model (4.5.1) across Pay-gates



Notes: This figure displays the relationship between the probabilities at pay-gate 40 and 60 of reaching the next pay-gate (left panel) and having a positive star gap (right panel). These binned scatter plots rely on the kNN estimates described in the main text around equation (4.5.1) and above in this Appendix. The sample includes all players in Group 40.

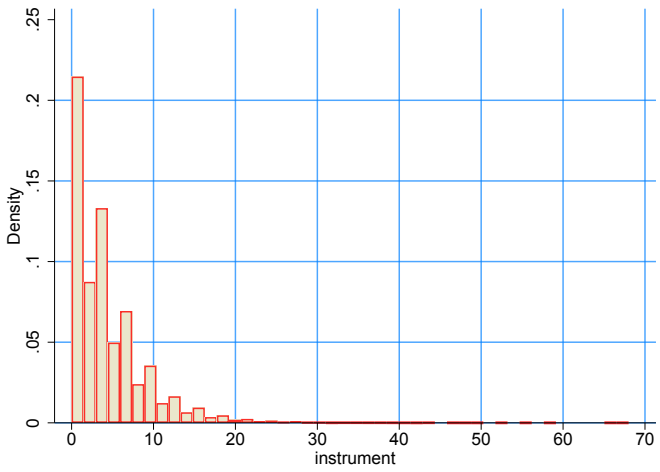
4.C Demand Estimates

4.C.1 First Step Estimates, Equation (4.5.5)

In this Appendix, we report the first step estimates of equation (4.5.5) and then assess the robustness of our instrument by considering two alternatives.

Figure 4.C.1 shows that the distribution of the instrument $Z_{i,t}$ based on equation (4.5.4) has fat tails, confirming the presence of wide sample variation.

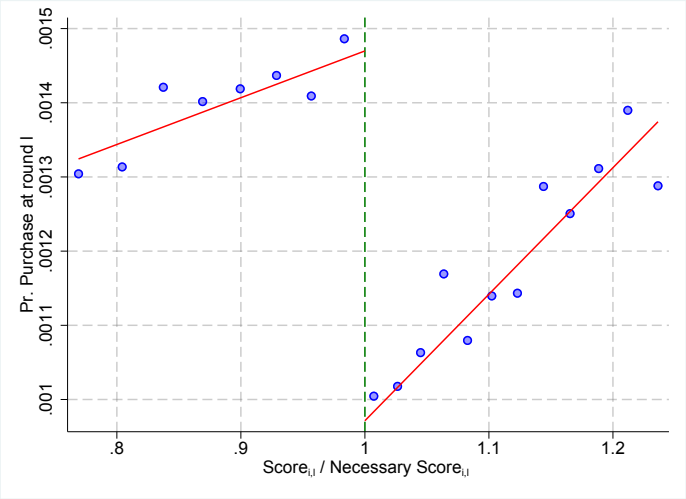
Figure 4.C.1: Histogram of the Instrument ($Z_{i,t}$)



Notes: This histogram displays the dispersion of the instrument $Z_{i,t}$ (based on equation (4.5.4)) used to estimate equation (4.5.5). The sample includes all observations in which a player from Group "No Star" faced a pay-gate.

Figure 4.C.2 displays a kink in the probability of a non pay-gate purchase when the player nears the necessary score cut-off, clarifying the type of variation leveraged by the proposed instrument.

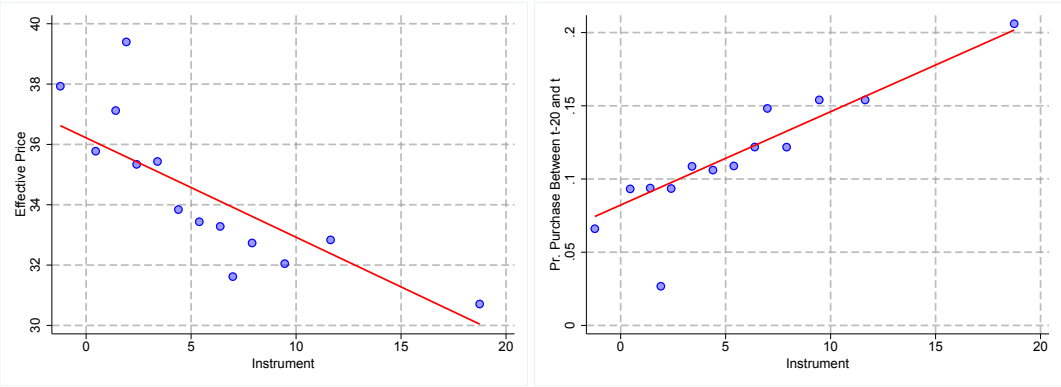
Figure 4.C.2: Binned Scatter Plot of Purchase Probability on the Instrument



Notes: This binned scatter plot displays the relationship between the probability of a purchase at a level ℓ against the instrument $Z_{i,t}$. The instrument $Z_{i,t}$ is constructed on the basis of equation (4.5.4) but adapted for visual inspection as the ratio between the player's score and the minimum required to pass the level attempted. On the y-axis, we plot the probability of making a purchase (outside of a pay-gate) for different values of the instrument on the x-axis. The sample includes all observations of all groups excluding observations corresponding to pay-gates.

Figure 4.C.3 confirms the intuition of the instrument: the left panel shows that the more often a player marginally failed a level, the more likely she is to face a lower effective price at the following pay-gate. The right panel illustrates that this is the result of players being pushed to purchase virtual coins to obtain additional lives or moves while trying to clear those challenging levels they marginally failed.

Figure 4.C.3: Binned Scatter Plot of Effective Price on the Instrument



Notes: These binned scatter plots display (left panel) the relationship between the average effective price $p_{i,t}$ against the instrument $Z_{i,t}$ and (right panel) the probability of purchasing virtual coins before reaching the next pay-gate against the instrument $Z_{i,t}$. The instrument $Z_{i,t}$ is constructed on the basis of equation (4.5.4). On the y-axis, we plot residualized averages (i.e, the average residual from a regression on pay-gate fixed effects) for various values of the instrument on the x-axis. The sample includes all observations in which players of Group "No Star" faced a pay-gate.

Table 4.C.1 reports the first step estimates of equation (4.5.5), which confirms the intuitive patterns from Figure 4.C.3. Conditional on controlling for pay-gate fixed effects and player-specific characteristics X_i (including i's ability), there is a negative and highly statistically significant relationship between the number of times a

player marginally failed a round of the game ($Z_{i,t}$) and the effective price ($p_{i,t}$). In particular, each marginal failure is found to lower the effective price by 0.457 virtual coins. The strength of the instrument is reflected in the large F-statistics. All in all, this Table suggests the instrument to be highly informative.

	(1) Effective Price	(2) Effective Price
Instrument $Z_{i,t}$ (δ)	-0.329***	-0.457***
	(0.0355)	(0.0359)
Pay-gate fixed effects (ζ_t)	Yes	Yes
Player-specific characteristics X_i (γ)	No	Yes
Observations	44,385	44,385
Num. of players	37,025	37,025
F-statistic	941.9	179.4

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.C.1: First-Step Estimation, Equation (4.5.5)

Notes: This table reports estimates of equation (4.5.5), a regression of the instrument $Z_{i,t}$ based on equation (4.5.4) on the effective price $p_{i,t}$. In the first column, we control for pay-gate fixed effects, while in the second we also control for player-specific characteristics X_i (described in Section 4.3.1). The specification in the second column is the one we use to construct the control function (equation (4.5.6)) for the estimation of mixed logit (4.5.7). The sample includes all observations in which a player in the Group "No Star" faced a pay-gate: 37,025 different players who faced a total of 44,385 pay-gates. Standard errors are clustered at the player level.

We now repeat the first step estimation of equation (4.5.5) by using two alternatives instruments. We denote by $Z_{i,t}^{(1)}$ our first alternative instrument, which also counts levels the player marginally cleared:

$$Z_{i,t}^{(1)} = \sum_{\ell=1}^{t} 1\left(0.95 \times \text{Necessary Score}_{\ell} < \text{Score}_{i,\ell} < 1.05 \times \text{Necessary Score}_{\ell}\right) \tag{4.C.1}$$

and by $Z_{i,t}^{(2)}$ the second alternative instrument, which decreases the threshold below which a player is considered to have marginally failed to pass a given level:

$$Z_{i,t}^{(2)} = \sum_{\ell=1}^{t} 1\left(0.90 \times \text{Necessary Score}_{\ell} < \text{Score}_{i,\ell} < \text{Necessary Score}_{\ell}\right). \tag{4.C.2}$$

Estimation results are reported in Table 4.C.2, which broadly confirm the robustness of the instrument $Z_{i,t}$ to alternative specifications. Varying the definition of the instrument does not qualitatively affect the negative and significant relationship with $p_{i,t}$, as also confirmed by the stability of the F-statistic across regressions. The magnitude of the estimated coefficient halves as a consequence of doubling the length of the interval used for the alternative instruments compared to $Z_{i,t}$.

4.C.2 Second Step Estimates, Equation (4.5.7)

Here we report the second step estimates of mixed logit model (4.5.7) and then assess their robustness using two alternative instruments described by equations (4.C.1) and (4.C.2) in Table 4.C.4.

Table 4.C.3 presents our main estimates of the parameters in equation (4.5.7). In terms of the utility of purchasing a key using virtual coins ($\text{buy}_{i,t} = 1$), we observe that the constant price coefficient α is negative and statistically significant. The coefficient θ on $\mu_{i,t}$, where $\theta\mu_{i,t}$ is the control function and $\mu_{i,t}$ is estimated using (4.5.5) (see Appendix 4.C.1), is positive and significant ($t \approx 11.28$)—confirming the presence of endogeneity in the effective prices.

	(1) Effective Price	(2) Effective Price	(3) Effective Price
Instrument $Z_{i,t}$ (δ)	-0.457***		
	(0.0359)		
Alternative instrument $Z_{i,t}^{(1)}$		-0.267***	
		(0.0196)	
Alternative instrument $Z_{i,t}^{(2)}$			-0.231***
			(0.0189)
Pay-gate fixed effects (ζ_t)	Yes	Yes	Yes
Player-specific characteristics X_i (γ)	Yes	Yes	Yes
Observations	44,385	44,385	44,385
Num. of players	37,025	37,025	37,025
F-statistic	179.4	181.3	178.8

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4.C.2: Robustness of First Step Estimation, Equation (4.5.5)

Notes: This table reports additional estimation results of equation (4.5.5), using alternative definitions of the instrument as in equations (4.C.1) and (4.C.2). In the first column, we report estimates using the basic definition of the instrument $Z_{i,t}$, while in the second and third we report estimates using the alternative definitions of the instrument $Z_{i,t}^{(1)}$ and $Z_{i,t}^{(2)}$. The sample includes all observations in which a player in the Group "No Star" faced a pay-gate: 37,025 different players who faced a total of 44,385 pay-gates. Standard errors are clustered at the player level.

We asses the robustness of these estimates using two alternative instruments. These instruments are described by equations (4.C.1) and (4.C.2) in Appendix 4.C.1. For each of these instruments, we re-estimate both the first step equation (4.5.5) and the mixed logit model (4.5.7). Table 4.C.4 reports our estimates. Compared with the estimates presented in Table 4.C.3, we do not observe any change in the signs across the different specifications for coefficients that are statistically significant. Both the constant associated with the price coefficient (α) and the control function (θ) are within a 95% confidence interval based on the estimates of Table 4.C.3 (i.e, respectively $[-1.01; -0.68]$ and $[-0.51; -0.36]$) suggesting limited differences across specifications. We conclude that our model estimates are robust to alternative specifications of the instrument for effective prices.

Table 4.C.3: Second Stage Estimates

Variable	Coefficient	Standard Error
Purchase (buy$_{i,t} = 1$)		
Intercepts (δ_1)		
Constant (δ_1)	64,620	6,365
Intercepts Pay-Gate Shifters ($\delta_{1,t}$)		
Pay-Gate 60 ($\delta_{1,60}$)	-4,680	0,628
Pay-Gate 80 ($\delta_{1,80}$)	4,054	0,132
Demographics (β_1)		
Australia and New Zealand	-0,452	0,043
Southern Asia	-1,455	0,170
Southern Europe	-0,112	0,088

Table 4.C.3: Second Stage Estimates

Variable	Coefficient	Standard Error
Sub Saharan Africa	-3,830	0,370
West Asia	-3,315	0,372
Eastern Asia	0,258	0,014
Eastern Europe	-0,302	0,022
Latin American and Caribbean	2,266	0,189
Northern Africa	-5,614	0,536
Northern America	-2,489	0,256
Northern Europe	-1,976	0,184
Other	-4,925	0,568
South Eastern Asia	0,334	0,052
Control Function (θ)	0,440	0,039
Log(GDP per Capita)	-1,768	0,179
iOS7	-2,225	0,126
iPad	1,487	0,116
Jailbroken	0,490	0,062
Ability	-3,929	0,350
Effective Price Intercept (α)		
Constant (α)	-0,852	0,083
Effective Price Pay-Gate Shifters		
Pay-Gate 60 (α_{60})	0,036	0,007
Pay-Gate 80 (α_{80})	0,053	0,008
Effective Price Demographic (π)		
Australia and New Zealand	0,017	0,001
Southern Asia	0,019	0,002
Southern Europe	0,004	0,001
Sub Saharan Africa	-0,014	0,000
West Asia	-0,015	0,001
Eastern Asia	-0,007	0,000
Eastern Europe	0,020	0,003
Latin American and Caribbean	0,022	0,002
Northern Africa	-0,222	0,017
Northern America	0,007	0,001
Northern Europe	0,005	0,000
Other	0,005	0,000
South Eastern Asia	0,018	0,002
Log(GDP per Capita)	0,020	0,002

Table 4.C.3: Second Stage Estimates

Variable	Coefficient	Standard Error
iOS7	-0,002	0,000
iPad	-0,001	0,000
Jailbroken	-0,018	0,002
Ability	0,006	0,001
Ask a Friend ($\text{buy}_{i,t} = 2$)		
Intercepts (δ_2)		
Constant (δ_2)	-43,650	6,696
Intercepts Pay-Gate Shifters ($\delta_{2,t}$)		
Pay-Gate 60 ($\delta_{2,60}$)	-8,078	1,100
Pay-Gate 80 ($\delta_{2,80}$)	-4,662	0,612
Demographics (β_2)		
Australia and New Zealand	-4,307	0,611
Southern Asia	-4,406	0,379
Southern Europe	-1,531	0,206
Sub Saharan Africa	-7,681	0,724
West Asia	-3,578	0,456
Eastern Asia	-6,415	0,865
Eastern Europe	-6,631	0,958
Latin American and Caribbean	-1,158	0,108
Northern Africa	0,998	0,312
Northern America	-5,635	0,810
Northern Europe	0,334	0,055
Other	-3,531	0,281
South Eastern Asia	3,459	0,544
Log(GDP per Capita)	4,458	0,668
iOS7	1,850	0,248
iPad	-1,509	0,223
Jailbroken	1,447	0,109
Ability	-0,348	0,036
Covariance Matrix		
$\sigma_{\delta_{2,i}}$	10,025	1,491
$\rho_{\delta_{2,i},\delta_{1,i}}$	0,432	0,081
$\sigma_{\delta_{1,i}}$	3,241	0,485
$\rho_{\delta_{2,i},\alpha_i}$	-0,043	0,004
$\rho_{\delta_{1,i},\alpha_i}$	-0,035	0,003
σ_{α_i}	-0,001	0,000

Notes: This table reports estimation results of equation (4.5.7). Standard errors are calculated using the method by Karaca-Mandic and Train (2003) to account for the two-step nature of the estimation procedure. The estimation procedure is described in Section 4.5.2 and uses 100 Halton draws per player as detailed by Bhat (2003). Each variable is defined in Section 4.3.1. The reference region is Western Europe. The sample includes all observations in which a player in the Group "No Star" faced a pay-gate: 37,025 different players who faced a total of 44,385 pay-gates.

Table 4.C.4: Robustness of Second Stage Estimates

	Alternative Instrument $Z_{i,t}^{(1)}$		Alternative Instrument $Z_{i,t}^{(2)}$	
Variable	Coefficient	Standard Error	Coefficient	Standard Error
Purchase ($\text{buy}_{i,t} = 1$)				
Intercepts (δ_1)				
Constant (δ_1)	69,907	4,767	57,527	4,329
Intercepts Pay-Gate Shifters ($\delta_{1,t}$)				
Pay-Gate 60 ($\delta_{1,60}$)	-5,653	0,470	-4,481	0,175
Pay-Gate 80 ($\delta_{1,80}$)	4,268	0,645	4,384	1,058
Demographics (β_1)				
Australia and New Zealand	-0,677	0,055	-0,643	0,119
Southern Asia	-1,748	0,109	-1,424	0,094
Southern Europe	-0,419	0,098	-0,181	0,059
Sub Saharan Africa	-4,257	0,252	-3,464	0,244
West Asia	-3,613	0,196	-2,716	0,128
Eastern Asia	0,061	0,035	-0,014	0,047
Eastern Europe	-0,312	0,064	-0,319	0,054
Latin American and Caribbean	2,141	0,571	2,024	0,337
Northern Africa	-5,905	0,542	-4,822	0,268
Northern America	-2,645	0,030	-1,983	0,095
Northern Europe	-2,168	0,106	-1,809	0,101
Other	-5,296	0,645	-4,658	0,241
South Eastern Asia	0,277	0,192	0,223	0,055
Control Function (θ)	0,483	0,060	0,418	0,047
Log(GDP per Capita)	-1,746	0,281	-1,406	0,055
iOS7	-2,452	0,239	-2,209	0,344
iPad	1,536	0,197	1,321	0,079
Jailbroken	0,601	0,070	0,513	0,037
Ability	-4,324	0,614	-3,644	0,351
Effective Price Intercept (α)				
Constant (α)	-0,913	0,006	-0,742	0,042
Effective Price Pay-Gate Shifters				
Pay-Gate 60 (α_{60})	0,045	0,001	0,030	0,005
Pay-Gate 80 (α_{80})	0,058	0,001	0,038	0,007
Effective Price Demographic (π)				
Australia and New Zealand	0,021	0,001	0,018	0,001
Southern Asia	0,019	0,011	0,011	0,001
Southern Europe	0,004	0,004	0,002	0,002
Sub Saharan Africa	-0,012	0,018	-0,012	0,003

Table 4.C.4: Robustness of Second Stage Estimates

	Alternative Instrument $Z_{i,t}^{(1)}$		Alternative Instrument $Z_{i,t}^{(2)}$	
Variable	Coefficient	Standard Error	Coefficient	Standard Error
West Asia	-0,015	0,005	-0,013	0,002
Eastern Asia	-0,006	0,010	-0,005	0,001
Eastern Europe	0,025	0,000	0,019	0,000
Latin American and Caribbean	0,025	0,005	0,017	0,001
Northern Africa	-0,297	2599	-0,623	5660246
Northern America	0,009	0,002	0,009	0,000
Northern Europe	0,006	0,001	0,005	0,001
Other	0,010	0,005	0,003	0,004
South Eastern Asia	0,017	0,003	0,011	0,001
Log(GDP per Capita)	0,019	0,010	0,015	0,000
iOS7	-0,003	0,000	-0,003	0,000
iPad	0,000	0,001	0,000	0,001
Jailbroken	-0,017	0,002	-0,015	0,001
Ability	0,007	0,002	0,004	0,000
Ask a Friend ($\text{buy}_{i,t} = 2$)				
Intercepts (δ_2)				
Constant (δ_2)	-65,547	28,772	-51,045	7,597
Intercepts Pay-Gate Shifters ($\delta_{2,t}$)				
Pay-Gate 60 ($\delta_{2,60}$)	-10,966	2,887	-8,690	0,982
Pay-Gate 80 ($\delta_{2,80}$)	-5,752	0,919	-4,531	0,219
Demographics (β_2)				
Australia and New Zealand	-6,424	2,581	-5,020	0,879
Southern Asia	-5,733	2,715	-4,081	3,227
Southern Europe	-2,080	0,645	-1,700	0,268
Sub Saharan Africa	-10,525	5,170	-9,185	5,016
West Asia	-5,090	1,872	-3,992	0,572
Eastern Asia	-8,789	2,244	-7,007	0,943
Eastern Europe	-9,015	1,214	-7,260	0,814
Latin American and Caribbean	-1,515	0,203	-1,301	0,335
Northern Africa	1,655	0,567	0,678	1,019
Northern America	-8,222	3,084	-6,457	1,013
Northern Europe	0,433	0,095	0,315	0,032
Other	-4,723	1,013	-3,657	0,426
South Eastern Asia	5,072	1,795	4,104	0,617
Log(GDP per Capita)	6,554	2,743	5,109	0,724
iOS7	2,622	0,876	2,114	0,343

Table 4.C.4: Robustness of Second Stage Estimates

	Alternative Instrument $Z_{i,t}^{(1)}$		Alternative Instrument $Z_{i,t}^{(2)}$	
Variable	Coefficient	Standard Error	Coefficient	Standard Error
iPad	-2,234	0,858	-1,752	0,318
Jailbroken	1,459	0,672	1,501	0,682
Ability	-0,469	0,176	-0,377	0,078
Covariance Matrix				
$\sigma_{\delta_{2,i}}$	15,174	5,826	11,948	2,230
$\rho_{\delta_{2,i},\delta_{1,i}}$	1,756	0,103	1,639	0,084
$\sigma_{\delta_{1,i}}$	-3,346	1,741	1,719	0,883
$\rho_{\delta_{2,i},\alpha_i}$	-0,061	0,004	-0,053	0,013
$\rho_{\delta_{1,i},\alpha_i}$	0,029	0,026	-0,013	0,024
σ_{α_i}	0,001	0,003	0,000	0,007

Notes: This table reports estimation results of equation (4.5.7). Standard errors are calculated using the method by Karaca-Mandic and Train (2003) to account for the two-step nature of the estimation procedure. The estimation procedure is described in Section 4.5.2 and uses 100 Halton draws per player as detailed by Bhat (2003). Each variable is defined in Section 4.3.1. The reference region is Western Europe. The sample includes all observations in which a player in the Group "No Star" faced a pay-gate: 37,025 different players who faced a total of 44,385 pay-gates.

4.D Price Elasticities and Counterfactual Simulations

4.D.1 Formulae

In this Appendix, we detail the formulae used to compute all our model predictions and simulations.

Price Elasticity of Demand. For player i at pay-gate t and given η_i, we refer to the multinomial logit formula as:

$$\text{MNL}_{i,t}(\eta_i) = \frac{\exp(V_{1,i,t}(\eta_i))}{1 + \exp(V_{1,i,t}(\eta_i)) + \exp(V_{2,i,t}(\eta_i))},$$

where, as described in Section 4.5.2, $V_{1,i,t}(\eta_i) = \delta_1 + \delta_{1,t} + \delta_{1,i} + X_i\beta_1 - (\alpha + \alpha_t + \alpha_i + X_i\pi)p_{i,t} + \theta\mu_{i,t}$ includes control function $\theta\mu_{i,t}$, based on (4.5.4), (4.5.5), and (4.5.6), to account for the potential endogeneity of $p_{i,t}$, and $V_{2,i,t}(\eta_i) = \delta_2 + \delta_{2,t} + \delta_{2,i} + X_i\beta_2$. Then, mixed logit model (4.5.7) implies the following price elasticity of demand:

$$\frac{p_{i,t}}{Pr_{i,t}(\text{buy}_{i,t} = 1|\text{lock}_{i,t} = 1, p_{i,t})} \frac{\partial Pr_{i,t}(\text{buy}_{i,t} = 1|\text{lock}_{i,t} = 1, p_{i,t})}{\partial p_{i,t}} = \tag{4.D.1}$$

$$\frac{-p_{i,t}(\alpha + \alpha_t + X_i\pi)}{\int \text{MNL}_{i,t}(\eta_i)\phi(\eta_i|\Sigma)d\eta_i} \int \alpha_i \text{MNL}_{i,t}(\eta_i)\left(1 - \text{MNL}_{i,t}(\eta_i)\right)\phi(\eta_i|\Sigma)d\eta_i,$$

where $\phi(\cdot|\Sigma)$ is the normal density of η_i in (4.5.3) with Σ denoting its variance-covariance matrix.

Per-Player Expected Revenue. Here we derive the formulae we use to compute the per-player expected revenue at level 0 in all our simulations with the exception of Figure 4.13 (which we instead discuss in the next sub-section). We calculate the per-player expected revenue at level 0 from pay-gate 40 (and none of the next pay-gates) as:

$$\mathbb{E}_{i,0}[R_{i,40}(p_{i,40})] = Pr_{i,0}(i \to 40, sg_{i,40} > 0) \times Pr_{i,40}(\text{buy}_{i,40} = 1|\text{lock}_{i,40} = 1, p_{i,40}) \times p_{i,40}$$

where $Pr_{i,0}(i \to 40, sg_{i,40} > 0)$ is i's probability of reaching pay-gate 40 with a positive star gap given that the player is at the beginning of the game, at level 0. This can be simply expressed in terms of the estimated probabilities in equation (4.5.1) as $Pr_{i,0}(i \to 40, sg_{i,40} > 0) = Pr_{i,40}(i \to 40, sg_{i,40} > 0)$, given that $t = 40$ is the first pay-gate i can encounter in the game. Per-player expected revenue at level 0 from pay-gate 60 (and none of the next pay-gates) is equal to:

$$\mathbb{E}_{i,0}[R_{i,60}(p_{i,40}, p_{i,60})] = Pr_{i,0}(i \to 60, sg_{i,60} > 0) \times Pr_{i,60}(\text{buy}_{i,60} = 1|\text{lock}_{i,60} = 1, p_{i,60}) \times p_{i,60}.$$

where $Pr_{i,0}(i \to 60, sg_{i,60} > 0)$ is i's probability of reaching pay-gate 60 with a positive star gap given that she is at level 0. We can again express this in terms of the estimated probabilities in models (4.5.1) and (4.5.7) as:

$$Pr_{i,0}(i \to 60, sg_{i,60} > 0) = Pr_{i,60}(i \to 60, sg_{i,60} > 0) \times Pr_{i,40}(i \to 40) \times$$
$$\left[Pr_{i,40}(sg_{i,40} > 0|i \to 40) \times (1 - Pr_{i,40}(\text{buy}_{i,40} = 0|\text{lock}_{i,40} = 1, p_{i,40}) + (1 - Pr_{i,40}(sg_{i,40} > 0|i \to 40))\right],$$

where $Pr_{i,60}(i \to 60, sg_{i,60} > 0)$ denotes i's probability of reaching pay-gate $t = 60$ with a positive star gap given that she unlocked pay-gate $t = 40$, and so on for the other probabilities. Similarly, we can write the

per-player expected revenue at level 0 from pay-gate 80 as:

$$\mathbb{E}_{i,0}[R_{i,80}(p_{i,40}, p_{i,60}, p_{i,80})] = Pr_{i,0}(i \to 80, sg_{i,80} > 0) \times Pr_{i,80}(\text{buy}_{i,80} = 1|\text{lock}_{i,80} = 1, p_{i,80}) \times p_{i,80}.$$

where $Pr_{i,0}(i \to 80, sg_{i,80} > 0)$ is i's probability of reaching pay-gate 80 with a positive star gap given that she is at level 0. This can be expressed in terms of the estimated probabilities in models (4.5.1) and (4.5.7) as:

$$\begin{aligned} Pr_{i,0}(i \to 80, sg_{i,80} > 0) = Pr_{i,80}(i \to 80, sg_{i,80} > 0) \times \\ Pr_{i,40}(i \to 40) \times \left[Pr_{i,40}(sg_{i,40} > 0 | i \to 40) \times (1 - Pr_{i,40}(\text{buy}_{i,40} = 0|\text{lock}_{i,40} = 1, p_{i,40}) + (1 - Pr_{i,40}(sg_{i,40} > 0 | i \to 40))\right] \times \\ Pr_{i,60}(i \to 60) \times \left[Pr_{i,60}(sg_{i,60} > 0 | i \to 60) \times (1 - Pr_{i,60}(\text{buy}_{i,60} = 0|\text{lock}_{i,60} = 1, p_{i,60}) + (1 - Pr_{i,60}(sg_{i,60} > 0 | i \to 60))\right]. \end{aligned}$$

Finally, we calculate the per-player expected revenue from player i at level 0 (from all pay-gates) given effective prices $p_i = (p_{i,40}, p_{i,60}, p_{i,80})$ as:

$$R_{i,0}(p_i) = \mathbb{E}_{i,0}[R_{i,40}(p_i) + R_{i,60}(p_i) + R_{i,80}(p_i)]. \tag{4.D.2}$$

This is the central expression at the basis of our counterfactual simulations, i.e. what the firm maximizes when choosing effective prices, and the main focus of the simulation method described in Appendix 4.D.2.

Revenue Decomposition in Figure 4.13. Here we derive the formulae used in Figure 4.13. While the computation of $Pr_{i,40}(\text{buy}_{i,40} = 1|\text{lock}_{i,40} = 1, p_{i,40}) \times p_{i,40}$ is immediate from the estimates of model (4.5.7), the computation of $(1 - Pr_{i,40}(\text{buy}_{i,40} = 0|\text{lock}_{i,40} = 1, p_{i,40})) \times R_{i,60}(p_{i,60}|p_{i,80})$ requires the calculation of $R_{i,60}(p_{i,60}|p_{i,80})$. This in turn can be expressed as a function of the probabilities estimated in models (4.5.1) and (4.5.7) as:

$$\begin{aligned} R_{i,60}(p_{i,60}|p_{i,80}) = Pr_{i,40}(i \to 60, sg_{i,60} > 0) \times Pr_{i,60}(\text{buy}_{i,60} = 1|\text{lock}_{i,60} = 1, p_{i,60}) \times p_{i,60} \\ + Pr_{i,60}(i \to 60) \times \left[Pr_{i,60}(sg_{i,60} > 0 | i \to 60) \times (1 - Pr_{i,60}(\text{buy}_{i,60} = 0|\text{lock}_{i,60} = 1, p_{i,60}) + (1 - Pr_{i,60}(sg_{i,60} > 0 | i \to 60))\right] \\ \times Pr_{i,80}(i \to 80, sg_{i,80} > 0) Pr_{i,80}(\text{buy}_{i,80} = 1|\text{lock}_{i,80} = 1, p_{i,80}) \times p_{i,80}, \end{aligned}$$

Per-Player Consumer Surplus. Here we derive the formulae we use to compute changes in per-player consumer surplus at level 0 in all our simulations. The derivations follow closely those for the per-player expected revenue above and here we rely on some of the objects defined there. We calculate the per-player consumer surplus at level 0 from pay-gate 40 (and none of the next pay-gates) as:

$$\mathbb{E}_{i,0}[CS_{i,40}(p_{i,40})] = Pr_{i,0}(i \to 40, sg_{i,40} > 0) \times CS_{i,40}(\text{lock}_{i,40} = 1|X_i, p_{i,40}).$$

Based on standard formulae for mixed logit models (Train, 2009), the per-player consumer surplus at level 40 from pay-gate 40 is given by:

$$CS_{i,40}(\text{lock}_{i,40} = 1|X_i, p_{i,40}) = C_{i,40} + \mathbb{E}_{i,40}\left(\frac{\ln(1 + \exp(V_{1,i,40}(\eta_i)) + \exp(V_{2,i,40}(\eta_i)))}{(\alpha + \alpha_{40} + \alpha_i + X_i\pi)}\right)$$

where $C_{i,40}$ is an unknown player-specific constant. Similarly, per-player consumer surplus at level 0 from pay-gate 60 (and none of the next pay-gates) is equal to:

$$\mathbb{E}_{i,0}[CS_{i,60}(p_{i,40}, p_{i,60})] = Pr_{i,0}(i \to 60, sg_{i,60} > 0) \times CS_{i,60}(\text{lock}_{i,60} = 1|X_i, p_{i,60}),$$

where

$$CS_{i,60}(\text{lock}_{i,60} = 1|X_i, p_{i,60}) = C_{i,60} + \mathbb{E}_{i,60}\left(\frac{\ln(1 + \exp(V_{1,i,60}(\eta_i)) + \exp(V_{2,i,60}(\eta_i)))}{(\alpha + \alpha_{60} + \alpha_i + X_i\pi)}\right).$$

In turn, we can write the per-player consumer surplus at level 0 from pay-gate 80 as:

$$\mathbb{E}_{i,0}[CS_{i,80}(p_{i,40}, p_{i,60}, p_{i,80})] = Pr_{i,0}(i \to 80, sg_{i,80} > 0) \times CS_{i,80}(\text{lock}_{i,80} = 1|X_i, p_{i,80}).$$

where

$$CS_{i,80}(\text{lock}_{i,80} = 1|X_i, p_{i,80}) = C_{i,80} + \mathbb{E}_{i,80}\left(\frac{\ln(1+\exp(V_{1,i,80}(\eta_i)) + \exp(V_{2,i,80}(\eta_i)))}{(\alpha + \alpha_{80} + \alpha_i + X_i\pi)}\right).$$

We calculate the per-player consumer surplus at level 0 (from all pay-gates) given effective prices $p_i = (p_{i,40}, p_{i,60}, p_{i,80})$ as:

$$CS_{i,0}(p_i) = \mathbb{E}_{i,0}[CS_{i,40}(p_{i,40}) + CS_{i,60}(p_{i,60}) + CS_{i,80}(p_{i,80})]. \tag{4.D.3}$$

Finally, for any given two vectors of effective prices p_i and p_i', we compute the associated change in per-player consumer surplus simply as:

$$\Delta CS_{i,0}(p_i, p_i') = CS_{i,0}(p_i) - CS_{i,0}(p_i').$$

Note that this difference at the player-level has the important advantage of removing, for each i, the unknown constants $C_{i,40}$, $C_{i,60}$, and $C_{i,80}$.

4.D.2 Simulation Method

Here we describe our simulation procedure both in the case of the observed pricing strategy chosen by the firm and in the case of the counterfactual pricing strategies we investigate. In general, to simulate the model, one needs to specify each player's effective price $p_{i,t}$ and corresponding residual $\mu_{i,t}$ from equation (4.5.5) at each possible pay-gate t.

Observed Pricing Strategy. Even in the case of the observed pricing strategy, because some players may have dropped out before reaching pay-gate 80, we cannot back out $(p_{i,t}, \mu_{i,t})$ directly from the data for all players and pay-gates. To address this missing data problem, we follow the approach by Jacobi and Sovinsky (2016) and treat the unobserved $(p_{i,t}, \mu_{i,t})$ as random effects to be integrated over their empirical distribution.

We are interested in simulating the firm's expected revenue (4.D.3) from player i before they start to play (at $t = 0$) for any given vector of effective prices $p_i = (p_{i,40}, p_{i,60}, p_{i,80})$ and corresponding residuals from equation (4.5.5), $\mu_i = (\mu_{i,40}, \mu_{i,60}, \mu_{i,80})$. To stress the dependence on both p_i and μ_i, due to the control function in mixed logit model (4.5.7), we extend the notation of per-player expected revenue (4.D.3) to explicitly account also for μ_i, $R_{i,0}(p_i, \mu_i)$. For those players who did not reach pay-gate t, we cannot directly back out $(p_{i,t}, \mu_{i,t})$ but assume that it follows the same empirical distribution $\hat{F}_{p,\mu,t}$ as among those players observed to reach pay-gate t. In particular, we compute the joint distribution $\hat{F}_{p,\mu,t}$ as $\hat{F}_{p,t}\hat{F}_{\mu|p,t}$, the product of the unconditional distribution of $p_{i,t}$ and the conditional distribution of $\mu_{i,t}$ given $p_{i,t}$. We then approximate the per-player expected revenue

$$\mathbb{E}_{\hat{F}_{p,\mu}}[R_{i,0}(p_i, \mu_i)] = \int R_{i,0}(p_i, \mu_i) d\hat{F}_{p,\mu,40}(p_{i,40}, \mu_{i,40}) d\hat{F}_{p,\mu,60}(p_{i,60}, \mu_{i,60}) d\hat{F}_{p,\mu,80}(p_{i,80}, \mu_{i,80}) \tag{4.D.4}$$

by taking 10,000 draws of (p_d, μ_d) from $\hat{F}_{p,\mu} = \hat{F}_{p,\mu,40}\hat{F}_{p,\mu,60}\hat{F}_{p,\mu,80}$ for each i and computing the average:

$$\hat{\mathbb{E}}_{\hat{F}_{p,\mu}}[R_{i,0}(p_i, \mu_i)] = \frac{1}{10,000} \times \sum_{d=1}^{10,000} R_{i,0}(p_d, \mu_d), \tag{4.D.5}$$

where $R_{i,0}(p_i, \mu_i)$ is derived above in Appendix 4.D.1, equation (4.D.3).

Counterfactual Pricing Strategies. In each of the counterfactual pricing strategies described in Section 4.6.2 and detailed below, the firm chooses effective prices so to maximize the sum of per-player expected revenue $R_{i,0}(p_i, \mu_i)$ in equation (4.D.3) (see Appendix 4.D.1) across players subject to some constraints. To simulate these counterfactuals, we assume that μ_i is also unobserved to firm and that its distribution is invariant to the specific pricing strategy used. More precisely, we assume that $\mu_{i,t}$ follows the same empirical distribution $\hat{F}_{\mu,t}$ as among those players observed to reach pay-gate t and that the firm uses $\hat{F}_{\mu,t}$ to form expectations with respect to $\mu_{i,t}$. We then approximate the per-player expected revenue for a given vector of effective prices p_i

$$\mathbb{E}_{\hat{F}_\mu}\left[R_{i,0}(p_i, \mu_i)\right] = \int R_{i,0}(p_i, \mu_i) d\hat{F}_{\mu,40}(\mu_{i,40}) d\hat{F}_{\mu,60}(\mu_{i,60}) d\hat{F}_{\mu,80}(\mu_{i,80}) \tag{4.D.6}$$

by taking 10,000 draws of μ_d from $\hat{F}_\mu = \hat{F}_{\mu,40}\hat{F}_{\mu,60}\hat{F}_{\mu,80}$ for each i and computing the average:

$$\hat{\mathbb{E}}_{\hat{F}_\mu}\left[R_{i,0}(p_i, \mu_i)\right] = \frac{1}{10,000} \times \sum_{d=1}^{10,000} R_{i,0}(p_i, \mu_d). \tag{4.D.7}$$

Alternative Pricing Strategies Considered. Here we detail the optimization problem of the firm in the simulation of each of the counterfactual pricing strategies described in Section 4.6.2. All counterfactuals are computed for the $44,660$ players in Group 40 (those who play the standard version of the game) over pay-gates 40, 60, and 80.

☐ Uniform (Optimal) Static Pricing:

$$(p^*, p^*, p^*) = \arg\max_{p} \sum_{i=1}^{44,660} \hat{\mathbb{E}}_{\hat{F}_\mu}\left[R_{i,0}(p, p, p, \mu_i)\right].$$

☐ Uniform (Optimal) Dynamic Pricing:

$$(p^*_{40}, p^*_{60}, p^*_{80}) = \arg\max_{p_{40}, p_{60}, p_{80}} \sum_{i=1}^{44,660} \hat{\mathbb{E}}_{\hat{F}_\mu}\left[R_{i,0}(p_{40}, p_{60}, p_{80}, \mu_i)\right].$$

☐ GDP per Capita Static Pricing for players in ventile $G = 1, \ldots, 20$ of GDP per capita:

$$(p^*_G, p^*_G, p^*_G) = \arg\max_{p_G} \sum_{i \in G} \hat{\mathbb{E}}_{\hat{F}_\mu}\left[R_{i,0}(p_G, p_G, p_G, \mu_i)\right],$$

where each ventile G gathers 5% of players in terms of the observed distribution of GDP per capita.

☐ GDP per Capita Dynamic Pricing for players in ventile $G = 1, \ldots, 20$ of GDP per capita:

$$(p^*_{G,40}, p^*_{G,60}, p^*_{G,80}) = \arg\max_{p_{G,40}, p_{G,60}, p_{G,80}} \sum_{i \in G} \hat{\mathbb{E}}_{\hat{F}_\mu}\left[R_{i,0}(p_{G,40}, p_{G,60}, p_{G,80}, \mu_i)\right],$$

where each ventile G gathers 5% of players in terms of the observed distribution of GDP per capita.

☐ Ability Static Pricing for players in ventile $A = 1, \ldots, 20$ of ability:

$$(p^*_A, p^*_A, p^*_A) = \arg\max_{p_A} \sum_{i \in A} \hat{\mathbb{E}}_{\hat{F}_\mu}\left[R_{i,0}(p_A, p_A, p_A, \mu_i)\right],$$

where each ventile A gathers 5% of players in terms of the observed distribution of ability.

☐ Ability Dynamic Pricing for players in ventile $A = 1, \ldots, 20$ of ability:

$$(p^*_{A,40}, p^*_{A,60}, p^*_{A,80}) = \arg\max_{p_{A,40}, p_{A,60}, p_{A,80}} \sum_{i \in A} \hat{\mathbb{E}}_{\hat{F}_\mu}\left[R_{i,0}(p_{A,40}, p_{A,60}, p_{A,80}, \mu_i)\right],$$

where each ventile A gathers 5% of players in terms of the observed distribution of ability.

□ Individual Static Pricing for player $i = 1, ..., 43660$:

$$(q_i^*, q_i^*, q_i^*) = \arg\max_{q_i} \hat{\mathbb{E}}_{\hat{F}_\mu} \left[R_{i,0}(q_i, q_i, q_i, \mu_i) \right].$$

□ Individual Dynamic Pricing for player $i = 1, ..., 43660$:

$$p_i^* = \arg\max_{p_i} \hat{\mathbb{E}}_{\hat{F}_\mu} \left[R_{i,0}(p_i, \mu_i) \right].$$

For simplicity, we solve each of these optimization problems using a simple grid search over effective prices. For each effective price the firm can choose, we specify a grid with intervals of 5 virtual coins going from 0 to 100, $[0, 5, 10, ..., 95, 100]$. For example, in Uniform (Optimal) Static Pricing this results in 21 possible combinations of effective prices, while in Individual Dynamic Pricing in $21^3 = 9,261$ combinations for each player $i = 1, ..., 43660$. We then evaluate the per-player expected revenue for each combination of effective prices and player and solve the above optimization problems. We do not extend the support of the optimizations above 100 virtual coins as we never found any optimal effective price to be larger than 70 virtual coins (which is also the maximum value we observe in the data). The step size of 5 virtual coins was selected as a trade-off between precision and required computational time.

4.E Model Validation

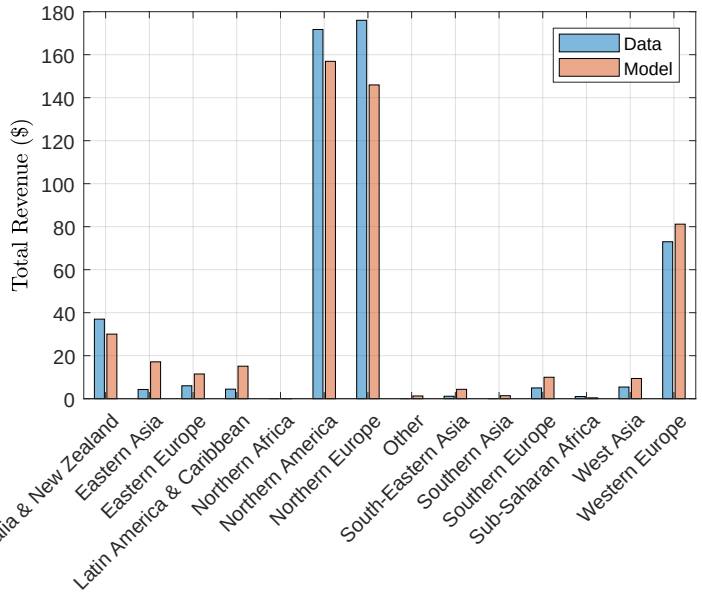
Our simulation exercises rely on the estimated model's ability to predict player behavior under counterfactual pricing strategies. In this section, we investigate the predictive power of the estimated model in terms of a player's expected revenue. As mentioned at the beginning of Section 4.6, while we relied on the players in Group 40 for the estimation of model (4.5.1) and on those in No Stars for the estimation of model (4.5.7), we test the model's predictive power and perform all counterfactual simulations only with respect to the players in Group 40. Players in Group 40 face the default design of the game, which corresponds to our discrete choice model in equation (4.5.1).

As can be seen in equation (4.4.3), to compute the per-player expected revenue at pay-gate t, we need to know the effective prices they would face at t and later pay-gates. However, for those players that drop out of the game before pay-gate t, we do not observe these effective prices. We address this "missing data" problem as in Jacobi and Sovinsky (2016) and treat the effective prices as another dimension of unobserved heterogeneity to be integrated over when calculating expectations. We assume the true t-specific distribution of $p_{i,t}$ across players, $F_t(p_{i,t})$, can be consistently estimated as the empirical distribution of the observed effective prices at t, $\hat{F}_t(p_{i,t})$, and then integrate per-player expected revenue over $\hat{F}_t(p_{i,t})$ for each i. In Appendix 4.D.1 we report the formulae used to compute per-player expected revenue (used also in the counterfactual simulations) and in Appendix 4.D.2 we describe the details of this simulation procedure (and of the procedure used for the counterfactual simulations). Below we report our validation results comparing the per-player average observed revenue with its counterpart as predicted by the estimated model.

The estimated model is very good at predicting the average observed per-player revenue of $\$0.011$ (from purchases of keys to unlock pay-gates), delivering a t-test as small as -0.09249. Note that this result is not mechanical, in that the estimated parameters are not chosen to minimize the distance between observed and predicted revenue, but rather the probabilities of models (4.5.1) and (4.5.7). Importantly for the investigation of price discrimination, Figures 4.E.1 and 4.E.2 illustrate the accuracy of the estimated model in predicting revenues for specific profiles of the observed player-specific characteristics X_i.

Figure 4.E.1 compares the total revenue by geographical region (in $) as observed in the data against that as predicted by the estimated model across different geographical regions. We calculate observed total revenue as the sum over 43,660 players in Group 40 of the revenue collected at pay-gates 40, 60, and 80 in each geographical region during the 15 days of our sample in 2013. The figure confirms the presence of geographical heterogeneity in total revenue and that the estimated model is good at capturing it. In general, the geographical regions for which more revenue is observed (and so for which we have more observations) are also those for which the estimated model delivers more accurate predictions.

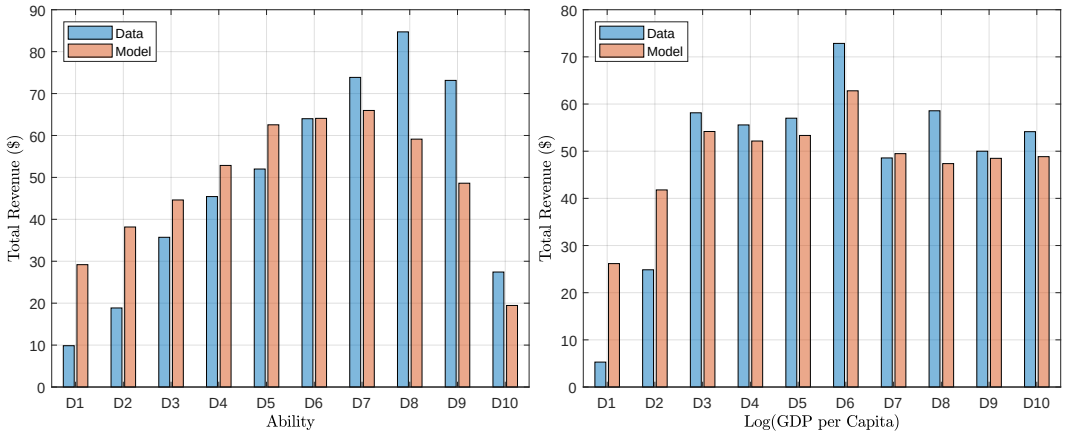
Figure 4.E.1: Observed Against Simulated Total Revenue by Region



Notes: This figure compares the total revenue by geographical region (in $) as observed in the data against that as predicted by the estimated model across different geographical regions. We calculate the observed total revenue as the sum over 43,660 players in Group 40 of the revenue collected at pay-gates 40, 60, and 80 in each geographical region during the 15 days of our sample in 2013. The expected total revenue as predicted by our estimated model is based on the simulation procedure detailed in Appendices 4.D.1 and 4.D.2. The geographical regions are described in Section 4.3.1 and Appendix 4.G.

Figure 4.E.2 compares the total revenue (in $) as observed in the data against that as predicted by the estimated model across players with different ability (left panel) and from countries with different GDP per capita (right panel). In particular, the left panel reports results by deciles (D1 being the lowest and D10 the highest decile) of ability while the right panel by deciles of log(GDP per capita). We calculate observed total revenue as the sum over 43,660 players in Group 40 of the revenue collected at pay-gates 40, 60, and 80 in each decile during the 15 days of our sample in 2013. The left panel shows that, when it comes to ability, the estimated model does a good job at predicting total revenue for 50% of players, those with ability between the third and the seventh decile. However, it tends to under-predict for most able players (the top three deciles) and to over-predict for the least able ones (the bottom two deciles). The right panel of Figure 4.E.2 confirms that, with the exception of the bottom two deciles of poorest countries, the estimated model is overall good at predicting total revenue in terms of players' log(GDP per capita). As we discuss at the end of Section 4.6 in the main text, in robustness checks reported in Appendix 4.F.1, we account for these predictive biases by limiting our counterfactual simulations to the sub-sample of players in Group 40 for whom the estimated model has better predictive power (deciles D3-D7 of ability and D3-D10 of GDP per capita).

Figure 4.E.2: Observed Against Predicted Revenue by Ability and GDP per Capita



Notes: These figures compare the total revenue (in $) as observed in the data against that as predicted by the estimated model across players with different ability (left panel) and from countries with different GDP per capita (right panel). In particular, the left panel reports results by deciles (D1 is the lowest decile and D10 the highest) of players' ability while the right panel by deciles of log(GDP per capita). Ability and log(GDP per capita) are described in Section 4.3.1. We calculate the observed total revenue as the sum over 43,660 players in Group 40 of the revenue collected at pay-gates 40, 60, and 80 in each decile during the 15 days of our sample in 2013. The expected total revenue as predicted by our estimated model is based on the simulation procedure detailed in Appendices 4.D.1 and 4.D.2.

4.F Additional Simulation Results

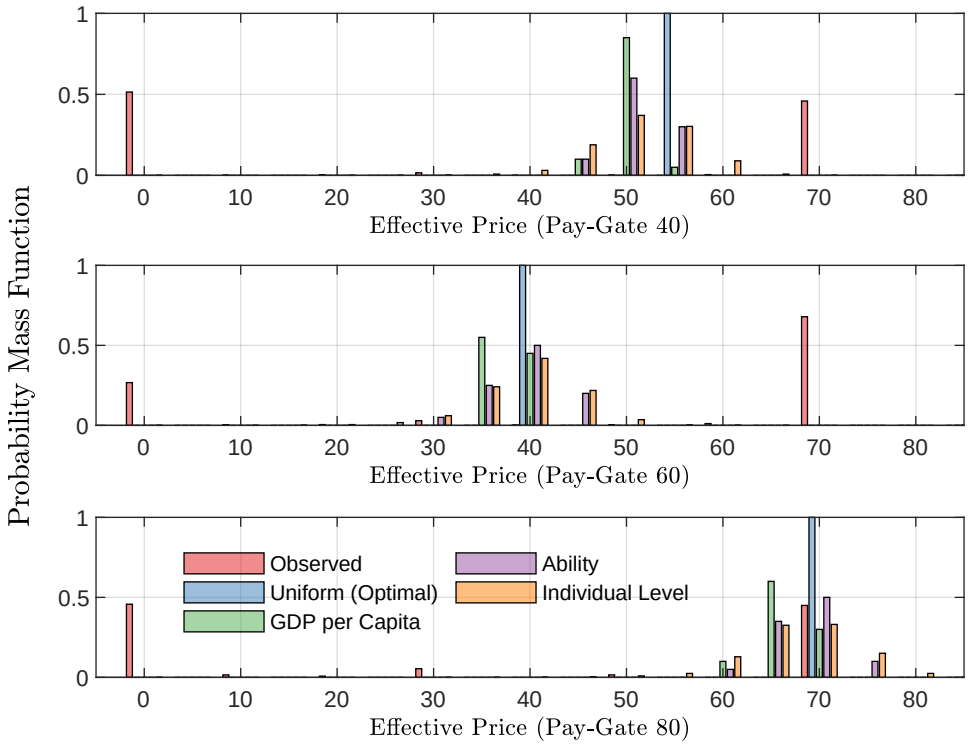
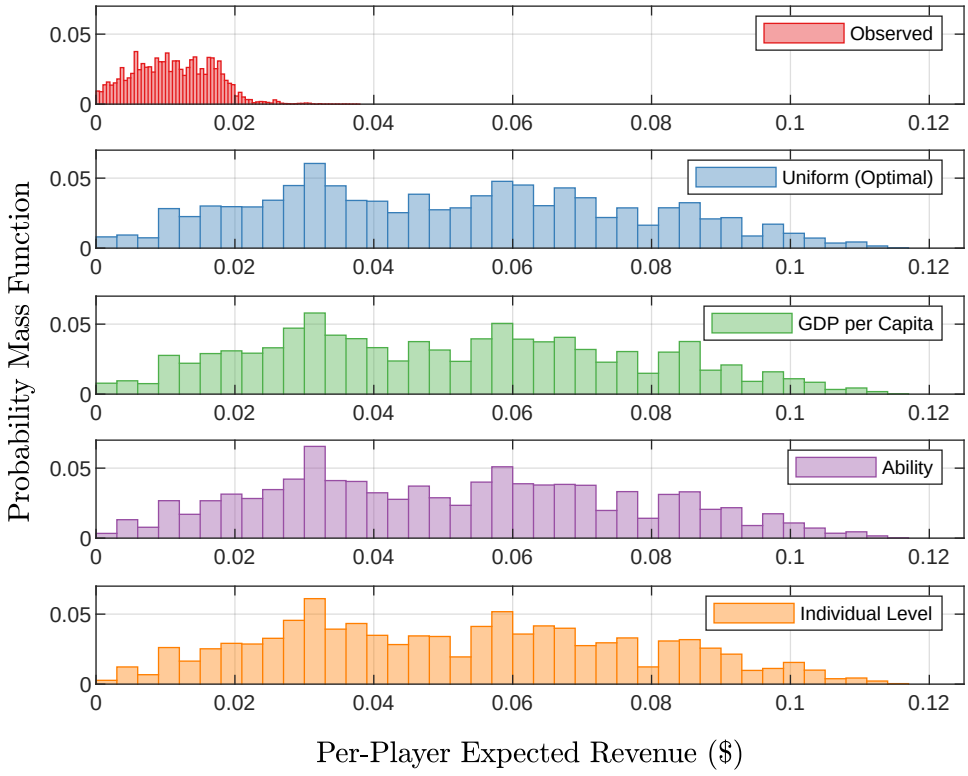


Figure 4.F.1: Distribution of Effective Prices in Dynamic Pricing Strategies

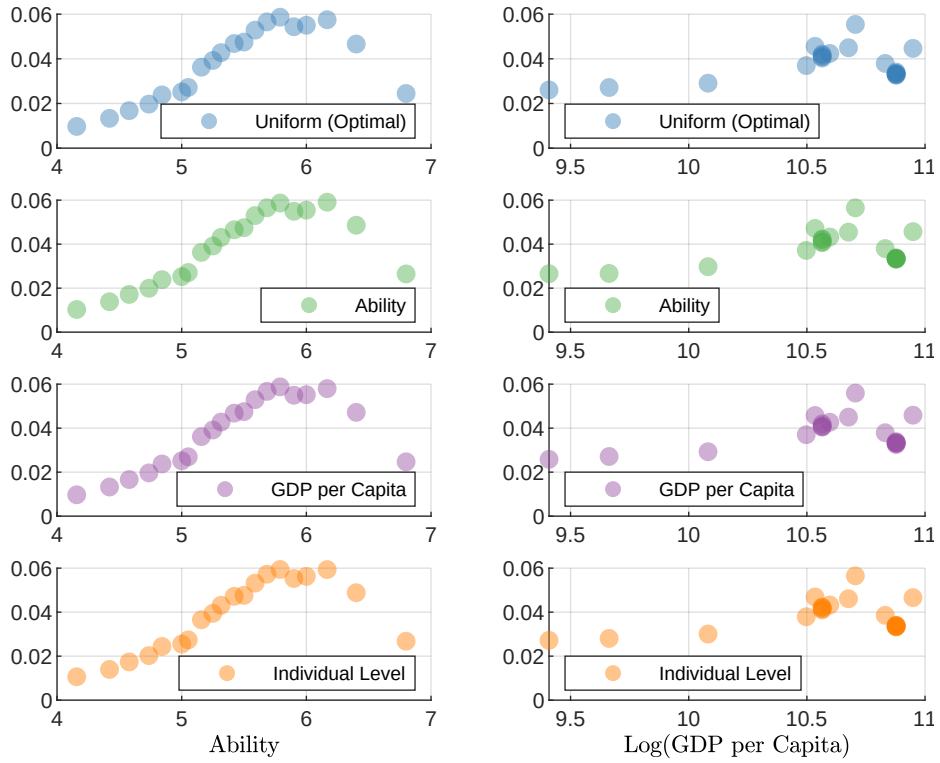
Notes: This figure shows the simulated distribution of effective prices (in virtual coins, where $\$1 \approx 70$ virtual coins) across players within each pay-gate for the "dynamic" pricing strategies considered in the right panel of Table 4.7. All pricing strategies are decreibed in detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices are allowed to change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.



Notes: This figure shows the simulated distribution of per-player expected revenue (in $) across players for the "dynamic" pricing strategies considered in the right panel of Table 4.7. All pricing strategies are described in detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices are allowed to change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

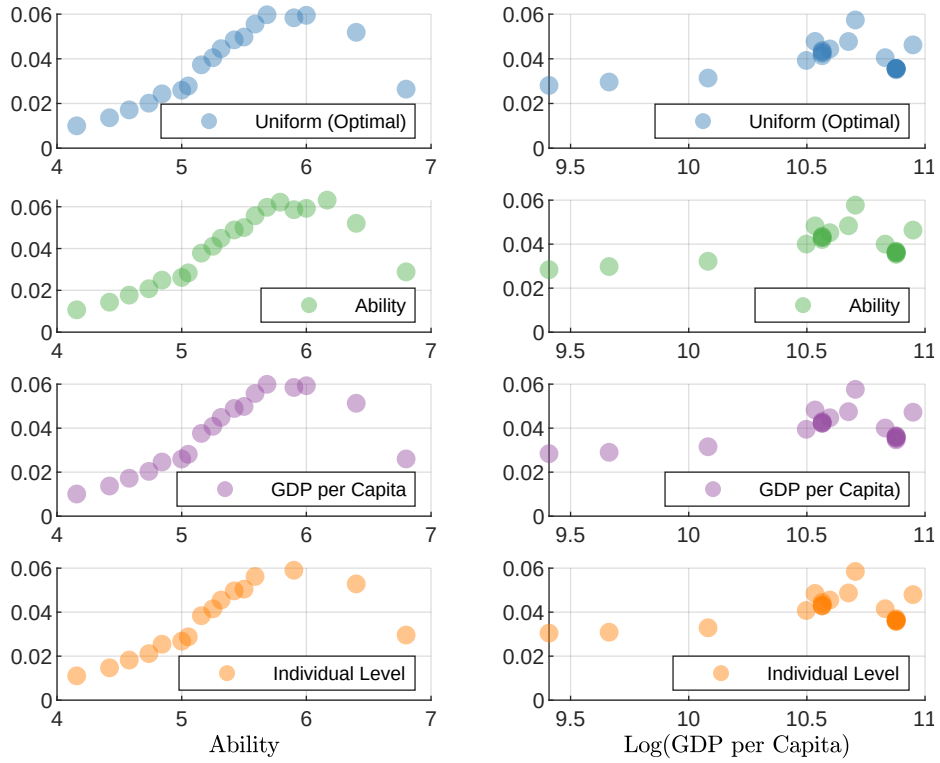
Figure 4.F.2: Distribution of Per-Player Expected Revenue, Dynamic Pricing Strategies

Figure 4.F.3: Δ Per-Player Expected Revenue, Static versus Observed Pricing



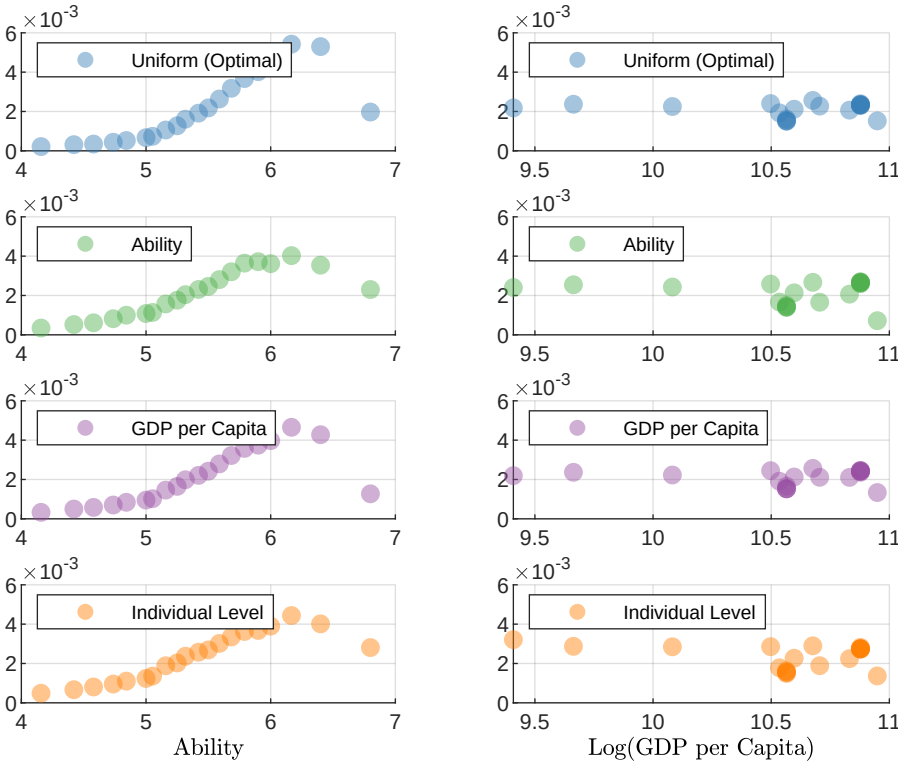
Notes: These binscatters display the average gains in per-player expected revenue (in $) of engaging in the "static" pricing strategies considered in the left panel of Table 4.7 as opposed to the observed pricing. For each static pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player expected revenue on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.F.4: Δ Per-Player Expected Revenue, Dynamic versus Observed Pricing

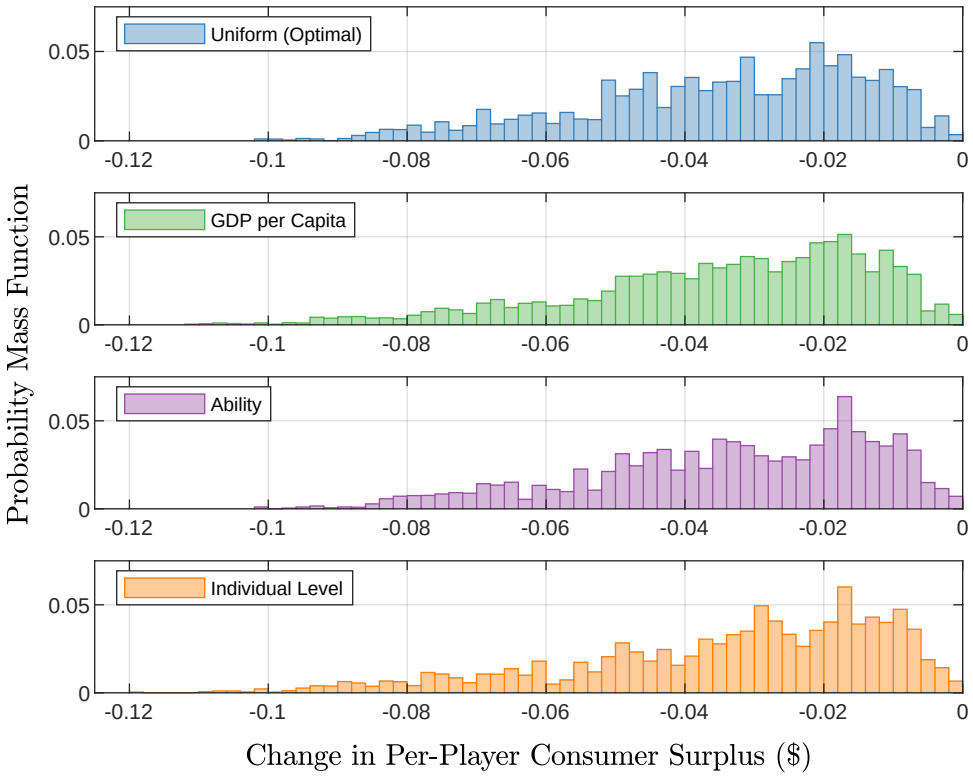


Notes: These binscatters display the average gains in per-player expected revenue (in $) of engaging in the "dynamic" pricing strategies considered in the right panel of Table 4.7 as opposed to the observed pricing. For each static pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player expected revenue on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are described in detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices are allowed to change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.F.5: Δ Per-Player Expected Revenue, Dynamic versus Static Pricing

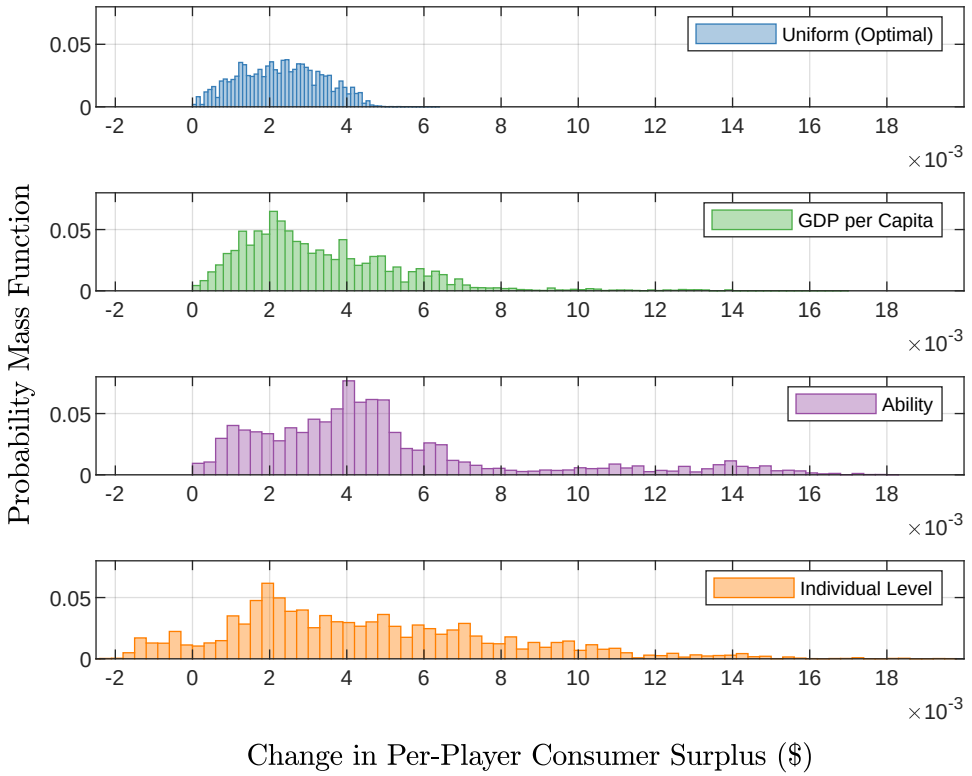


Notes: These binscatters display the average gains in per-player expected revenue (in $) of engaging in the "dynamic" versus the "static" versions of each of the pricing strategies considered in Table 4.7. For each pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player expected revenue on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. Dynamic pricing strategies are instead those in which effective prices are also allowed to change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.



Notes: This figure shows the simulated distribution of changes in per-player consumer surplus (in $) across players for the "dynamic" pricing strategies considered in the right panel of Table 4.8. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

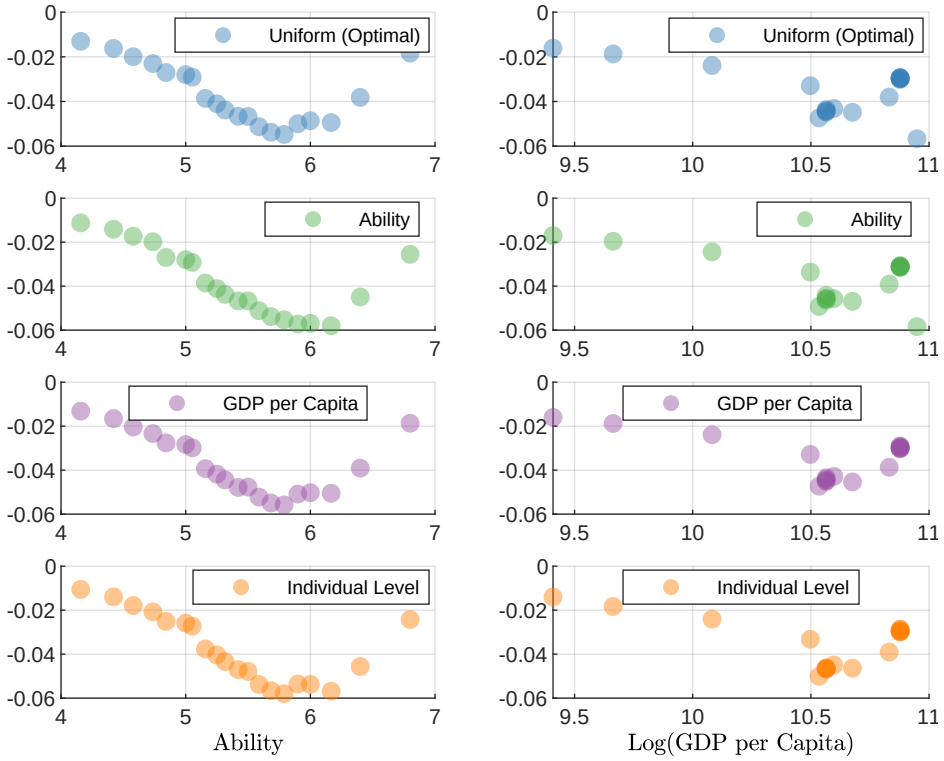
Figure 4.F.6: Distribution of Δ Per-Player Consumer Surplus, Dynamic versus Observed Pricing



Notes: This figure shows the simulated distribution of changes in per-player consumer surplus (in $) across players of engaging in the "dynamic" versus the "static" versions of each of the pricing strategies considered in Table 4.8. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices change among pay-gates. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

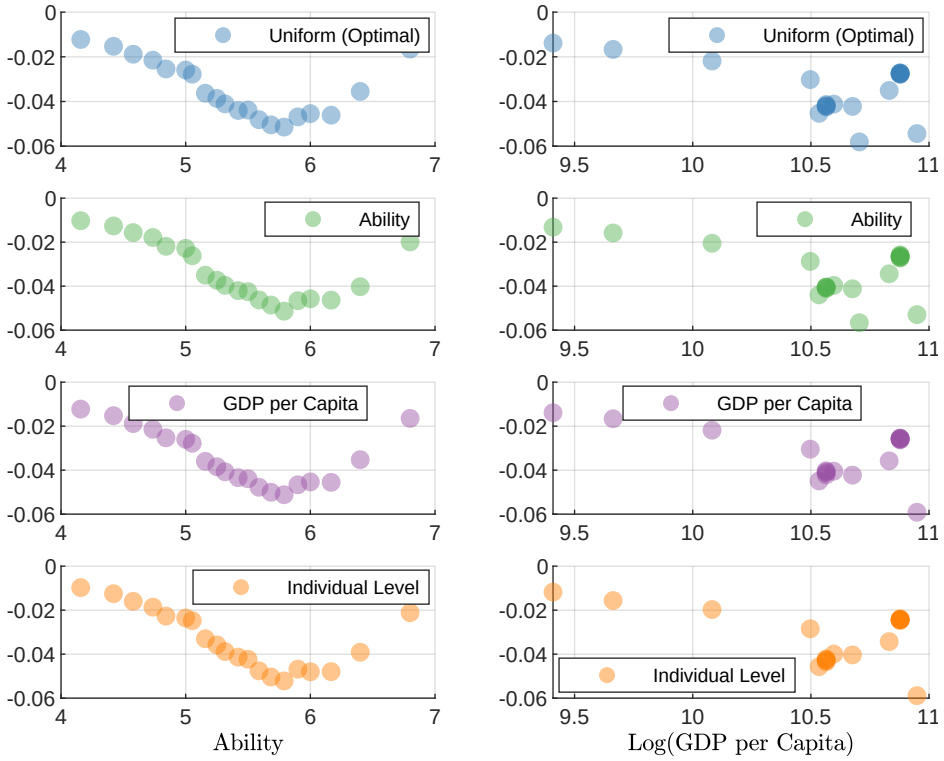
Figure 4.F.7: Distribution of Δ Per-Player Consumer Surplus, Dynamic versus Static Pricing

Figure 4.F.8: Δ Per-Player Consumer Surplus, Static versus Observed Pricing



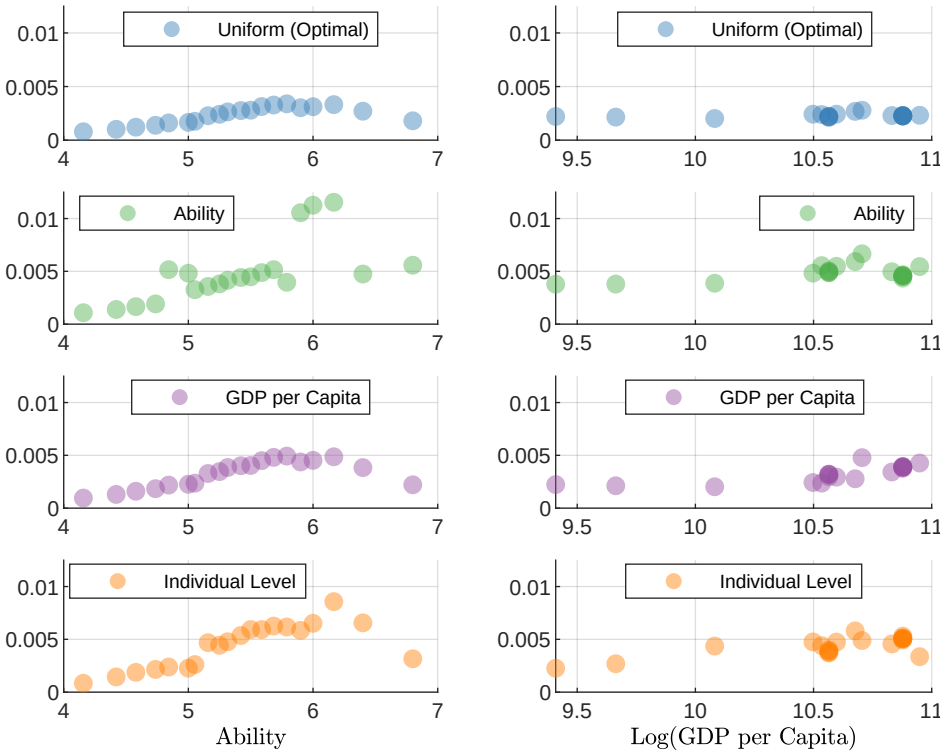
Notes: These binscatters display the average change in per-player consumer surplus (in $) of engaging in the "static" pricing strategies considered in the left panel of Table 4.8 as opposed to the observed pricing. For each static pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player expected consumer surplus on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.F.9: Δ Per-Player Consumer Surplus, Dynamic versus Observed Pricing

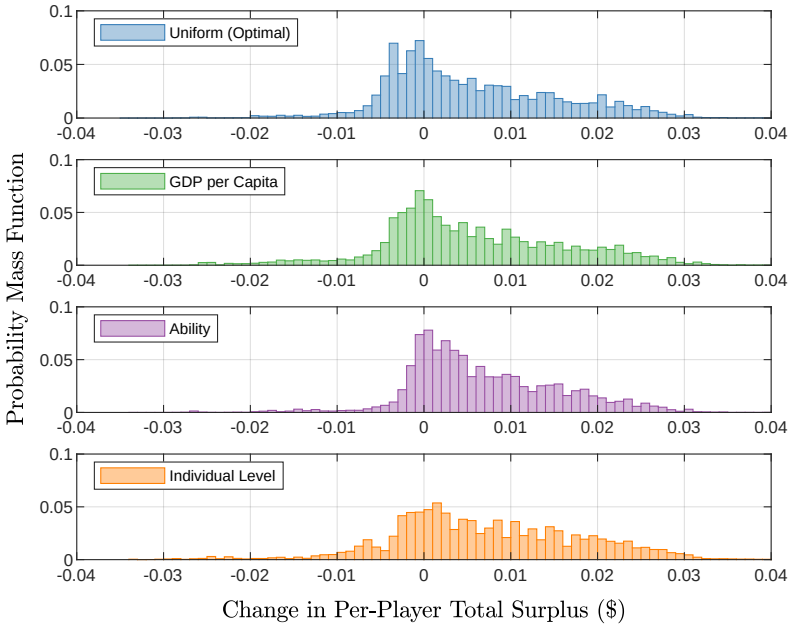


Notes: These binscatters display the average change in per-player consumer surplus (in $) of engaging in the "dynamic" pricing strategies considered in the right panel of Table 4.8 as opposed to the observed pricing. For each dynamic pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player consumer surplus on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.F.10: Δ Per-Player Consumer Surplus, Dynamic versus Static Pricing



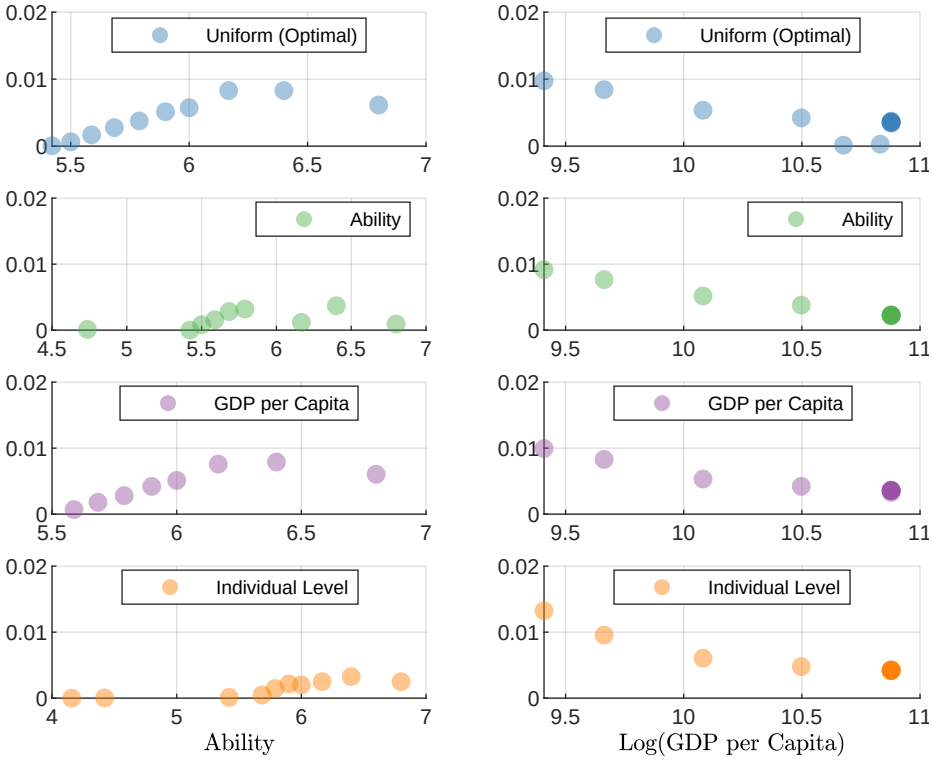
Notes: These binscatters display the average change in per-player consumer surplus (in $) of engaging in the "dynamic" versus the "static" versions of each of the pricing strategies considered in Table 4.8. For each pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player consumer surplus on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices change among pay-gates. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.



Notes: This figure shows the simulated distribution of changes in per-player total surplus (in $) across players for the "dynamic" pricing strategies considered in the right panel of Table 4.8 as opposed to the observed pricing. Changes in per-player total surplus are computed as the sum between changes in per-player expected revenues and in per-player consumer surplus. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

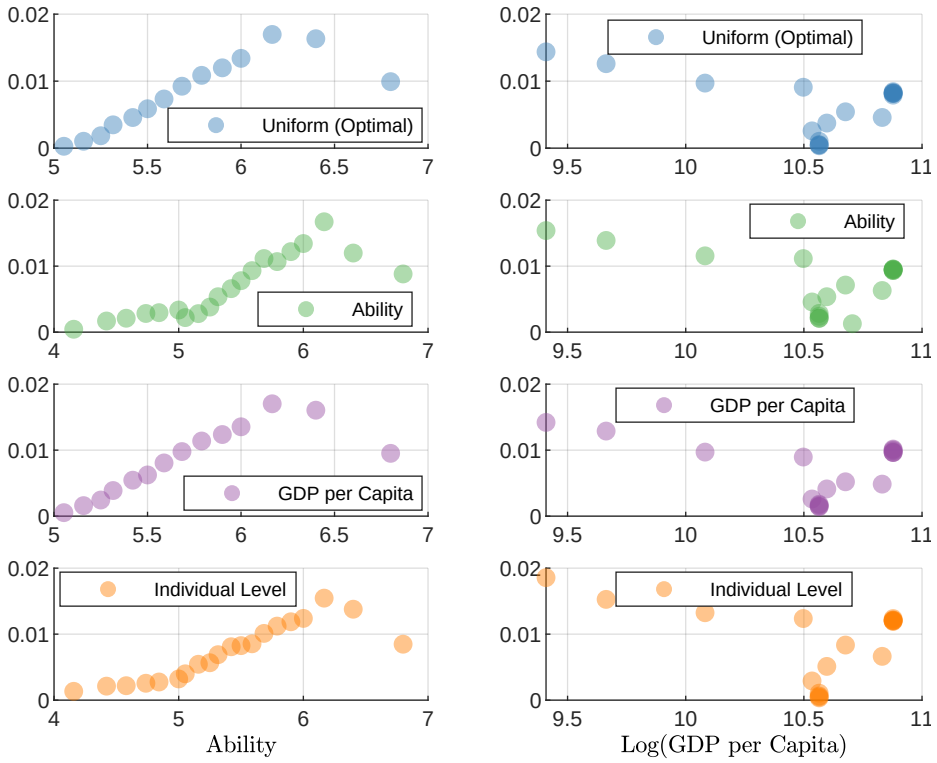
Figure 4.F.11: Distribution of Δ Per-Player Total Surplus, Dynamic versus Observed Pricing

Figure 4.F.12: Δ in Per-Player Total Surplus, Static versus Observed Pricing



Notes: These binscatters display the average gains in per-player total surplus (in $) of engaging in the "static" pricing strategies considered in the left panel of Table 4.8 as opposed to the observed pricing. For each static pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player expected revenue on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. Static pricing strategies are those in which effective prices do not change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Figure 4.F.13: Δ in Per-Player Total Surplus, Dynamic versus Observed Pricing



Notes: These binscatters display the average gains in per-player total surplus (in $) of engaging in the "dynamic" pricing strategies considered in the right panel of Table 4.8 as opposed to the observed pricing. For each dynamic pricing strategy, we construct 20 groups of players based on players' ability (left panel) and log(GDP per capita) (right panel) and plot the average group-specific difference in per-player expected revenue on the y-axis. The definitions of ability and log(GDP per capita) are provided in Section 4.3.1. All pricing strategies are described in detail in Appendix 4.D.2. Dynamic pricing strategies are those in which effective prices are allowed to change among pay-gates. All simulations are based on our estimates of models (4.5.1) and (4.5.7) and on the 43,660 players in Group 40 during the 15 days of our sample in 2013. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

4.F.1 Robustness Checks

In this Appendix, we repeat all counterfactual simulations limiting the sample of players in Group 40 to those for whom the estimated model displays the best predictive power in terms of expected revenue, namely the players with ability in deciles D3-D7 and GDP per capita in deciles D3-D10. We do this in order to account for the predictive biases of our estimated model as documented in Appendix 4.E. Overall, these robustness checks show no qualitative difference in any of our results and suggest that the predictive biases documented in Appendix 4.E do not play a crucial role in our simulations exercises.

Table 4.F.1: Robustness Check : Simulation of Effective Prices and Expected Revenue with Restricted Sample

Pricing Strategy	Static Pricing					Dynamic Pricing				
	Effective Price		Per-Player Revenue ($)			Effective Price		Per-Player Revenue ($)		
	mean	s.d.	mean	s.d.	%	mean	s.d.	mean	s.d.	%
Observed	35,566	34,529	0,014	0,120	-	-	-	-	-	-
Uniform (70)	70,000	-	0,028	0,008	98.1%	-	-	-	-	-
Uniform (Optimal)	45,000	-	0,056	0,020	292.4%	10,000	2,450	0,057	0,021	302.3%
GDP per Capita	44,385	1,642	0,056	0,020	292.2%	51,451	11,839	0,058	0,021	304.8%
Ability	45,001	0,075	0,056	0,020	292.4%	53,323	12,059	0,058	0,021	305.1%
Individual Level Pricing	45,603	3,086	0,056	0,020	294.1%	53,418	12,073	0,058	0,021	308.1%

Notes: This table summarizes our counterfactual simulation results in terms of effective prices and per-player expected revenues using a restricted sample. Each row refers to a pricing strategy and summarizes the simulated effective prices chosen by the firm (in virtual coins, where $\$1 \approx 70$ virtual coins) and the corresponding per-player expected revenues (in $). The columns denoted by "%" report the percentage increase in per-player expected revenue implied by the row pricing strategy with respect to the observed pricing chosen by the firm (i.e., 0% means same average as the observed pricing). All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. The left panel summarizes results for the case in which effective prices do not change among pay-gates (static pricing). The right panel instead summarizes results for the case in which effective prices are allowed to change also among pay-gates (dynamic pricing). All simulations are based on our estimates of models (4.5.1) and (4.5.7). The sample excludes players in Group 40 who are below the 2nd decile in terms of GDP per Capita, below the third decile in terms of ability, and above the seventh decile in ability. There are 17,719 remaining players. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

Table 4.F.2: Robustness Check : Simulation of Consumer Surplus and Total Surplus with Restricted Sample

Pricing Strategy	Static Pricing				Dynamic Pricing			
	Δ Consumer Surplus ($)		Δ Total Surplus ($)		Δ Consumer Surplus ($)		Δ Total Surplus ($)	
	mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
Uniform (70)	-0,0227	0,0128	-0,0088	0,0115	-	-	-	-
Uniform (Optimal)	-0,0441	0,0187	-0,0025	0,0059	-0,0417	0,0181	0,0014	0,0068
GDP per Capita	-0,0451	0,0202	-0,0035	0,0077	-0,0412	0,0192	0,0022	0,0085
Ability	-0,0441	0,0187	-0,0025	0,0059	-0,0396	0,0178	0,0038	0,0066
Individual Level	-0,0443	0,0215	-0,0025	0,0079	-0,0397	0,0207	0,0042	0,0094

Notes: This table summarizes our counterfactual simulation results in terms of per-player consumer surplus and per-player total surplus, computed as the sum between changes in per-player expected revenue and in per-player consumer surplus using a restricted sample. Each row refers to a pricing strategy and summarizes the simulated change in per-player consumer surplus and in per-player total surplus (both in $) with respect to the observed pricing. All pricing strategies are briefly described in the text and explained in more detail in Appendix 4.D.2. The left panel summarizes results for the case in which effective prices do not change among pay-gates (static pricing). The right panel instead summarizes results for the case in which effective prices are allowed to change also among pay-gates (dynamic pricing). All simulations are based on our estimates of models (4.5.1) and (4.5.7). The sample excludes players in Group 40 who are below the 2nd decile in terms of GDP per Capita, below the third decile in terms of ability, and above the seventh decile in ability. There are 17,719 remaining players. Details of the formulae and simulation procedures used can be found in Appendices 4.D.1 and 4.D.2.

4.G Data

Table 4.G.1: Countries and Region Assignment

Country	Region
Afghanistan	Southern Asia
Åland Islands	Northern Europe
Albania	Southern Europe
Algeria	Northern Africa
American Samoa	Other
Andorra	Southern Europe
Angola	Sub-Saharan Africa
Anguilla	Latin America and the Caribbean
Antarctica	Other
Antigua and Barbuda	Latin America and the Caribbean
Argentina	Latin America and the Caribbean
Armenia	Western Asia
Aruba	Latin America and the Caribbean
Australia	Australia and New Zealand
Austria	Western Europe
Azerbaijan	Western Asia
Bahamas	Latin America and the Caribbean
Bahrain	Western Asia
Bangladesh	Southern Asia
Barbados	Latin America and the Caribbean
Belarus	Eastern Europe
Belgium	Western Europe
Belize	Latin America and the Caribbean
Benin	Sub-Saharan Africa
Bermuda	Northern America
Bhutan	Southern Asia
Bolivia (Plurinational State of)	Latin America and the Caribbean
Bonaire, Sint Eustatius and Saba	Latin America and the Caribbean
Bosnia and Herzegovina	Southern Europe
Botswana	Sub-Saharan Africa
Bouvet Island	Latin America and the Caribbean
Brazil	Latin America and the Caribbean
British Indian Ocean Territory	Sub-Saharan Africa
Brunei Darussalam	South-eastern Asia
Bulgaria	Eastern Europe
Burkina Faso	Sub-Saharan Africa
Burundi	Sub-Saharan Africa
Cabo Verde	Sub-Saharan Africa
Cambodia	South-eastern Asia
Cameroon	Sub-Saharan Africa
Canada	Northern America
Cayman Islands	Latin America and the Caribbean
Central African Republic	Sub-Saharan Africa
Chad	Sub-Saharan Africa

Chile	Latin America and the Caribbean
China	Eastern Asia
Christmas Island	Australia and New Zealand
Cocos (Keeling) Islands	Australia and New Zealand
Colombia	Latin America and the Caribbean
Comoros	Sub-Saharan Africa
Congo	Sub-Saharan Africa
Congo, Democratic Republic of the	Sub-Saharan Africa
Cook Islands	Other
Costa Rica	Latin America and the Caribbean
Cote d'Ivoire	Sub-Saharan Africa
Croatia	Southern Europe
Cuba	Latin America and the Caribbean
Curacao	Latin America and the Caribbean
Cyprus	Western Asia
Czechia	Eastern Europe
Denmark	Northern Europe
Djibouti	Sub-Saharan Africa
Dominica	Latin America and the Caribbean
Dominican Republic	Latin America and the Caribbean
Ecuador	Latin America and the Caribbean
Egypt	Northern Africa
El Salvador	Latin America and the Caribbean
Equatorial Guinea	Sub-Saharan Africa
Eritrea	Sub-Saharan Africa
Estonia	Northern Europe
Eswatini	Sub-Saharan Africa
Ethiopia	Sub-Saharan Africa
Falkland Islands (Malvinas)	Latin America and the Caribbean
Faroe Islands	Northern Europe
Fiji	Other
Finland	Northern Europe
France	Western Europe
French Guiana	Latin America and the Caribbean
French Other	Other
French Southern Territories	Sub-Saharan Africa
Gabon	Sub-Saharan Africa
Gambia	Sub-Saharan Africa
Georgia	Western Asia
Germany	Western Europe
Ghana	Sub-Saharan Africa
Gibraltar	Southern Europe
Greece	Southern Europe
Greenland	Northern America
Grenada	Latin America and the Caribbean
Guadeloupe	Latin America and the Caribbean
Guam	Other
Guatemala	Latin America and the Caribbean
Guernsey	Northern Europe
Guinea	Sub-Saharan Africa

Guinea-Bissau	Sub-Saharan Africa
Guyana	Latin America and the Caribbean
Haiti	Latin America and the Caribbean
Heard Island and McDonald Islands	Australia and New Zealand
Holy See	Southern Europe
Honduras	Latin America and the Caribbean
Hong Kong	Eastern Asia
Hungary	Eastern Europe
Iceland	Northern Europe
India	Southern Asia
Indonesia	South-eastern Asia
Iran (Islamic Republic of)	Southern Asia
Iraq	Western Asia
Ireland	Northern Europe
Isle of Man	Northern Europe
Israel	Western Asia
Italy	Southern Europe
Jamaica	Latin America and the Caribbean
Japan	Eastern Asia
Jersey	Northern Europe
Jordan	Western Asia
Kazakhstan	Other
Kenya	Sub-Saharan Africa
Kiribati	Other
Korea (Democratic People's Republic of)	Eastern Asia
Korea, Republic of	Eastern Asia
Kuwait	Western Asia
Kyrgyzstan	Other
Lao People's Democratic Republic	South-eastern Asia
Latvia	Northern Europe
Lebanon	Western Asia
Lesotho	Sub-Saharan Africa
Liberia	Sub-Saharan Africa
Libya	Northern Africa
Liechtenstein	Western Europe
Lithuania	Northern Europe
Luxembourg	Western Europe
Macao	Eastern Asia
Madagascar	Sub-Saharan Africa
Malawi	Sub-Saharan Africa
Malaysia	South-eastern Asia
Maldives	Southern Asia
Mali	Sub-Saharan Africa
Malta	Southern Europe
Marshall Islands	Other
Martinique	Latin America and the Caribbean
Mauritania	Sub-Saharan Africa
Mauritius	Sub-Saharan Africa
Mayotte	Sub-Saharan Africa
Mexico	Latin America and the Caribbean

Other (Federated States of)	Other
Moldova, Republic of	Eastern Europe
Monaco	Western Europe
Mongolia	Eastern Asia
Montenegro	Southern Europe
Montserrat	Latin America and the Caribbean
Morocco	Northern Africa
Mozambique	Sub-Saharan Africa
Myanmar	South-eastern Asia
Namibia	Sub-Saharan Africa
Nauru	Other
Nepal	Southern Asia
Netherlands	Western Europe
New Caledonia	Other
New Zealand	Australia and New Zealand
Nicaragua	Latin America and the Caribbean
Niger	Sub-Saharan Africa
Nigeria	Sub-Saharan Africa
Niue	Other
Norfolk Island	Australia and New Zealand
North Macedonia	Southern Europe
Northern Mariana Islands	Other
Norway	Northern Europe
Oman	Western Asia
Pakistan	Southern Asia
Palau	Other
Palestine, State of	Western Asia
Panama	Latin America and the Caribbean
Papua New Guinea	Other
Paraguay	Latin America and the Caribbean
Peru	Latin America and the Caribbean
Philippines	South-eastern Asia
Pitcairn	Other
Poland	Eastern Europe
Portugal	Southern Europe
Puerto Rico	Latin America and the Caribbean
Qatar	Western Asia
Reunion	Sub-Saharan Africa
Romania	Eastern Europe
Russian Federation	Eastern Europe
Rwanda	Sub-Saharan Africa
Saint Barthelemy	Latin America and the Caribbean
Saint Helena, Ascension and Tristan da Cunha	Sub-Saharan Africa
Saint Kitts and Nevis	Latin America and the Caribbean
Saint Lucia	Latin America and the Caribbean
Saint Martin (French part)	Latin America and the Caribbean
Saint Pierre and Miquelon	Northern America
Saint Vincent and the Grenadines	Latin America and the Caribbean
Samoa	Other
San Marino	Southern Europe

Sao Tome and Principe	Sub-Saharan Africa
Saudi Arabia	Western Asia
Senegal	Sub-Saharan Africa
Serbia	Southern Europe
Seychelles	Sub-Saharan Africa
Sierra Leone	Sub-Saharan Africa
Singapore	South-eastern Asia
Sint Maarten (Dutch part)	Latin America and the Caribbean
Slovakia	Eastern Europe
Slovenia	Southern Europe
Solomon Islands	Other
Somalia	Sub-Saharan Africa
South Africa	Sub-Saharan Africa
South Georgia and the South Sandwich Islands	Latin America and the Caribbean
South Sudan	Sub-Saharan Africa
Spain	Southern Europe
Sri Lanka	Southern Asia
Sudan	Northern Africa
Suriname	Latin America and the Caribbean
Svalbard and Jan Mayen	Northern Europe
Sweden	Northern Europe
Switzerland	Western Europe
Syrian Arab Republic	Western Asia
Taiwan, Province of China	Eastern Asia
Tajikistan	Other
Tanzania, United Republic of	Sub-Saharan Africa
Thailand	South-eastern Asia
Timor-Leste	South-eastern Asia
Togo	Sub-Saharan Africa
Tokelau	Other
Tonga	Other
Trinidad and Tobago	Latin America and the Caribbean
Tunisia	Northern Africa
Turkey	Western Asia
Turkmenistan	Other
Turks and Caicos Islands	Latin America and the Caribbean
Tuvalu	Other
Uganda	Sub-Saharan Africa
Ukraine	Eastern Europe
United Arab Emirates	Western Asia
United Kingdom of Great Britain and Northern Ireland	Northern Europe
United States of America	Northern America
United States Minor Outlying Islands	Other
Uruguay	Latin America and the Caribbean
Uzbekistan	Other
Vanuatu	Other
Venezuela (Bolivarian Republic of)	Latin America and the Caribbean
Viet Nam	South-eastern Asia
Virgin Islands (British)	Latin America and the Caribbean
Virgin Islands (U.S.)	Latin America and the Caribbean

Wallis and Futuna	Other
Western Sahara	Northern Africa
Yemen	Western Asia
Zambia	Sub-Saharan Africa
Zimbabwe	Sub-Saharan Africa

Figure 4.G.1: Share of Players per Region

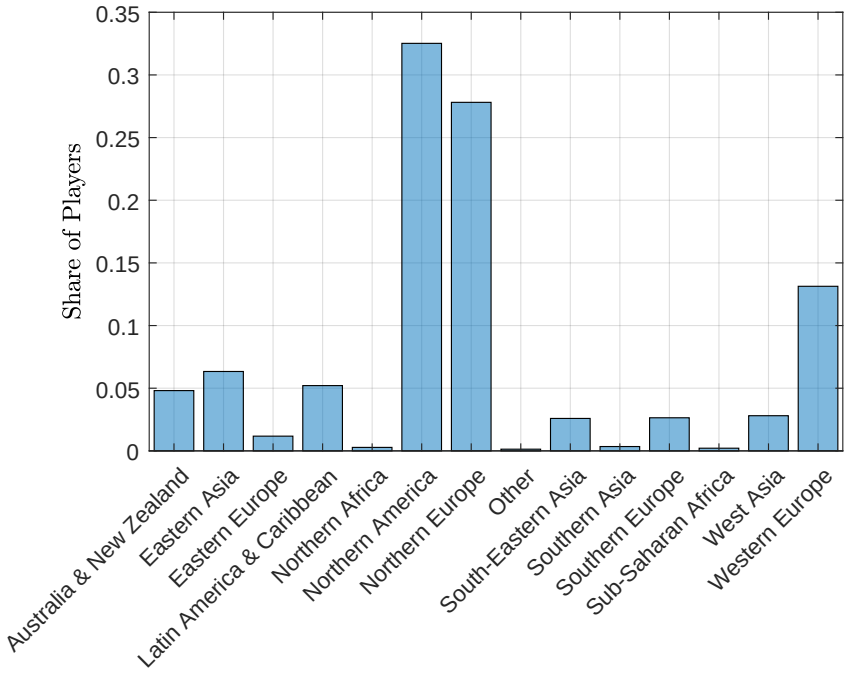


Table 4.G.2: Balance test: Comparison across groups

	1st gate at level 20		1st gate at level 40				Differences		
	Stars		Stars		No stars				
	(1)		(2)		(3)		(1)-(2)	(1)-(3)	(2)-(3)
	Mean	Std. dev.	Mean	Std. dev.	Mean	Std. dev.			
Avg. Snake Length	5,226	0,817	5,221	0,810	5,223	0,811	0,005	0,003	-0,002
Avg. Move Count	10,399	2,331	10,398	2,315	10,388	2,319	0,002	0,012	0,010
Avg. Final Score	24973,640	6554,295	24963,570	6544,456	24940,420	6570,346	10,070	33,220	23,150
Avg. Rounds per Level	3,932	5,665	3,989	5,787	3,934	5,703	-0,058	-0,002	0,055*
Number of Players	43,218		43,660		205,415				

*$p < 0.10$, **$p < 0.05$, ***$p < 0.01$.

Notes: This table provides evidence of balance between the different experimental groups. The sample includes all players in each group, as explained in section 4.3.2. It provides averages and associated t-tests (with unequal variances) across these groups. For the purposes of comparison, averages are calculated across the first 20 levels of the game, dropping rounds of the game spent on levels already previously cleared. The average snake length is the per-player average number of consecutive jellies assembled in a round of the game. The average move count counts instead the number different moves played in the game. The average final score reflects the aggregate performance of a player in a given round of the game. The average rounds per level is the number of attempts before first success.

Chapter 5

Dealing with Logs and Zeros in Regression Models

This chapter is joint work with Christophe Bellégo and David Benatia.

Abstract: Log-linear models are prevalent in empirical research. Yet, how to handle zeros in the dependent variable remains an unsettled issue. This article clarifies it and addresses the "log of zero" by developing a new family of estimators called iterated Ordinary Least Squares (iOLS). This family nests standard approaches such as log-linear and Poisson regressions, offers several computational advantages, and somehow reconciles the log(1 + Y) with econometric theory. We extend it to the endogenous regressor setting (i2SLS) and overcome other common issues with Poisson models, such as controlling for many fixedeffects. We also develop specification tests to help researchers select between alternative estimators. Finally, our methods are illustrated through numerical simulations and replications of recent publications.

5.1 Introduction

The log-linear and log-log models are among the most frequent specifications used in empirical research.[1] However, having to deal with the (natural) logarithm of a zero in the response variable is a common issue faced by practitioners. There is, unfortunately, a lack of consensus about the best practice to address those zeros, as evidenced by the many alternative solutions used in recent leading publications. This paper not only clarifies this issue, it also develops a new family of estimators and a model selection procedure. Our estimators are simple iterative extensions of ordinary least squares (OLS) and two-stage least-squares (2SLS). They are consistent, asymptotically normal, computationally simple, and can accommodate many fixed-effects. We also develop specification tests aimed at verifying the external validity of the model with respect to the observed patterns of zeros in the data. Those tests prove to be helpful for selecting the most suitable approach to address the log of zero in any given setting.

The log transformation is popular because (1) the parameter estimate is related to an elasticity;[2] (2) logs can linearize a theoretical model, e.g. a Cobb-Douglas production function (Goldberger, 1968) or a gravity equation

[1] In this paper, we focus on the log-linear model and address the minor differences of the log-log model as an extension.

[2] In a log-log model such as $\log(y) = \beta \log(x) + \epsilon$, the elasticity of y with respect to x is given by $\frac{\partial \log(y)}{\partial \log(x)} = \frac{\partial y}{\partial x}\frac{x}{y} = \beta$.

(Head and Mayer, 2019); (3) logs can make heteroskedasticity vanish in some settings, e.g. when the variance of a variable is proportional to its squared mean (Carroll and Ruppert, 1984); (4) the data is sometimes *naturally* related by a log-linear relationship (Ciani and Fisher, 2018); or even (5) it provides a concave transformation (MacKinnon and Magee, 1990).

However, the variable taken in logs may contain non-positive values. For example, a company can employ no worker, a product can have no sales or two countries zero trade in a given year. In these cases, the log is undefined and a fix is needed. Although this problem is quite common, the solution to be adopted is still unclear to many empirical researchers. We have reviewed all articles published in the American Economic Review (AER) between 2016 and 2020 to support this statement. Figure 5.1 summarizes our findings. It shows that nearly 40% of empirical papers used a log-specification and 36% of these faced the problem of the log of zero. It corresponds to an average of 10 publications per year dealing with the log of zero in the AER.

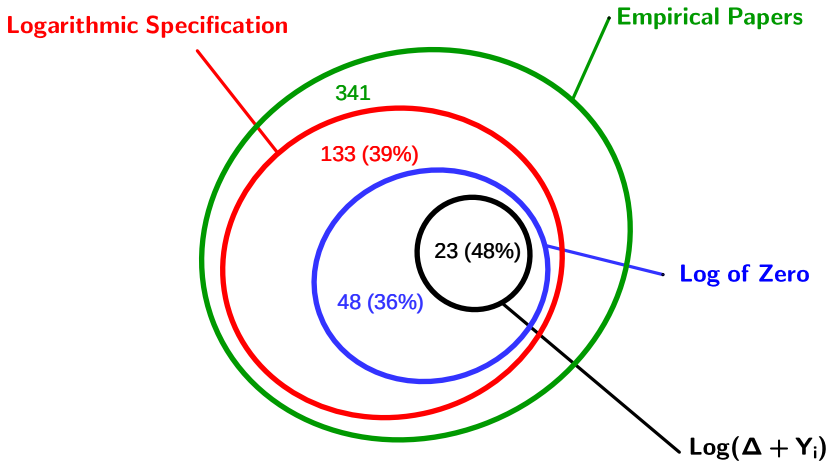


Figure 5.1: Prevalence of the Log of Zero in the AER (2016-2020)

No *single* solution has achieved consensus. In most publications, the authors chose to keep the zero observations and opted to either (1) add a positive discretionary value to the dependent variable (48%), (2) use Poisson-type estimators (35%), or (3) apply the inverse hyperbolic sine (IHS) transformation (15%). Discarding non-positive observations occurred in 31% of publications. We also note that in around 20% of cases, the authors compared several methods in order to gauge the robustness of their results.[3]

Moreover, researchers seldom report all their intermediary results leading to the submission of an article. To uncover existing practices, we have conducted a survey in three online seminars in economic departments asking "What would you do when facing the log of zero?".[4] Among the 28 respondents (including 21% of Ph.D students), 42% opt for the popular fix, 35% for mixture models (Tobit, Heckit, etc.), and 18% for abandoning the use of a log-like specification. Putting the latter individuals aside, only 46% would compare multiple approaches, on average 2.7 each. It is interesting to note that the (somewhat large) stated preference for mixture models is not reflected in recent AER publications.

The issue of the log of zero extends well beyond economics. The question "Log transformation of values

[3]This excludes cases where the authors decided to use a linear specification by fault of having to use such a fix. See Table 5.4.2 in the Appendix for additional details and information regarding data collection.

[4]See Appendix 5.4.3 for the survey and for the exhaustive set of results.

that include 0 (zero) for statistical analyses?" asked in 2014 on the forum ResearchGate, a multidisciplinary research-oriented social network, has received 38 contributions from researchers in various fields, including, but not limited to, medicine, biology, statistics and engineering. The thread has been read 120,000 times as of August 2020.[5] The prevalence of each solution is comparable to that in the AER. Adding a positive constant is suggested 50% of the time. Poisson, mixture models, and transformations like IHS are each recommended only 12.5% of the time.

There are hence five main solutions to the "log of zero". The most common fix consists in adding a positive constant to all observations (MaCurdy and Pencavel, 1986). This approach will thereafter be referred to as the "popular fix". A second solution is to delete the non-positive observations from the sample (Young and Young, 1975). A third solution uses transformations of the response variable, such as IHS, akin to the log function (MacKinnon and Magee, 1990; Burbidge et al., 1988; Johnson, 1949). A fourth solution consists in adopting mixture models (e.g. Tobit or Heckit) where a sample selection process explains the occurrence of non-positive observations (Heckman, 1979b; Eaton and Tamura, 1994; Helpman et al., 2008). Finally, Poisson models (Gourieroux et al., 1984) handle the presence of zeros well in many settings. They are especially popular in international trade where it is the workhorse model for the estimation of gravity equations (Head and Mayer, 2019). To the best of our knowledge, Santos Silva and Tenreyro (2006) were the first to argue in favor of Poisson regression to address the log of zero.

However, to deal with the log of zero and select among these models one must first address the critical question "why do the data contain zeros?".[6] It could be due to either data problems, such as measurement errors of small values, or a "true zero", for example when a product has exactly zero sale. In any case, one must make distributional assumptions about the zeros either explicitly (e.g. Tobit or Heckit) or implicitly through moment restrictions (e.g. PPML or IHS). We will discuss the assumptions made by existing methods, and propose a model of the latter type. The main advantage of this approach is that it does not require to specify a selection process explaining the occurrence of zeros.

The main focus of our paper is the identification of the model parameters rather than the prediction of an outcome. Identification is key for the estimated parameters to have an economic interpretation. It typically relies on exogeneity restrictions in the form of moment conditions between the errors and regressors, like OLS or Poisson regression. Our discussions with empirical researchers revealed that many opt for the popular fix approach because they do not feel comfortable assuming the exogeneity restriction imposed by Poisson models.[7] Instead, they seem to believe that adding a constant to the outcome before taking the log function yields an error satisfying an exogeneity condition close to that of OLS in a log-linear model. Unfortunately, it does not.

Our approach consists in adding an observation-specific value to the outcome instead of a constant. It makes use of an exogeneity condition, either user-chosen or data-driven, in a range of possible conditions between that of the log-linear model and Poisson. Our estimators are then computed thanks to an iterative procedure. We rely for that on the asymptotic theory developed in Dominitz and Sherman (2005) to prove the

[5] See https://www.researchgate.net/post/Log_transformation_of_values_that_include_0_zero_for_statistical_analyses2, and Figure 5.4.1 in the Appendix.

[6] This question echoes that of Heckman (1979b) about missing data: "why are the data missing?".

[7] In February 2021, Jeffrey Wooldridge tweeted *"Poisson regression can get one so far with so little trouble, why do so many still resist? [...]"* (https://twitter.com/jmwooldridge/status/1363828456136523779?s=20.) Ten years earlier, the President of StataCorp, William Gould, wrote a blog post arguing that researchers should use Poisson regression rather than OLS with a log outcome: https://blog.stata.com/2011/08/22/use-poisson-rather-than-regress-tell-a-friend/.

consistency and asymptotic normality of our estimators. Iterative estimation methods are frequently encountered in physics and machine learning, where Iterated Reweighted Least-Squares (IRLS) are widely used for robust estimation (Dembinski et al., 2019). Although less popular in economics, iterative estimators are used in some settings. For instance, Blundell and Robin (1999) propose an iterative solution for demand estimation to improve computational efficiency with respect to non-linear methods. Another example is the iterative estimation strategy in Head and Mayer (2014) of the structural gravity model of Anderson and van Wincoop (2003).

We make three principal contributions. First, we clarify the log of zero issue in a didactic way by reviewing existing practices. Second, we develop a new family of solutions, referred to as iterated OLS (iOLS). They consist in adding a data-dependent value to each observation and iterating OLS on the transformed model until convergence. They have multiple advantages: (a) they can be estimated by ordinary least squares, hence are computationally fast and easy to implement[8]; (b) robust standard errors are readily available; (c) they do not suffer from highly dispersed response variables; (d) they extend naturally to the endogenous setting using iterated 2SLS (i2SLS); and (e) they are amenable to different identifying assumptions. Finally, we develop a procedure to select which solution should be preferred in any given setting. This procedure helps choosing the most plausible model(s) given the data at hand. It consists in testing the implicit assumption about the patterns of zeros made by each approach. More formally, it is a test of whether the conditional probability of having a zero implied by the model is consistent with the data.

Our methodological contributions are illustrated through numerical simulations and (partial) replications of three recent publications in top-tier economics journals. First, Santos Silva and Tenreyro (2006) compare various estimators to estimate gravity models of trade and argue in favor of Poisson regression. Second, Michalopoulos and Papaioannou (2013) adds a positive constant to the response variable in order to examine the role of pre-colonial ethnic institutions on economic development. Third, Card and DellaVigna (2020) investigate the preferences of academic journal editors with the IHS transformation. Our tests reveal that no single solution is preferred in all settings. Nevertheless, iOLS tends to be selected more often than other methods in those examples.

The remaining of the paper is organized as follows. Section 5.2 clarifies the log of zero issue and discusses existing practices found in empirical research. Section 5.3 develops a new family of solutions. Section 5.4 presents specification tests and a data-driven model selection procedure. Numerical simulations are presented in Section 5.5. Partial replications of leading publications are proposed in Section 5.6. Section 5.7 concludes the paper.

The Appendix section also contains several useful extensions to our methods. First, we adapt it to the endogenous setting in Appendix 5.2.1. Second, we address the case where discarding zeros does not jeopardize identification in Appendix 5.2.2. Third, we show in Appendix 5.2.3 how to deal with negative values in Y. Appendix 5.2.4 discusses log-log specifications with zeros in the independent variables. Appendix 5.2.5 develops a computationally fast "within" iOLS estimator to avoid the incidental parameter problem when many fixed-effects are included. Appendix 5.2.6 shows how to deal with the log of a ratio of two response variables. Appendix 5.2.7 makes use of yet another alternative exogeneity condition close to that of the log-linear model. Finally, Appendix 5.2.8 details the testing procedures in the endogenous regressors setting.

[8] Stata packages for iOLS$_\delta$ and i2SLS$_\delta$ (with potentially high dimensional fixed effects) are available from www.https://github.com/ldpape.

5.2 Existing Practices

Let us consider an iid sample of observations $\{Y_i, X_i\}_{i=1}^n$, where n denotes the sample size, generated by the "true" model given by

$$Y_i = \exp(X_i'\beta + \varepsilon_i)\xi_i, \tag{5.2.1}$$

where β is a fixed parameter of interest in $\mathbb{R}^K$, with $K \geq 1$, ε_i is an iid mean-zero error term, and $\xi_i \in \{0, 1\}$ is a Bernoulli random error and, without loss of generality, we take $E[\exp(\varepsilon_i)\xi_i] = 1$. Let X denote the $n \times K$ matrix comprised of the K-dimensional column vector X_i with elements X_{ki}, for $1 \leq k \leq K$. Let us assume that $E(X_i X_i') < \infty$, and X has full column rank.

Y_i can either be equal to zero, when $\xi_i = 0$, or take positive values, when $\xi_i = 1$. Taking logs on both sides of (5.2.1) is allowed only if Y_i (and thus ξ_i) takes only strictly positive values. Doing so yields the log-linear model given by

$$\log(Y_i) = X_i'\beta + \varepsilon_i. \tag{5.2.2}$$

For parsimony, we will rely on the more compact *multiplicative* representation,

$$Y_i = \exp(X_i'\beta)U_i, \tag{5.2.3}$$

where $U_i = \exp(\varepsilon_i)\xi_i$ has mean one, and refer to the equivalent *additive* model

$$Y_i = \exp(X_i'\beta) + \epsilon_i, \tag{5.2.4}$$

with $\epsilon_i = \exp(X_i'\beta)(U_i - 1)$ treated as a mean-zero error.

5.2.1 The popular fix: to add a positive constant

The most popular solution is to add a positive constant Δ to all observations Y_i so that $\tilde{Y}_i = Y_i + \Delta > 0$ and the log-transformation becomes feasible. The choice of Δ is discretionary and may arbitrarily bias the estimates and their standard errors. Moreover, the size of the bias will depend on the data at hand, suggesting that adding the smallest possible constant is not necessarily the least "harmful" choice.[9]

To understand the bias, consider the model specified in (5.2.1). Adding $\Delta > 0$ and applying the log function yields after rearrangement

$$\log(Y_i + \Delta) = X_i'\beta + \log\left(U_i + \frac{\Delta}{\exp(X_i'\beta)}\right) \tag{5.2.5}$$

where the error term $\omega_i = \log\left(U_i + \frac{\Delta}{\exp(X_i'\beta)}\right)$ is correlated with X_i by construction, even when U_i and X_i are statistically independent, and creates an endogeneity bias. Although the choice of Δ matters, $\exp(X_i'\beta)$ can be arbitrarily close to zero hence leading to possibly large biases. Thus, the "popular fix" estimator is (in general) not consistent.[10]

Anecdotal evidence reveals that empiricists sometimes believe this bias to be negligible for small values of Δ, or for $\Delta = 1$. This belief holds true only under strong and unverifiable restrictions about the underlying

[9] Other variants include adding a constant solely to the non-positive values and including an additional dummy variable indicating such a treatment, generating the same kind of troubles. Alternatively, Johnson and Rausser (1971) propose to estimate the constant along with the other parameters. However, their method does not guarantee unbiased estimates.

[10] This estimator is consistent under the condition $E(\omega_i|X) = constant$ which implies strong assumptions of the joint distribution of U_i and X_i.

DGP. To illustrate this point, we rely on numerical simulations based on the design detailed in Section 5.5 (DGP 1). The objective is to estimate the parameters $\beta_1 = \beta_2 = 1$. Figure 5.2 presents the mean estimates using the popular fix, i.e. the OLS estimate of (5.2.5), as a function of the value of Δ. For this parameter, the mean squared bias is minimized at $\Delta = 0.7$, but the bias of each parameter varies with the constant and remains substantial. The "best" value for Δ is hence neither arbitrarily small nor equal to 1, contrary to common belief.

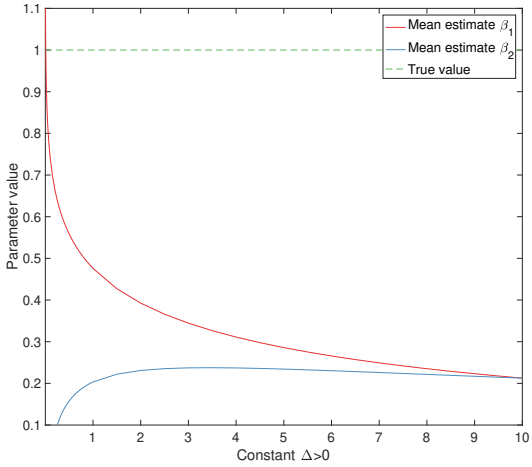


Figure 5.2: Bias against Δ

5.2.2 Other methods

Among the models which can address the log of zero, Non-linear methods are popular because they offer a valid approach in many settings. There are also approaches which should usually be avoided.

Poisson models. The model presented in (5.2.1) is non-linear in variables and parameters. The parameters are identified and non-linear estimators, such as the generalized method of moments (GMM), yield consistent estimates of β under the strict exogeneity restriction $E(U_i|X_i) = 1$ which implies the unconditional moments[11]

$$E(X_i(Y_i - \exp(X_i'\beta))) = 0. \tag{5.2.6}$$

which allow the estimation of β by maximizing the Pseudo log-likelihood of the Poisson model (Gourieroux et al., 1984). This approach is computationally efficient because it is a well-defined concave problem. Santos Silva and Tenreyro (2006) were the first to argue for Pseudo-Poisson Maximum Likelihood (PPML) as a potential solution for the appearance of zeros in Y_i. This approach is based on the additive representation of the model in (5.2.4) assuming $E(\epsilon_i|X_i) = 0$, which is equivalent to $E(U_i|X_i) = 1$.

Nevertheless, these Poisson regression has several shortcomings. First, existence of a solution is not guaranteed leading to convergence issues. Second, their precision can be sensitive to the dispersion of Y_i because of the exponential function. Third, they can be difficult to estimate with many fixed-effects. Fourth, instrumental variables require stronger assumptions and may dramatically increase computational complexity.[12]

[11] Choosing the "best" unconditional moments, or rather picking the optimal instruments, from a conditional moment restriction is beyond the scope of this paper. The interested reader is referred to Chamberlain (1987).

[12] Non-linear IV estimators require strict exogeneity between the errors and instruments unlike linear estimators.

Most of these issues have been discussed and addressed in a series of papers (Santos Silva and Tenreyro, 2010, 2011; Correia et al., 2019).[13]

Mixture models and Heckman's correction. Censorship models, such as Tobit models (Tobin, 1958), provide another non-linear solution. They consist in modeling the selection explicitly, $Y_i = 0$ or $Y_i > 0$, using a latent variable approach under chosen distributional assumptions. This approach is not often used to address the log of zeros but has been relied upon in the context of gravity equations. For example, Eaton and Tamura (1994) implement a Tobit approach to model thresholds above which trade starts to be measured.

The Heckman's ("Heckit") correction (Heckman, 1979b) is seldom used for the log of zero. In the setting provided by model (5.2.1), it assumes that $\xi_i = 1$ if $X_i'\gamma + \nu_i > 0$, and $\xi_i = 0$ otherwise. $X_i'\gamma + \nu_i$ is hence referred to as the "selection equation". The key identifying restriction is that ε_i and ν_i are bivariate normal, so that $E[\varepsilon_i|U_i > 0, X]$ admits the closed-form expression

$$E[\varepsilon_i|\nu_i > -X_i'\gamma, X] = \lambda\frac{\phi(-X_i'\gamma)}{\Phi(X_i'\gamma)}, \tag{5.2.7}$$

for $\phi(\cdot)$ and $\Phi(\cdot)$ denoting the Gaussian probability density and distribution functions, respectively. λ and γ are estimable parameters.

Estimation takes two steps. First, a probit model of $Y_i > 0$ conditional on X_i yields $\hat{\gamma}$. Second, the log-linear regression with an additional term, as specified by

$$\log(Y_i) = X_i'\beta + \lambda\frac{\phi(-X_i'\hat{\gamma})}{\Phi(X_i'\hat{\gamma})} + e_i, \tag{5.2.8}$$

is estimated by OLS to obtain β and λ. The relevance of the correction term can be tested using a t-test to check whether $\hat{\lambda}$ is different from zero. When $\hat{\lambda}$ is zero, the mechanism generating the zeros is not correlated to the outcome and OLS regression using the positive values of Y_i will provide a consistent estimate of β. Therefore, this simple two-step approach can be used to investigate whether discarding zeros would threaten identification. Note that, however, this approach is heavily dependent on the distributional assumption in absence of instrumental variables in the selection equation.

Discarding zeros. The simplest solution is to delete the zero observations and estimate (5.2.2) directly with OLS. Formally, discarding zeros introduces a selection bias unless the following condition holds,

$$E[\varepsilon|\xi = 1, X] = constant. \tag{5.2.9}$$

Similarly, one could discard zeros and estimate (5.2.4) with PPML assuming

$$E[\exp(\varepsilon)|\xi = 1, X] = constant. \tag{5.2.10}$$

Doing so assumes away any role played by the zeros and has context-dependent consequences; rendering it inadvisable at least since Young and Young (1975). At the very least, it will change the scope of the study by narrowing down the focus to observations for which $Y_i > 0$. The economic interpretation of the error term should always be discussed when making such an assumption. For instance, some empirical studies relying on the mincer equation for the purpose of estimating the returns to schooling use the log wage and discard unemployed individuals. Unemployed agents have unobserved wage rates which can be labelled as zeros. If ε_i

[13] The authors also have a dedicated website with helpful resources about Poisson regression (https://personal.lse.ac.uk/tenreyro/lgw.html).

captures the unobserved ability of individual i, it will undoubtedly be correlated with her employment outcome $\xi_i = 1$ or $\xi_i = 0$, hence introducing a sample selection bias when discarding the zeros.

Transformations. An alternative approach relies on log-like transformations applicable to non-positive values. The most popular are the "popular fix", presented earlier, and the IHS (MacKinnon and Magee, 1990; Burbidge et al., 1988; Johnson, 1949).[14] It consists in transforming Y_i into $\tilde{Y}_i = \log(\theta Y_i + \sqrt{\theta^2 Y_i^2 + 1})/\theta$ and estimating $\tilde{Y}_i = X'\beta + \omega_i$ by OLS. If the underlying model writes in log, then this transformation will likely yield biased estimates.[15] Nearly all economic applications set θ to 1 such that $\tilde{Y}$ tends toward $\log(2Y)$ for large values of Y. There is also a version with a location parameter as discussed in MacKinnon and Magee (1990). This transformation essentially consists in adding a positive *observation-specific* value to the response variable before applying the log function. Its similarity with the $\log$ function may lead to treating them interchangeably. However, for small values of Y_i, these transformations can behave differently. Besides, as shown in Bellemare and Wichman (2020), the interpretation of the coefficients is not trivial and the underlying elasticity is potentially biased or undefined.[16] It is hence satisfactory in contexts where applying a concave transformation is the main objective, e.g. for prediction models, where identification is not an issue, or when the exogeneity restriction can be justified as discussed later on.

5.3 Iterated Ordinary Least Squares (iOLS)

In this section, we develop a new approach based on the popular fix. This new approach yields a family of estimators requiring only OLS to implement. For clarity, we first show how our estimation procedure arises in the context of the log of zeros. Second, we derive its asymptotic properties. Third, we detail how minor modifications can accommodate alternative exogeneity conditions.

5.3.1 Fixing the popular fix (iOLS$_\delta$)

We let Δ_i vary across observations such that $Y_i + \Delta_i > 0$. From (5.2.5), we have

$$\log(Y_i + \Delta_i) = X_i'\beta + \log\left(U_i + \frac{\Delta_i}{\exp(X_i'\beta)}\right). \tag{5.3.1}$$

Letting $\Delta_i = \delta \exp(X_i'\beta)$, for some $\delta > 0$, this equation becomes

$$\log(Y_i + \delta \exp(X_i'\beta)) = X_i'\beta + \upsilon_i. \tag{5.3.2}$$

where the new error term $\upsilon_i = \log(\delta + U_i)$ is assumed to satisfy an exogeneity restriction (discussed later). This shows that adding a constant value to Y_i falls short of the varying $\Delta_i = \delta \exp(X_i\beta)$ required to suppress bias.

The DGP specified in (5.2.1) assumes $E[U_i] = 1$,[17] implying that the transformed error υ_i is not mean-zero. Instead, we have $E[\log(\delta + U_i)] = c$, where c is an unknown constant depending on higher-order moments

[14] An extended concave version of this transformation is provided by Ravallion (2017).

[15] Considering model (5.2.1), having consistent estimates requires a moment condition like $E(\log(\theta U_i + \frac{\sqrt{\theta Y_i^2+1}}{\exp(X_i'\beta}) | X) = 0$, which may be difficult to justify.

[16] The authors show that in $\tilde{Y}_i = X_i'\beta + \epsilon_i$, the elasticity $\hat{\zeta}_{yx} = \hat{\beta} x \frac{\sqrt{y^2+1}}{y}$ is a function of x, y, or is not defined for $y = 0$. β is an elasticity only if $x = 1$ and y is large.

[17] This assumption is only useful to identify the intercept term.

of U_i. To see this, consider the Taylor expansion of $\log(\delta + U_i)$ around $\log(1+\delta)$ to obtain

$$c = \log(1+\delta) - \frac{1}{2(1+\delta)^2} E[(U_i - 1)^2] + \frac{1}{3(1+\delta)^3} E[(U_i - 1)^3] + ..., \tag{5.3.3}$$

where the second and third terms are respectively the variance and third centered moment of U_i. The first centered moment is assumed to be zero. Thus, this transformation introduces a nuisance parameter in the form of an extra constant term.

5.3.2 Identification

Demeaning this new error term is required to identify the parameters. Let us assume the exogeneity condition $E[X_i \overline{\upsilon}_i] = 0$, where $\overline{\upsilon}_i = \upsilon_i - c$ denotes the centered error term of the linearized model. This condition yields the set of $k+1$ equations

$$E\left[X_i \left(\log(Y_i + \delta \exp(X_i'\beta)) - c\right)\right] = E\left[X_i X_i'\right]\beta, \tag{5.3.4}$$

with $k+2$ unknowns. This system identifies β only if c is known. Fortunately, the multiplicative model in (5.2.1) provides the additional restriction necessary for identification. Let us write $X_i'\beta = \beta^1 + X_i^{r'}\beta^r$, where β^1 is the constant term and the other term represents the non-deterministic part. We rewrite (5.2.1) into

$$Y_i = \exp(\beta^1 + X_i^{r'}\beta^r)U_i = \exp(\beta^1)\exp(X_i^{r'}\beta^r)U_i. \tag{5.3.5}$$

Rearranging, taking expectations and applying the log function gives the following expression for the intercept given the other parameters

$$\beta_\beta^1 = \log(E[Y_i \exp(-X_i^{r'}\beta^r)]). \tag{5.3.6}$$

Therefore, the parameters are identified and the nuisance c can be written as[18]

$$c(\beta) = E[\log(Y_i + \delta \exp(\beta_\beta^1 + X_i^{r'}\beta^r)) - \beta_\beta^1 - X_i^{r'}\beta^r]. \tag{5.3.7}$$

5.3.3 Estimation by iOLS

The following transform of the response variable yields a (seemingly) linear model:

$$\tilde{Y}_i^{iOLS_\delta}(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c(\beta) = X_i'\beta + \overline{\upsilon}_i \tag{5.3.8}$$

We refer to this model as iOLS$_\delta$, because it depends on the choice of the parameter δ, which will be discussed shortly together with the exogeneity restriction. The moment condition $E[X_i\overline{\upsilon}_i] = 0$ yields

$$\beta = E[X_i X_i']^{-1} E\left[X_i \tilde{Y}_i(\beta)\right], \tag{5.3.9}$$

which characterizes β as the solution of a fixed-point problem. Based on this insight, we propose an iterative least-squares estimator.

Algorithm 1 (iOLS estimator)**.** *The iOLS estimator is defined as the following iterative procedure:*

1. *Initialize t at 0 and let $\hat{\beta}_0$ be an initial estimate, as obtained for example with the "popular fix" estimator $\hat{\beta}^{PF} = [X'X]^{-1}X'\log(Y+\Delta) \in \mathbb{R}^K$, for some $\Delta > 0$;*

[18] In our practical implementation, we solve the identification problem by using the consistent estimator defined for any ϕ as $\hat{c}(\phi) = \frac{1}{n}\sum_{i=1}^n \log(Y_i + \delta\exp(\tilde{\phi}_\phi^1 + X_i^{r'}\phi^r)) - \frac{1}{n}\sum_{i=1}^n(\tilde{\phi}_\phi^1 + X_i^{r'}\phi^r)$, where the constant parameter estimate is replaced by the estimator $\tilde{\phi}_\phi^1 = \log(n^{-1}\sum_{i=1}^n Y_i \exp(-X_i^{r'}\phi^r))$.

2. *Transform the dependent variable into* $\tilde{Y}_{iOLS_\delta}(\hat{\beta}_t)$ *using* (5.3.8)*;*

3. *Compute the OLS estimate* $\hat{\beta}_{t+1} = [X'X]^{-1}X'\tilde{Y}(\hat{\beta}_t)$*, and update* t *to* $t+1$*;*

4. *Iterate steps 2 and 3 until* $\hat{\beta}_t$ *converges.*

We illustrate the algorithm in Figure 5.3 using the numerical simulations presented in Section 5.5 (DGP 1). The iterative estimation procedure converges to a solution within 15 to 20 iterations on average. Moreover, only a few iterations are required to suppress most of the bias of the popular fix estimator. Remark also that $X'X$ needs only be inverted once.

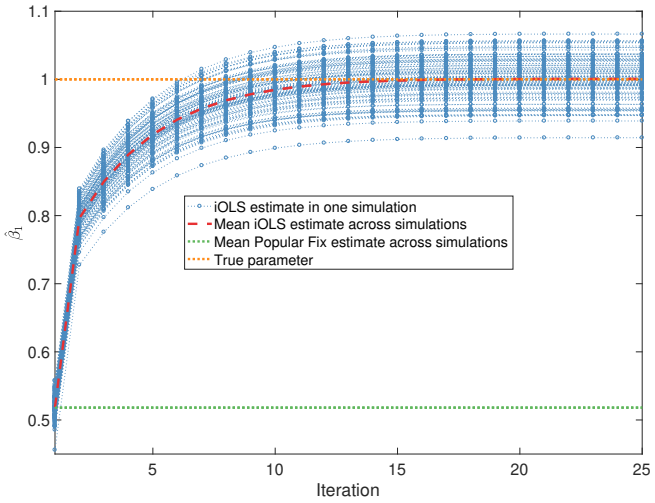


Figure 5.3: Convergence of iOLS$_{\delta=1}$ (DGP 1, n =1,000)

5.3.4 Asymptotic Properties

We establish the asymptotic properties of iOLS$_\delta$ in the following theorem.

Theorem 2 (Consistency and Normality of iOLS$_\delta$)**.** *Under the above assumptions, the iOLS*$_\delta$ *estimator is consistent and achieves the parametric rate of convergence* $n^{-1/2}$*. Formally, we have* $n^{1/2}|\hat{\beta}_{t(n)} - \beta| = O_p(1)$ *as* $n \to \infty$ *for any* $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$*, where* $\kappa \in [0,1)$ *is the modulus of the associated contraction mapping from* $\mathbb{R}^K$ *to* $\mathbb{R}^K$*. In addition, iOLS*$_\delta$ *is asymptotically normally distributed such that* $\sqrt{n}\left(\hat{\beta}_{t(n)} - \beta\right) \xrightarrow{d} N(0,\Omega)$*, as* $n \to \infty$*, where* Ω*, as given in the proof, corresponds to the asymptotic covariance of the OLS estimator in the last iteration up to minor modifications.*

This asymptotic result guarantees root-n consistent estimates and, for any fixed n, the iterative process converges after a finite number of iterations: $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, where $\kappa \in [0,1)$ is the modulus of the associated contraction mapping. The numerical convergence will hence be slower for larger sample sizes n and modulus κ closer to 1. κ depends on the DGP and is decreasing with δ. Note that there may exist values of δ such that the algorithm does not converge in finite time. This occurs when δ implies a κ very close to 1, hence a very slow convergence. However, choosing a larger δ will mechanically decrease κ and solve this issue.[19]

[19]The algorithm must include a safety check to ensure that κ is sufficiently smaller than 1. In practice, we take the median across estimates obtained at each iteration by $\hat{\kappa} = |\beta_{t+1} - \beta_t|/|\beta_t - \beta_{t-1}|$.

The asymptotic distributions of iOLS$_\delta$ and of OLS in the last iteration (once the estimator has converged) are similar. Although the standard errors of the latter are incorrect for iOLS$_\delta$, a reweighting of the corresponding covariance matrix using simple algebra is sufficient and allows to use any HAC-robust covariance estimator.[20]

5.3.5 Moment Selection

Our approach relies on an exogeneity condition about the error $\log(\delta + U_i)$. In absence of zeros, the condition is about $\log(\delta + \exp(\varepsilon_i))$ hence about ε_i when $\delta \to 0$, like in the log-linear model (5.2.2), whereas the Poisson condition is about $\exp(\varepsilon_i)$. Our understanding of the survey results presented in the introduction is that economists concerned about identification prefer conditions about ε_i rather than $\exp(\varepsilon_i)$, and that is why they often opt for the popular fix.

5.3.5.1 The role of δ

The condition $E[X_i\overline{\upsilon}_i] = 0$ is different from the condition $E[U_i|X_i] = 1$ assumed in Poisson models. Ultimately, which conditional moment restriction is satisfied depends on the context and is unverifiable *ex-ante*. However, as will be detailed later, one can test whether the restriction yields estimates that verify the implicit assumption about the pattern of zeros.

The parameter δ allows selecting a restriction among the *family* of moments $E[X_i \log(\delta + U_i)] = c$. Indeed, we observe at one extreme that when $\delta \to 0$, we have that $\lim_{\delta\to 0} \log(\delta + U)$ is exogenous to X_i, a moment condition similar to the one assumed in the standard log-linear model. At the other extreme, when $\delta \to \infty$, we have a condition equivalent to $E[X_iU_i] =$ constant, which corresponds to the (unconditional) moment condition used in (multiplicative) Poisson regressions. In other words, δ allows one to pick any condition (strictly) in-between these two extremes.

To see this, observe the Taylor expansion of $E[X_i \log(\delta + U)]$ around $U_i = 1$

$$E[X_i \log(\delta + U)] = E[X_i] \log(1 + \delta) + \left\{ \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k(1+\delta)^k} E[(U_i - 1)^k X_i] \right\}. \tag{5.3.10}$$

Assuming $E[X_i(\log(\delta + U) - c)] = 0$ to be true means the weighted sum of moment conditions between U_i and X_i is constant; where weights depend on the parameter δ. As δ goes to infinity, the weighted sum on the right-hand-side becomes negligible and $\log(\delta + U) \to \log(\delta) + U$, so the limiting moment condition is $E[X_iU_i] =$ constant.

This flexibility is of significant relevance. Indeed, the researcher usually lacks any *a priori* knowledge of the right exogeneity condition (and associated δ). We thus provide a data-driven selection method for δ based on testing the model's validy with respect to the pattern of zeros.

5.3.5.2 Poisson regression as iOLS

If the *Poisson condition* ($E[U_i|X_i] = 1$) holds in the data, then iOLS$_\delta$ will deliver a reasonable approximation for an arbitrarily large δ. Nevertheless, we now show how to enforce this condition directly in iOLS$_\delta$.

[20] A simple approximation of the standard errors for iOLS$_\delta$ consists in multiplying those of the last step OLS by a factor $1 + \delta$.

Multiplicative Poisson (iOLS$_U$). First, we consider the multiplicative version of the model. It relies on the identifying assumption $E(U_i|X_i) = 1$, but only requires $E((U_i - 1)X_i) = 0$ for consistency. To enforce this condition, we can add $\frac{1}{1+\delta}(U_i - 1)$ on both sides of (5.3.2) and rearrange to obtain

$$\log(Y_i + \delta \exp(X_i'\beta)) - \left(\log(\delta + U_i) - \frac{1}{1+\delta}(U_i - 1)\right) = X_i'\beta + \frac{1}{1+\delta}(U_i - 1). \tag{5.3.11}$$

with $U_i = Y_i \exp(-X_i'\beta)$, the second term on the left-hand-side can be rewritten into

$$c_i(\beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i \exp(-X_i'\beta) - 1), \tag{5.3.12}$$

to obtain a new transformed dependent variable

$$\tilde{Y}_i^{iOLS_U}(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c_i(\beta). \tag{5.3.13}$$

and associated model

$$\tilde{Y}_i^{iOLS_U}(\beta) = X_i'\beta + \eta_i, \tag{5.3.14}$$

where $\eta_i = \frac{1}{1+\delta}(U_i - 1)$ is a mean-zero error term, and is exogenous to X_i under the assumption $E[U_i|X_i] = 1$. This estimator will be referred to as iOLS$_U$. The choice of δ will be discussed shortly.

Additive Poisson (iOLS$_\epsilon$). Similarly, one can enforce the additive representation based on model (5.2.4), which assumes $E[\epsilon_i|X_i] = 0$, where $\epsilon_i = Y_i - \exp(X_i'\beta)$. This assumption is equivalent to $E[U_i|X] = 1$ but leads to a different least-squares objective function. iOLS can be adapted to this setting by adding and substracting $\frac{1}{1+\delta}(Y_i - \exp(X_i'\beta))$ to (5.3.2) and changing $c_i(\beta)$ in (5.3.12) into

$$c_i(\beta) = \log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i - \exp(X_i'\beta)). \tag{5.3.15}$$

This estimator, hereafter referred to as iOLS$_\epsilon$, is equivalent to PPML but can yield numerically different estimates. However, it may be less sensitive to the dispersion of the dependent variable since it does not require computing the gradient, as in PPML, or the Hessian, as in the IRLS implementation of Poisson regression (Correia et al., 2019). We derive the asymptotic properties of both estimators in the following theorem.

Theorem 3 (Consistency and Normality of iOLS$_U$ and iOLS$_\epsilon$)**.** *Under the above assumptions, the two estimators are consistent and achieve the parametric rate of convergence $n^{-1/2}$. Formally, we have $n^{1/2}|\hat{\beta}_{t(n)} - \beta| = O_p(1)$ as $n \to \infty$ for any $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, where $\kappa \in [0, 1)$ is the modulus of the associated contraction mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$. In addition, they are asymptotically normally distributed such that $\sqrt{n}\left(\hat{\beta}_{t(n)} - \beta\right) \xrightarrow{d} N(0, \Omega)$, as $n \to \infty$, where Ω, as given in the proof, differs for iOLS$_U$ and iOLS$_\epsilon$ but corresponds to the asymptotic covariance of the OLS estimator in the last iteration up to minor modifications.*

This result shows that our approach is flexible with respect to the choice of both the identifying restriction and objective criterion without significant consequences in large samples, except for minor modifications to the covariance matrix.

For both estimators, the parameter δ does not modify the relevant moment condition but is key to guarantee the convergence of the algorithm. The modulus κ is a function of δ with two important features. First, the algorithm will diverge for too small values of δ, which ultimately depends on the underlying DGP, because it implies κ above 1. Second, a too large δ implies κ very close to 1, hence a slow convergence. Therefore, the optimal δ is large enough to guarantee convergence but small enough so that convergence is fast. We address these issues by starting with a small value which we multiply by 10 if the algorithm diverges, or if our estimate

of κ is above 1, and repeat this incrementation until convergence.

Various extensions. In Appendix 5.2, we propose several extensions to the iOLS procedure. First, we adapt it to the endogenous setting in Appendix 5.2.1. Second, we address the case where discarding zeros does not jeopardize identification in Appendix 5.2.2. Third, we show in Appendix 5.2.3 how to deal with negative values in Y. Appendix 5.2.4 discusses log-log specifications with zeros in the independent variables. Appendix 5.2.5 develops a computationally fast "within" iOLS estimator to avoid the incidental parameter problem when many fixed-effects are included. Appendix 5.2.6 shows how to deal with the log of a ratio of two response variables. Finally, Appendix 5.2.7 proposes an alternative solution to use the exogeneity condition of the log-linear model $E(\varepsilon_i|X) = 0$, or an approximation of that condition.

5.4 Specification testing and model selection

Empirical researchers facing the log of zero usually compare several estimators to gauge the sensitivity of their results. Yet, each estimator is only valid under specific identifying assumptions. The latter can be systematically investigated through their implications regarding the patterns of zeros in order to substantiate the choice of an estimation procedure.

The tests developed in this section offer an opportunity for an ex-post evaluation of the identifying restrictions used for moment-based estimators. However, they are not useful to gauge the validity of the explicit distributional assumptions made in mixture models. Our tests are specification tests used to evaluate the validity of conditional moment restrictions, like $E(U_i|X_i) = 1$ for Poisson models.[21] They are, as such, similar to the RESET test of Ramsey (1969) for linear regression and its application for Poisson models by Santos Silva and Tenreyro (2006).[22] Our approach is, however, fundamentally different because it relies on testing the validity of the model with respect to the conditional probability of observing a zero. It also provides a much more powerful test of the conditional restrictions in this context as will be shown in the simulations.

A common limit of these tests is their focus on the conditional moment restrictions (e.g. $E(U_i|X_i) = 1$) rather than the unconditional restrictions (e.g. $E((U_i - 1)X_i) = 0$). The former is a *sufficient* condition whereas the latter is a *necessary* condition for consistency. We argue that statistical evidence in favor of a sufficient condition is still valuable information that the associated model bears some validity. The main issue is, however, that a rejection of the sufficient condition is not evidence against the necessary condition. In other words, these tests may lead to reject a correct model. Bearing these limits in mind, we proceed to present our methods.

5.4.1 Specification testing

Testing the Poisson condition. For clarity, we first look at the implications made by Poisson models regarding the pattern of zeros.[23] A related approach will be applied for other restrictions including for iOLS. Noting that a

[21] Santos Silva et al. (2015) proposed a radically different approach based on non-nested hypothesis tests (Davidson and MacKinnon, 1981) which consists in testing two competing models against each other.

[22] Extensions of the RESET test are proposed in Wooldridge (1997).

[23] Appendix 5.2.8 details how these tests can be implemented in the endogenous setting.

zero can only be observed if $U_i = 0$, the Poisson restriction ($E(U_i|X_i) = 1$) can be decomposed into

$$E[U_i|X_i] = E[U_i|X_i, U_i > 0]Pr(U_i > 0|X_i) = E(U_i), \tag{5.4.1}$$

since $E[U_i|X_i, U_i = 0] = 0$. There are only two possibilities for this condition to hold true. First, $E[U_i|X_i, U_i > 0]$ and $Pr(U_i > 0|X_i)$ vary with X_i in such a way that the condition holds. It happens, for example, if U_i is conditionally Poisson, or more generally, if it follows a mixture distribution with a mass probability at zero such that the condition holds. Second, this condition also holds if, instead, $E[U_i|X_i, U_i > 0]$ and $Pr(U_i > 0|X_i)$ are constant. The former is an exogeneity restriction between X_i and U_i, conditional of the error being positive, which assumes away any selection bias. The latter means that the occurrence of a zero does not depend on X_i. In this case, discarding zeros or not before estimation is irrelevant for identification.

This equation reveals the implicit relation between zeros and non-zero observations,

$$E[U_i|X_i, U_i > 0] = \frac{E(U_i),}{Pr(U_i > 0|X_i)}, \tag{5.4.2}$$

which means that the conditional error term for non-zero observations is inversely proportional to the conditional probability of having a non-zero observation.[24] We propose to investigate whether this implication matches what is observed in the data. To do so, we develop a test to assess whether the residuals implied by the chosen model satisfy this relationship where the conditional probability is estimated outside the model. This amounts to evaluating the null hypothesis[25]

$$H_0 : E[U_i|X_i, U_i > 0] = \frac{E[U]}{Pr(U_i > 0|X_i)}, \tag{5.4.3}$$

which implies that $E[U_i|X_i, U_i > 0]$ and $Pr(U_i > 0|X_i)^{-1}$ are proportional.[26]

Under the null, one can model the error term U_i as

$$U_i = \lambda E[U]Pr(U_i > 0|X_i)^{-1} + \nu_i \tag{5.4.4}$$

for $U_i > 0$ with $\lambda = 1$ and $E[\nu_i|U_i > 0, X_i] = 0$. Therefore, one can evaluate H_0 by testing whether $\lambda = 1$. This test is done in 4 steps: (1) obtain a consistent estimator of $Pr(U_i > 0|X_i)$ denoted $\hat{P}(X)$, which is possible because $U_i > 0$ if and only if $Y_i > 0$; (2) compute Poisson estimates $\hat{\beta}$ for the multiplicative model, for instance with iOLS$_U$; (3) recover the residuals $\hat{U}_i = Y_i \exp(-X'\hat{\beta})$;[27] and (4) estimate the following regression model

$$\hat{U}_i = \lambda W_i + \nu_i, \tag{5.4.5}$$

for strictly positive errors only, and where $W_i = \hat{E}[U]\hat{P}(X_i)^{-1}$ and $\hat{E}[U]$ is the unconditional mean of $\hat{U}_i$ across both positive and zero observations.

The following t-stat allows evaluating the model's validity:

$$t = \frac{\hat{\lambda} - 1}{\hat{\sigma}_\lambda}. \tag{5.4.6}$$

Under the null, the OLS estimate of λ is consistent since $\hat{U}_i$ and $\hat{P}(X_i)$ are consistent and t will hence

[24] It is worth noting that Heckman's correction model enforces a comparable conditional moment restriction: $E[\log(U_i)|X_i, U_i > 0] = \frac{\lambda\phi(-X_i'\gamma)}{Pr(U_i>0|X_i)}$. More generally, moment-based methods typically make implicit assumptions about the selection process, whereas sample-selection models enforce explicit restrictions.

[25] $E[U]$ is used to address the general framework where $E[U]$ could differ from 1.

[26] We also need to test whether the probability of observing a zero depends on any X_i by assessing $H_{0b} : Pr(U_i > 0|X_i) = p$, for any constant p. The latter is easily checked by estimating a logit or probit model and testing the statistical significance of each X_i's. The null is rejected if any coefficient is found to differ significantly from zero.

[27] For the additive model (PPML), one must use the "additive error" $\hat{\epsilon}_i + 1 = \hat{U}_i / \exp(X_i'\hat{\beta}^{PPML})$, and regress $\hat{\epsilon}_i + 1 = \lambda E[U]Pr(U_i > 0|X_i)^{-1} + \nu_i$.

converge to zero. In finite samples, however, the standard error estimates will need to be adjusted to account for the additional noise introduced by first-step estimates.[28] We opt for a pairs bootstrap to estimate this test statistic in our practical implementation. This approach yields t_{PPML} t_{iOLS_U} and t_{iOLS_ϵ}.

Testing the iOLS restriction. The same reasoning can be applied to the iOLS$_\delta$ condition $E[\log(\delta+U_i)|X_i] = c$. The null hypothesis is now

$$H_0 : E[\log(\delta + U_i)|X_i, U_i > 0] - \log(\delta) = \frac{c - \log(\delta)}{Pr(U_i > 0|X_i)}, \tag{5.4.7}$$

and the corresponding regression given by $\log(\delta + \hat{U}_i) - \log(\delta) = \lambda W_i + \nu_i$, for strictly positive errors only, where $\hat{U}_i = Y_i \exp(-X'\hat{\beta}^{iOLS_\delta})$ and $W_i = (\hat{c} - \log(\delta))\hat{P}(X_i)^{-1}$ based on $\hat{c}$ obtained from iOLS$_\delta$. The rest of the testing procedure is unchanged. This approach yields t_{iOLS_δ}.

Testing other restrictions. The same reasoning can be applied to the popular fix or the IHS. Using (5.2.5), for the popular fix, the null hypothesis becomes

$$H_0 : E[\omega_i|X_i, U_i > 0] = (X'\beta - \log(\Delta))\frac{1 - Pr(U_i > 0|X_i)}{Pr(U_i > 0|X_i)}, \tag{5.4.8}$$

and the corresponding regression model is given by $\hat{\omega}_i = \lambda W_i + \nu_i$, for strictly positive errors only. For the popular fix, we would have $\hat{\omega}_i = \log(Y_i + \Delta) - X'\hat{\beta}^{PF}$ and $W_i = (X'\hat{\beta}^{PF} - \log(\Delta))(1 - \hat{P}(X_i))\hat{P}(X_i)^{-1}$.[29] The t-stat t_{PF} and t_{IHS} are obtained as above.

Testing whether zeros can be dropped. Discarding zeros is not recommended in general but can be valid in some settings. Once zeros are dropped, researchers generally estimate either the log-linear model (5.2.2) by OLS, or PPML based on (5.2.4).

In the former case, Heckman's model is particularly useful. Statistical significance of the parameter λ associated with the correction term, using the t-stat t_{HECK}, is evidence that dropping zeros introduces a selection bias. In the latter case, one can test for such bias by substituting (5.4.4) for $\lambda = 1$ into (5.2.3) to obtain

$$Y_i = \exp(X_i'\beta)\frac{E(U)}{Pr(U > 0|X_i)}\eta_i, \tag{5.4.9}$$

where $\eta_i - 1 = \nu_i \exp(X_i'\beta)\frac{E(U)}{Pr(U>0|X_i)})^{-1}$. This expression simplifies to $Y_i = \exp(X'\beta - \log(Pr(U > 0|X_i)))\eta_i$. Therefore, testing whether zeros can be discarded from PPML is possible by estimating the augmented model on the strictly positive observations

$$Y_i = \exp(X'\beta + \theta \log(\hat{P}(X_i))\eta_i, \tag{5.4.10}$$

and evaluate $H_0 : \theta = 0$ with the t-statistic t_{PPML0} for the new regressor $\log(\hat{P}(X_i))$. Under the Poisson condition, we should observe $\theta = -1$ but any deviation from $\theta = 0$ may signal that zeros play a non-negligible role, even if $E[U|X] = 1$ does not hold.

[28] The main difficulty in deriving a closed-form expression for $\hat{\sigma}_\lambda$ is to account for the correlation between $\hat{P}(X_i)$ and $\hat{\beta}^{iOLS_U}$ which are separately estimated. We do not address this issue.

[29] For the IHS, we would use the null hypothesis $H_0 : E[\omega_i|X_i, U_i > 0] = X'\beta\frac{1-Pr(U_i>0|X_i)}{Pr(U_i>0|X_i)}$. We would have $\hat{\omega}_i = \tilde{Y}_i - X'\hat{\beta}^{IHS}$ and $W_i = X'\hat{\beta}^{IHS}(1 - \hat{P}(X_i))\hat{P}(X_i)^{-1}$.

Conditional probability estimation. Those tests require a consistent estimate of the conditional probability function $P(U > 0|\cdot)$. Specifying a parametric model, like the logit, probit or even (ex-post bounded) linear probability model, provides a simple option. However, the misspecification of $P(U > 0|\cdot)$ may distort the test's size and performance. A nonparametric estimate of the conditional probability should hence be preferred whenever possible. Although consistent, nonparametric estimate can have poor small-sample behaviors especially at the tails. We use a k-nearest neighbors (kNN) algorithm (Hastie et al., 2009) and "trim" observations associated with predicted probabilities outside the 10% and 90% quantiles to correct for this issue.[30]

5.4.2 Model selection

We propose to select the most suitable approach using the previous tests.

First, iOLS$_\delta$ for any $\delta \in (0, \infty)$ is based on condition (A1):

$$E[\log(\delta + U)|X] = constant, \tag{A1}$$

which depends on the choice of δ. The "best" model within this category minimizes t^{iOLS_δ} with respect to δ. This rule will select the model with the least deviation between the implied and observed patterns of zeros. Second, PPML, iOLS$_U$ and iOLS$_\epsilon$ are based on condition (A2):

$$E[U|X] = constant, \tag{A2}$$

Third, OLS and PPML without zeros (or iOLS$_S$ in the Appendix 5.2.2) are based on either:

$$E[\log(U)|U > 0, X] = constant, \tag{A3}$$

or $E[U|U > 0, X] = constant$, which states that zeros can be discarded. Fourth, the Heckman's correction model is based on

$$(\varepsilon_i, \nu_i)' \sim \text{bivariate Gaussian}, \tag{A4}$$

which is not readily testable. Fifth, the popular fix or the IHS transformation relies on assumptions of the form

$$E[\omega_i|X] = constant, \tag{A5}$$

where ω_i is a known function of U_i, X_i and β.

We propose a model selection procedure predicated on first using models based on moment conditions rather than explicit distributional assumptions. This implies that we advocate for using more complex estimators only when the simpler ones are rejected. The selection procedure is as follows:

1. Compute t_{iOLS_U}, t_{PPML}, t_{iOLS_δ} for a range of δ, t_{PF} and t_{IHS}, and select the model with the smallest t-statistic in absolute value, denoted $|t_1|$. If $|t_1| < 1.96$, stop and select this model;

2. Else, compute t_{PPML0}, t_{HECK} and take the maximum in absolute value to define $|t_2|$. If $|t_2| < 1.96$, stop and report the estimates of PPML and OLS without zeros;[31]

3. Else, select Heckman's correction model or another mixture model.

[30] See Cameron and Trivedi (2005) section 9.5.3 for a discussion of this common practice in nonparametric estimation.

[31] By taking the maximum of the two t-stats, we require that both tests suggest that zeros can be dropped before recommending to do so. The rationale is that both tests have power against different alternatives, hence combining them enlarges statistical power.

5.5 Simulations

Let us specify the dependent variable as

$$Y_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}) U_i, \tag{5.5.1}$$

where $\beta_0 = \beta_1 = \beta_2 = 1$, $U_i = \exp(\varepsilon_i)\xi_i$ with $\xi_i = 0$ or 1, and $Pr(\xi_i = 0|X_i) = P(X_i) = \frac{1}{1+\exp(\gamma_0+\gamma_1 X_{1i}+\gamma_2 X_{2i})}$, with $\gamma_0 = -0.5$, $\gamma_1 = 0.5$ and $\gamma_2 = -0.5$. We consider six DGPs specified as follows:

- DGP 1 (A1): $E[X_i'(\log(1+U_i) - c)] = 0$. This DGP is useful to illustrate iOLS$_\delta$ with $\delta = 1$. Let us assume that $\log(1+\varepsilon_i)$ is uniformly distributed as $U[\frac{c}{2P(X_i)}, \frac{3c}{2P(X_i)}]$ with X_{1i} and X_{1i} also uniformly distributed as $U[-1, 2]$. Choosing $c = 0.41512$ yields the desired condition $E[X_i'(\log(1+U_i)-c)] = 0$ with $E(U_i) = 1$.

- DGP 2 (A2): $E[U_i|X_i] = 1$. This DGP is aimed at comparing the alternative modelling approaches to PPML. We assume that $(X_{1i}, X_{2i})'$ is bivariate normal with mean zero, variance $\sigma^2_{X1} = \sigma^2_{X2} = 1$ and covariance $\sigma_{X1X2} = -0.3$. We further assume that ε_i is Gaussian with mean $-\log(P(X_i)) - 1/2$ and variance 1 so that $\exp(\varepsilon_i)$ is log-normal with conditional mean $1/P(X_i)$.

- The other DGPs are detailed in the Appendix. DGP 3 (A3) ($E[U_i|U_i > 0, X_i] = 1$) is such that discarding zeros and using PPML yields consistent estimates. DGP 4 (A4) is such that Heckman's model applies. DGP 5 (A5) is designed so that applying the IHS transform yields consistent OLS estimates. Finally, DGP 6 (IV) assumes $E[U_i|X_i] \neq 1$ but $E[U_i|Z_i] = 1$ which corresponds to the Poisson condition with endogenous regressors.

We simulate 10,000 times each DGP, for two sample sizes ($n = 1,000$ and $n = 10,000$), and report the mean and standard deviations for the following estimators: iOLS$_{\delta=1}$, iOLS$_{\delta=100}$, iOLS$_U$ (multiplicative Poisson), iOLS$_\epsilon$ (additive Poisson), PPML (additive Poisson), iOLS$_S$ (see Appendix 5.2.5), OLS and PPML without zeros (PPML0), Heckman's corrected model, OLS after performing the inverse hyperbolic sine transform (IHS), and the popular fix with $\Delta = 0.7$ (PF).[32] For DGP 6, where regressors are endogenous, we report the 2SLS analog of the above estimators. The results for DGP 3 to 6 are included in Appendix 5.3.

[32] This is the "best" value for Δ which minimizes the mean square bias, see Section 5.2.1.

Table 5.1: Simulations: DGP 1 (A1: $E[X_i'(\log(1+U_i)-c)]=0$)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	0.99	1.01	0.99	1.00	1.00	1.00
		(0.09)	(0.10)	(0.10)	(0.03)	(0.03)	(0.03)
	$iOLS_{\delta=100}$	0.92	0.72	1.28	0.92	0.72	1.28
		(0.07)	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)
(A2)	$iOLS_U$	0.91	0.71	1.29	0.92	0.70	1.30
		(0.07)	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)
	$iOLS_\varepsilon$	1.00	0.61	1.27	1.01	0.60	1.28
		(0.14)	(0.13)	(0.12)	(0.04)	(0.04)	(0.04)
	$PPML$	1.00	0.61	1.27	1.01	0.60	1.27
		(0.14)	(0.13)	(0.12)	(0.04)	(0.04)	(0.04)
(A3)	$iOLS_S$	0.97	0.46	1.54	0.97	0.46	1.54
		(0.06)	(0.04)	(0.04)	(0.02)	(0.01)	(0.01)
	OLS	1.77	0.50	1.50	1.77	0.50	1.50
		(0.04)	(0.03)	(0.04)	(0.01)	(0.01)	(0.01)
	$PPML0$	2.04	0.27	1.58	2.05	0.26	1.58
		(0.08)	(0.09)	(0.07)	(0.02)	(0.03)	(0.02)
(A4)	$Heckman$	−8.18	2.48	−0.48	−7.92	2.47	−0.47
		(3.48)	(0.56)	(0.56)	(1.02)	(0.17)	(0.17)
Others	$IHST$	0.89	0.56	0.18	0.89	0.56	0.18
		(0.05)	(0.07)	(0.07)	(0.02)	(0.02)	(0.02)
	PF	0.46	0.52	0.18	0.46	0.52	0.18
		(0.05)	(0.06)	(0.06)	(0.01)	(0.02)	(0.02)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP1. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard errors are presented in parentheses.

Bias and variance. Table 5.1 reports the results for DGP 1 based on the true identifying condition $E[X_i'(\log(1+U_i)-c)]=0$. All estimators but iOLS$_{\delta=1}$ are biased, confirming that the identifying conditions of iOLS$_\delta$ indeed differ from those assumed by PPML. This bias is severe for the inverse hyperbolic sine transformation, the popular fix, and the Heckman correction. PPML exhibits a smaller bias than existing alternative estimators and is found to have greater variance than iOLS$_U$. As expected, iOLS$_\epsilon$ corresponds exactly to PPML estimates. These results also illustrate the $\sqrt{n}$-consistency of the estimators as the standard errors are divided by 10 as the sample size increases by 100-fold.

Table 5.2: Simulations: DGP 2 (A2: $E[U_i|X_i] = 1$)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	1.06	1.26	0.74	1.08	1.25	0.75
		(0.19)	(0.11)	(0.11)	(0.06)	(0.04)	(0.04)
	$iOLS_{\delta=100}$	0.98	1.03	0.97	1.00	1.02	0.98
		(0.17)	(0.10)	(0.10)	(0.05)	(0.03)	(0.03)
(A2)	$iOLS_U$	0.98	1.01	0.99	1.00	1.00	1.00
		(0.17)	(0.10)	(0.10)	(0.05)	(0.03)	(0.03)
	$iOLS_\varepsilon$	1.02	0.99	0.97	1.00	1.00	1.00
		(0.47)	(0.17)	(0.21)	(0.19)	(0.06)	(0.09)
	$PPML$	1.02	0.99	0.97	1.01	1.00	0.99
		(0.47)	(0.17)	(0.21)	(0.19)	(0.06)	(0.09)
(A3)	$iOLS_S$	1.08	0.74	1.27	1.09	0.73	1.27
		(0.14)	(0.07)	(0.07)	(0.04)	(0.02)	(0.02)
	OLS	1.52	0.73	1.27	1.52	0.73	1.27
		(0.10)	(0.06)	(0.05)	(0.03)	(0.02)	(0.02)
	$PPML0$	2.05	0.69	1.29	2.05	0.69	1.30
		(0.36)	(0.14)	(0.17)	(0.15)	(0.06)	(0.07)
(A4)	$Heckman$	-1.45	1.30	0.70	-1.40	1.30	0.70
		(1.87)	(0.35)	(0.35)	(0.56)	(0.11)	(0.11)
Others	$IHST$	0.82	0.72	0.05	0.82	0.72	0.05
		(0.11)	(0.07)	(0.07)	(0.03)	(0.02)	(0.02)
	PF	0.38	0.68	0.06	0.38	0.68	0.06
		(0.10)	(0.07)	(0.07)	(0.03)	(0.02)	(0.02)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP2. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard errors are presented in parentheses.

Table 5.2 reports the results for DGP 2 when $E[U_i|X_i] = 1$ is the true identifying condition. We first observe that only PPML, iOLS$_U$ and iOLS$_\epsilon$ are consistent. Second, we nonetheless see that iOLS$_{\delta=100}$ exhibits a small bias but does not exclude the true value from its confidence interval. Third, we note that iOLS$_U$ dominates PPML in terms of precision under this simulation design. Finally, we observe that the estimates of PF and IHS have large biases.

Tests and model selection. The simulations are also useful to study our testing procedures. The conditional probabilities to have a zero are logistic in all DGPs but (A4). In what follows, we mainly focus on the correct parametric specification to compute the conditional probability of observing zero values (logit). We also report and discuss some results when using a nonparametric approach (kNN) or a misspecified model (probit).

First, we report the frequency of selecting each δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ for DGPs 1 and

2, based on the smallest t-stat (section 5.4.1), When the true restriction is that of iOLS$_{\delta=1}$ (DGP 1), this approach selects $\delta = 1$ correctly 50% of the time for $n = 10,000$ as opposed to 17% of the time when $n = 1000$. In comparison, when the exogeneity condition of PPML is correct (DGP 2), a large δ is selected in most simulations.[33].

Table 5.3: Simulations: Data-driven selection of δ (iOLS$_\delta$)

		δ						
n	DGP	0.1	0.5	1	5	10	50	100
1000	1	28%	18%	17%	16%	9%	5%	7%
	2	14%	9%	8%	9%	7%	5%	48%
10,000	1	3%	32%	50%	14%	1%	0%	0%
	2	0%	0%	2%	8%	14%	19%	57%

Notes: This table shows the relative frequency with which a given δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ was chosen on the basis of the 10,000 simulations. These simulations vary by sample size n and by DGP. These test assume the probability model to be logistic. Interpretation: when the sample size is n=10,000 and the data was generated using DGP1, $t_{iOLS_{\delta=1}}$ was the smallest 50% of the time, so $\delta = 1$ was selected 50% of the time.

Second, we show the empirical size and power of each test for all DGPs in Table 5.4 for a nominal size of 5%. In DGP 1, the t-test $t_{\delta=1}$ rejects $\delta = 1$ only 5% of the time which corresponds to the nominal test size. The tests for $iOLS_{\delta=100}$, $iOLS_U$ and $PPML$ have power against this alternative with a 100% rejection rate in large samples. The results for DGP 2 show that the tests for iOLS$_{\delta=100}$, iOLS$_U$, PPML and iOLS$_\epsilon$ are correctly sized, and that the other tests have satisfactory power.[34] Finally, the test for IHS is correctly sized in DGP 5. It is worth noting that not all tests have power against each of the considered alternative, even in large samples. This is the case for $t_{\delta=1}$ and t_{PF} in DGP 5. Finally, the last column reports the empirical rejection rates of the RESET test for PPML, including 3 polynomials terms (Santos Silva and Tenreyro, 2006; Ramsey, 1969). Findings reveal that the test is slightly oversized and lacks power against all considered alternatives. For comparison, we report the empirical sizes and powers when using kNN instead of logit in Table 5.5.[35] The tests' sizes are slightly distorted but exhibit satisfactory power.

[33] Results are similar for kNN and Probit (Appendix 5.3)

[34] Although not reported here for readability, t_{HECK} and t_{PPML0} are correctly sized with only 5% rejection rates in DGP 3 where zeros can be dropped, and exhibit some power under the alternatives.

[35] Results for Probit are reported in Table 5.3.7.

Table 5.4: Simulations: Specification testing (Logit)

n	DGP	$t_{\delta=1}$	$t_{\delta=100}$	t_U	t_ε	t_{PPML}	t_{IHST}	$RESET$
	1	6%	26%	25%	57%	57%	19%	12%
	2	9%	5%	5%	4%	4%	22%	6%
1000	3	11%	8%	8%	9%	9%	18%	5%
	4	10%	7%	7%	13%	13%	39%	6%
	5	6%	7%	7%	19%	19%	6%	9%
	1	5%	100%	100%	100%	100%	87%	9%
	2	47%	5%	5%	5%	5%	100%	7%
10,000	3	75%	33%	39%	62%	62%	100%	6%
	4	61%	25%	30%	71%	70%	100%	6%
	5	6%	22%	24%	94%	94%	5%	9%

Notes: This table shows the relative rejection frequency of each null hypothesis for 10,000 simulations. These simulations vary by sample size (as reported by the column "n") and by Data Generating Process (as reported in the column "DGP"). These test assume the probability model to be logistic. RESET refers to the t-statistic associated with the joint significance of three polynomial terms. Interpretation: when the sample size is n=1000 and the data was generated using DGP1, $t_{\delta=1}$ was rejected 6% of the time.

Finally, Table 5.6 reports selection rates using our procedure. The correct model is chosen more often as the sample size enlarges. However, the estimates' precision largely drives this selection. For example, the Heckman model is only selected 14% of the time when n=10,000 in DGP 4, requiring more observations to approach 100%.

Table 5.5: Simulations: Specification testing (kNN)

n	DGP	$t_{\delta=1}$	$t_{\delta=100}$	t_U	t_ε	t_{PPML}	t_{IHST}	$RESET$
	1	8%	16%	18%	9%	9%	6%	12%
	2	39%	9%	8%	6%	6%	14%	6%
1000	3	66%	37%	35%	16%	16%	39%	5%
	4	62%	29%	26%	9%	9%	70%	5%
	5	5%	6%	6%	4%	4%	6%	9%
	1	8%	98%	99%	63%	63%	21%	9%
	2	98%	12%	8%	7%	7%	71%	7%
10,000	3	100%	95%	93%	53%	53%	100%	6%
	4	100%	93%	88%	23%	23%	100%	6%
	5	7%	13%	14%	5%	5%	5%	9%

Notes: This table shows the relative rejection frequency of each null hypothesis for 10,000 simulations. These simulations vary by sample size (as reported by the column "n") and by Data Generating Process (as reported in the column "DGP"). These test are based on a non-parametric KNN probability model, trimmed of the top and bottom 10% observations. RESET refers to the t-statistic associated with the joint significance of three polynomial terms. Interpretation: when the sample size is n=1000 and the data was generated using DGP1, $t_{\delta=1}$ was rejected 8% of the time.

Table 5.6: Simulations: Model selection

n	DGP	$(A1)$	$(A2)$	$(A3)$	$(A4)$	$(A5)$
1000	1	52%	25%	0%	0%	23%
	2	24%	67%	4%	0%	5%
	3	18%	60%	20%	0%	3%
	4	25%	56%	19%	0%	0%
	5	14%	14%	63%	0%	9%
10,000	1	93%	0%	0%	4%	2%
	2	5%	95%	0%	0%	0%
	3	0%	31%	67%	2%	0%
	4	1%	67%	19%	13%	0%
	5	41%	14%	3%	0%	41%

Notes: This table shows the selection frequency of each identifying restriction for 10,000 simulations. These simulations vary by sample size (as reported by column "n") and by Data Generating Process (as reported in column "DGP"). Selection is done assuming the probability model to be logistic. Interpretation: when the sample size is n=1000 and generated by DGP1, a model with moments (A1) is chosen 51% of the time.

5.6 Application

We now revisit three empirical studies published in top-tier economic journals where the log of zero had to be addressed. First, Santos Silva and Tenreyro (2006) compare various estimators to estimate gravity models of trade. Second, Michalopoulos and Papaioannou (2013) use the popular fix to examine the role of pre-colonial ethnic institutions on economic development. Third, Card and DellaVigna (2020) investigate the preferences of academic journal editors with the IHS transformation in the context of endogenous regressors.

For brevity, we report only the main estimates for the most relevant estimators. We focus on the data-driven selected δ for iOLS$_\delta$, along with iOLS$_U$, PPML and IHS. Standard errors, as reported in parenthesis, are obtained using 300 pairs bootstrap. Comprehensive results for all estimators and tests discussed in the paper are given in the Appendix.

5.6.1 Santos Silva and Tenreyro (2006)

First, we study the gravity model of Santos Silva and Tenreyro (2006). Their Table 3 reports models of bilateral trade for data covering 136 countries in 1990. For importer (I) and exporter (X) countries, they control for log(GDP), log(GDP per capita), log(Distance), along with dummies for contiguity, shared language, colonial ties, access to the oceans, remoteness (which measures the access to other trading partners), free trade agreement (FTA), and the existence of a preferential trade agreement. They advocate for PPML over a log-linear model arguing that the latter is biased in presence of heteroskedastic errors.

We report the t-statistic and main estimates in Table 5.7 in order to compare the different approaches.[36] The tests provide evidence against iOLS estimators but fail to reject PPML. Although the test for PPML0 fail to reject that zeros can be discarded, the test for Heckman provides evidence of the opposite. Therefore, dropping

[36] The associated $\hat{\lambda}$ statistics for all specifications are provided in Table 5.4.4 in the Appendix.

zeros is not recommended and PPML should be preferred when using the parametric test with logit probability. However, it should be noted that using a nonparametric approach (kNN) to estimate the conditional probability of observing zeros for the test provides a different conclusion. PPML is no longer the preferred model and is rejected in favor of iOLS$_{\delta=100}$ and iOLS$_U$ (see Table 5.4.4 in the Appendix).

Table 5.7: Estimates from Santos Silva and Tenreyro (2006)'s Table 3

	iOLS$_{\delta=100}$	PPML*	iOLS$_U$	Heckman	PPML0
Log(Distance)	-1.52	**-0.78**	-1.48	-1.26	-0.78
	(0.08)	**(0.06)**	(0.08)	(0.04)	(0.06)
Contiguity	0.13	**0.19**	0.20	0.15	0.20
	(0.38)	**(0.10)**	(0.38)	(0.13)	(0.10)
Language	0.76	**0.75**	0.69	0.77	0.75
	(0.13)	**(0.14)**	(0.13)	(0.07)	(0.14)
Colonial	0.41	**0.03**	0.39	0.44	0.02
	(0.15)	**(0.15)**	(0.15)	(0.07)	(0.15)
FTA	1.45	**0.18**	1.55	0.46	0.18
	(0.49)	**(0.10)**	(0.57)	(0.11)	(0.10)
$\hat{\lambda}$	0.46	**1.26**	0.44	0.81	-0.22
	(0.06)	**(0.39)**	(0.06)	(0.09)	(0.37)
t-Stat.	[-9.82]	**[0.68]**	[-9.24]	[8.75]	[-0.59]

Notes: This table displays the main coefficients, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for several models of trade gravity, based on using a logistic probability model. The full list of control variables is provided in Section 5.6.1. $iOLS_{\delta=100}$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. The symbol *denotes the specification recommended by the authors in their original article. Our preferred specification (i.e. with the smallest t-stat) is in bold.

5.6.2 Michalopoulos and Papaioannou (2013)

Michalopoulos and Papaioannou (2013) examine the relationship between pre-colonial political centralization and contemporary development in African countries. The latter is proxied using light density at night at the regional level and used as the response variable through the "popular fix": $\log(0.01+Y_i)$. The authors focuses on the coefficient associated with Murdock's 1967 index of jurisdictional hierarchy.[37] The cross-sectional unit is ethnicity-by-country. They control cumulatively for population density, location, and geography,[38] and find positive and significant estimates.

[37] Ranging between 0 and 4, it provides the number the number of jurisdictions above the local level for each ethnicity as reported in 1967. A large number indicates the presence of a centralized political organization.

[38] We focus on columns (2)-(4) of their Table 2. Full results along with a replication of their Table 3 (Panel A, column (1)-(4)), which includes additional country fixed effects, are provided in the Appendix. The same conclusions apply.

Table 5.8: Estimates from Michalopoulos and Papaioannou (2013)'s Table 2

	PF*	iOLS$_{\delta=0.05}$	iOLS$_{\delta=100}$	PPML	iOLS$_U$
Pop.					
$\hat{\beta}$	0.35	**0.53**	0.44	0.29	0.41
	(0.07)	**(0.11)**	(0.13)	(0.12)	(0.14)
$\hat{\lambda}$	-0.88	**1.01**	1.08	2.62	1.09
	(0.29)	**(0.02)**	(0.06)	(0.57)	(0.07)
t-Stat.	[-6.40]	**[0.67]**	[1.36]	[2.85]	[1.30]
Pop. & Loc.					
$\hat{\beta}$	0.32	**0.40**	0.36	0.14	0.35
	(0.06)	**(0.09)**	(0.09)	(0.11)	(0.10)
$\hat{\lambda}$	-0.56	**1.00**	1.03	3.68	1.03
	(0.24)	**(0.04)**	(0.05)	(1.24)	(0.05)
t-Stat.	[-6.58]	**[-0.04]**	[0.58]	[2.16]	[0.55]
Pop. & Loc. & Geo.					
$\hat{\beta}$	0.19	0.09	0.11	0.00	**0.10**
	(0.05)	(0.11)	(0.09)	(0.10)	**(0.09)**
$\hat{\lambda}$	-0.18	0.82	0.85	1.94	**0.85**
	(0.12)	(0.19)	(0.21)	(0.91)	**(0.22)**
t-Stat.	[-9.46]	[-0.91]	[-0.69]	[1.04]	**[-0.68]**

Notes: This table displays the coefficient associated with jurisdictional hierarchy, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. These tests are based on using a logistic probability model. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). Three specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). Full estimates are available in Table 5.4.8 of the Appendix. The symbol *denotes the specification used by the authors in their original article. Our preferred specifications (i.e. with the smallest t-stat) are in bold.

The results are reported in Table 5.8. The popular fix provides overly precise estimates but is always rejected by our tests with λ far from 1. In comparison, iOLS yields values of λ fairly close to 1. Adding geographic controls lowers the estimated β in all cases to the point where it is not statistically significant anymore. We fail to reject PPML in this specification although with a fairly imprecise estimate of λ. Those results reveal a much weaker statistical relationship between the variables under study than considered in the original paper.[39] Interestingly, the best model suggested by the test differs depending on the explanatory variables included in the specification. Indeed, the exogeneity condition directly depends on X, which implies that the right model to use for the estimation of the effects can change from one specification to another, simply by adding a new variable in X, even though these specifications are very close.[40]

[39] In subsequent work, Michalopoulos and Papaioannou (2014) study the link between contemporary political institutions in Africa and economic development. They find an absence of statistical significance using both the popular fix, OLS in level, and PPML (Table 6 in the Appendix). In contrast, we find an absence of significance in terms of pre-ethnic political hierarchy and economic development.

[40] As seen in Table 5.4.7 in the Appendix, the nonparametric probability model used for the tests yields qualitatively similar results for our main specification of interest (i.e, with population, location, and geographical controls). Indeed, iOLS$_{\delta=100}$ and iOLS$_U$ remain our favored specification, associated in both cases with λ close to one.

5.6.3 Card and DellaVigna (2020)

Finally, we revisit Card and DellaVigna (2020)'s study of journal editors. The data contains submission-level information from four leading economics journal matched to Google Scholar citations. The authors address the log of zero with the IHS transformation and study the role of many control variables. For simplicity, we focus our attention on measuring the impact of receiving an invitation to *Revise & Resubmit* on the number of citations, which is considered an endogenous variable. The authors take a control function approach, instrumented using the "leave-out mean R&R rate of the editor".[41]

Table 5.9 reports estimates in three cases: without correcting for the endogeneity, using the control function approach, and with instrumental variables. In all cases, $iOLS_{\delta=50}$ is selected and PPML is rejected. The IHS is also rejected although with a λ fairly close to 1.[42] Accounting for the endogeneity of this variable yields a lower estimate, even negative when using iOLS. Yet, it is never statistically significant.

This effect is interpreted as the mechanical publication effect and has a positive sign in the original paper: an invitation for R&R should yield additional citations. Although not statistically significant, its sign changes when using $iOLS_{\delta}$ and $iOLS_{U}$ or using 2SLS instead of the control function. We interpret this negative sign as follows. An editor which is more likely to offer R&R will mechanically do so for papers with lesser potential to attract citations. Specification tests presented in the Appendix point in favor of $iOLS_{\delta=50}$ in all specifications, and reject all other estimators based on the Poisson condition or discarding zeros.

[41] This instrument measures the frequency with which the same editor has invited *other* authors to revise their manuscript before reassessment.

[42] The kNN nonparametric probability model favors iOLS$_{\delta=50}$ in the instrumental variable case but rejects all models in the other cases, as shown in Table 5.4.14 in the Appendix.

Table 5.9: Estimates from Card and DellaVigna (2020)

	IHS*	$\text{iOLS}_{\delta=50}$	PPML	iOLS_{U}
No correction for Endogeneity				
$\hat{\beta}$	0.57	**0.48**	0.53	0.48
	(0.05)	**(0.04)**	(0.04)	(0.04)
$\hat{\lambda}$	0.97	**1.00**	1.12	1.00
	(0.00)	**(0.00)**	(0.01)	(0.00)
t-Stat.	[-6.64]	**[0.73]**	[8.63]	[-3.06]
Control Function				
$\hat{\beta}$	0.07	**-0.09**	0.11	-0.08
	(0.14)	**(0.13)**	(0.13)	(0.13)
$\hat{\lambda}$	0.97	**1.00**	1.13	1.00
	(0.00)	**(0.00)**	(0.02)	(0.00)
t-Stat.	[-5.84]	**[0.83]**	[8.36]	[-2.93]
Instrumental Variables				
$\hat{\beta}$	-1.77	**-1.20**	-1.36	-1.11
	(1.32)	**(1.86)**	(0.91)	(1.76)
$\hat{\lambda}$	0.97	**1.00**	1.14	1.00
	(0.01)	**(0.10)**	(0.02)	(0.00)
t-Stat.	[-3.53]	**[-0.04]**	[7.54]	[-0.88]

Notes: This table displays the coefficient associated with an invitation to revise & resubmit (R&R), standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of citations. $iOLS_{\delta}$ and $iOLS_{U}$ are defined in Section 5.3. Three specifications are presented: no correction for endogeneity contrasts with control function and instrumental variables which rely on the Editor leave-out mean R&R rate for identification. The symbol *denotes the specification used by the authors in their original article. Our preferred specifications (i.e. with the smallest t-stat) are in bold.

5.7 Conclusion

This paper developed multiple contributions to address a common yet unresolved issue faced in empirical research: the log of zero. First, we have attempted to clarify some issues and misconceptions with respect to existing practices, such as adding an arbitrary constant to the dependent variable. Second, we have derived a new family of estimators to estimate log-linear models when the dependent variable can take non-positive values. Those estimators have several advantages, including: 1) computational simplicity, 2) a natural extension to instrumental variables, 3) robustness to the inclusion of many fixed effects, and 4) their flexibility to exogeneity restrictions. Third, we have developed testing procedures to verify the underlying exogeneity restrictions imposed by our estimators and other well-known approaches, such as PPML or IHS. We show how these tests can be helpful to guide empirical research. Fourth, all methods are illustrated through numerical simulations and replications of recent publications in top-tier economics journals. We find that the exogeneity restrictions used by our estimators are rarely rejected and often selected as the best solution in those applications.

The main takeaway from our research should be that no single method works for all settings, hence different methods can lead to different conclusions. Hopefully, empirical researchers are now better equipped

to substantiate their preferred method in any given setting. The methodology developed in this paper should help find a consensus among practitioners about the best practice to address the log of zero. There are also many possible extensions, including semi-parametric models of sample selection and regularized models like the lasso, which we leave for future research.

5.A Mathematical Appendix

***Proof of Theorem 2**: Consistency.* Recall that the parameter $\beta \in \mathbb{R}^K$ is characterized by the fixed-point equation

$$\beta = E[X_i X_i']^{-1} E\left[X_i \tilde{Y}_i(\beta)\right], \tag{5.A.1}$$

where $\tilde{Y}_i(\beta) = \log(Y_i + \exp(X_i'\beta)) - c(\beta)$ is the transformed dependent variable. The mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$ which characterizes the parameter is hence defined $\forall \phi \in \mathbb{R}^K$ as

$$M(\phi) = E[X_i X_i']^{-1} E\left[X_i \tilde{Y}_i(\phi)\right]. \tag{5.A.2}$$

The sample counterpart of this mapping is given by

$$\hat{M}_n(\phi) = [X'X]^{-1} X' \hat{\tilde{Y}}_i(\phi), \tag{5.A.3}$$

where $\hat{\tilde{Y}}_i(\phi) = \log(Y_i + \exp(X_i'\phi)) - \hat{c}(\phi)$, with $\hat{c}(\phi) = \frac{1}{n}\sum_{i=1}^n \log(Y_i + \exp(\hat{\phi}_1(\phi) - \phi_1 + X_i'\phi)) - \log(\frac{1}{n}\sum_{i=1}^n (\hat{\phi}_1(\phi) - \phi_1 + X_i'\phi))$ for $\hat{\phi}_1(\phi) = \log(n^{-1}\sum_{i=1}^n Y_i \exp(-X_i\phi + \phi_1))$

Our proof follows Dominitz and Sherman (2005) who develop a convergence theory for iterative estimators. Following their theory, the convergence of iOLS requires that $M(\cdot)$ and $\hat{M}_n(\cdot)$ be contraction mappings, asymptotically.[43]

In order to show the convergence result $n^{1/2}|\hat{\beta}_{t(n)} - \beta| = O_p(1)$ as $n \to \infty$ by applying Theorem 1 in Dominitz and Sherman (2005), we need to show that the following conditions hold:

(i) $\left\{\hat{M}_n(\cdot) : n \geq 1, \omega \in S\right\}$ is an asymptotic contraction mapping on (B_0, E_K), where S is a sample space, E_K is the Euclidean metric on R^K and B_0 is the closed ball centered at β_0 of radius $|\hat{\beta}_0 - \beta|$;[44]

(ii) $n^{1/2}|\beta_{t(n)} - \beta| = O_p(1)$ as $n \to \infty$;

(iii) $n^{1/2} \sup_{\phi \in B_0} |\hat{M}_n(\phi) - M(\phi)| = O_p(1)$ as $n \to \infty$; and

(iv) $\sup_{\phi \in B_0} ||\hat{V}_n(\phi) - V(\phi)|| = o_p(1)$ as $n \to \infty$.

For condition (i), we adapt the proof of Lemma 5 in Dominitz and Sherman (2005) as follows. The first step is to consider that X is prewhitened so that $X'X = nI_k$. This assumption is useful to establish the local contraction mapping property. From a multivariate Taylor expansion argument, Dominitz and Sherman (2005) show that condition (i) boils down to showing that the largest eigenvalue of $\nabla_\phi \hat{M}_n(\beta) = \hat{V}_n(\beta)$ is strictly less that unity as $n \to \infty$. Note that we have

$$\begin{aligned} \hat{V}_n(\phi) &= [X'X]^{-1} X' \nabla_\phi \hat{\tilde{Y}}(\phi) \\ &= n^{-1} X' \nabla_\phi \hat{\tilde{Y}}(\phi), \end{aligned} \tag{5.A.4}$$

[43] The reader is referred to Dominitz and Sherman (2005) for a formal definition of an asymptotic contraction mapping.

[44] Note that Dominitz and Sherman (2005)'s condition (i) is about $M(\cdot)$ and not $\hat{M}_n(\cdot)$. However those conditions imply each other under conditions (iii) and (iv) by applying their Lemma 3 with trivial modifications.

where the second equality uses prewhitening and $\nabla_\phi \hat{\hat{Y}}_i(\phi)$ has element (i,k) defined as

$$\left[\nabla_\phi \hat{\hat{Y}}(\phi)\right]_{i,k} = \frac{\exp(X_i'\phi)X_{ki}}{Y_i + \exp(X_i'\phi)} - \frac{\partial \hat{c}(\phi)}{\partial \phi_k}. \tag{5.A.5}$$

Let us denote $X_{1i} = 1$, for all i as the constant. By prewhitening, we have $\sum_{j=1}^n X_{1j} = n$ and $\sum_{j=1}^n X_{kj} = 0$ for $k > 1$.

$$\frac{\partial \hat{c}(\phi)}{\partial \phi_k} = n^{-1}\sum_{i=1}^n \frac{\exp(X_i^{r'}\phi^r + \hat{\phi}^1)(\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki})}{Y_i + \exp(X_i^{r'}\phi^r + \hat{\phi}^1)} - n^{-1}\sum_{i=1}^n (\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki}), \tag{5.A.6}$$

for $k > 1$ and $\frac{\partial \hat{c}(\phi)}{\partial \phi_1} = 0$. This expression simplifies when evaluated at $\phi = \beta$, as shown by

$$\frac{\partial \hat{c}(\beta)}{\partial \phi_k} = n^{-1}\sum_{i=1}^n \frac{X_{ki}}{1+U_i} + O_p(1), \tag{5.A.7}$$

for $k > 1$ because $\hat{\phi}^1(\beta) = \log(n^{-1}\sum_{i=1} Y_i \exp(-X_i^{r'}\beta^r)) = \beta_1 + \log(n^{-1}\sum_{i=1} U_i)$, where $\log(n^{-1}\sum_{i=1} U_i) = O_p(1)$ by iid assumption and $E[U_i] = 1$, and $n^{-1}\sum_{i=1}^n X_{ki} = 0$ by prewhitening. Thus, we have $\frac{\partial \hat{\phi}_1(\beta)}{\partial \phi_k} = 0$.

Therefore, each element (k,l) of $\hat{V}_n(\beta)$ writes

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{X_{ki}X_{li}}{1+U_i} - n^{-1}\sum_{i=1}^n X_{ki} n^{-1}\sum_{j=1}^n \frac{X_{lj}}{1+U_j}, \tag{5.A.8}$$

for $l > 1$ and

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{X_{ki}}{1+U_i}, \tag{5.A.9}$$

for $l = 1$. Remark that for $k = 1, \forall l > 1$ we have $[V_n(\beta)]_{1,l} = 0$, and for $k = 1, l = 1$, we have $\left[\hat{V}_n(\beta)\right]_{1,1} = n^{-1}\sum_{i=1}^n \frac{1}{1+U_i} < 1$. Therefore, the eigenvalue associated with the constant term is strictly below 1, and proving the convergence amounts to showing that the largest eigenvalue of the $(K-1)\times(K-1)$ lower right-hand submatrix of $\hat{V}_n(\beta)$ is strictly less than unity. All elements (k,l) for $k,l > 1$ of this matrix writes

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{X_{ki}X_{li}}{1+U_i}. \tag{5.A.10}$$

because of prewhitening. We can write this in matrix form as

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'WX, \tag{5.A.11}$$

where W is a diagonal matrix with elements (i,i) acting as weights given by $\frac{1}{1+U_i} \in (0,1]$. Note that those weights become $\frac{\delta}{\delta+U_i} \in [0,1)$ for $\delta \neq 1$. We can thus write $W = W^{1/2}W^{1/2}$, and rewrite the submatrix of interest as the quadratic form

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'W^{1/2}W^{1/2}X. \tag{5.A.12}$$

Consequently, this matrix is nonnegative definite and so must have all nonnegative eigenvalues. We can alternatively write the weight matrix $W = I_n - D$, where D is also a diagonal matrix with elements $\frac{U_i}{1+U_i} \in [0,1)$, or more generally $\frac{U_i}{\delta+U_i} \in [0,1)$. Therefore, we have the alternative expression

$$\left[\hat{V}_n(\beta)\right]_{k,l>1} = n^{-1}X'(I_n - D)X = I_{K-1} - n^{-1}X'D^{1/2}D^{1/2}X, \tag{5.A.13}$$

where the second term is also a quadratic form. It follows that as $n \to \infty$, the maximum eigenvalue is equal to

$$\max_{|a|=1} a'\left[\hat{V}_n(\beta)\right]_{k,l>1} a = \max_{|a|=1} 1 - a'X'D^{1/2}D^{1/2}Xa. \tag{5.A.14}$$

Assuming the data distribution is non-degenerate, $a'X'D^{1/2}D^{1/2}Xa$ is positive and bounded away from zero for all unit vectors $a \in R^{K-1}$. Thus, as $n \to \infty$, the maximum eigenvalue of $\hat{V}_n(\beta)$ is strictly less than unity. This proves the result.

Let us turn to condition (ii). Following Dominitz and Sherman (2005), a sufficient condition to satisfy (ii) is $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, where $\kappa \in [0,1)$ is the modulus of the contraction $M(\cdot)$, which can be estimated as the mean or median of $\hat{\kappa} = |\hat{\beta}_{t+1} - \hat{\beta}_t|/|\hat{\beta}_t - \hat{\beta}_{t-1}|$ across several iterations.

For condition (iii), we want to show that $n^{1/2}\sup_{\phi\in B_0}|\hat{M}_n(\phi) - M(\phi)| = O_p(1)$ as $n \to \infty$. For any $\phi \in B_0$, recall that $\hat{M}_n(\phi) = X'X^{-1}X'\hat{\tilde{Y}}_i(\phi)$. Under the iid assumption and assuming $E[X_iX_i'] < \infty$, applying the weak law of large numbers and Slutsky's theorem yield $n^{-1}X'X^{-1} \xrightarrow{p} E[X_iX_i']^{-1}$ and $\hat{c}(\phi) \xrightarrow{p} c(\phi)$ as $n \to \infty$, and thus $n^{-1}X'\hat{\tilde{Y}}_i(\phi) \xrightarrow{p} E\left[X_i\tilde{Y}_i(\phi)\right]$ as $n \to \infty$. Therefore, $\hat{M}_n(\phi) \xrightarrow{p} M(\phi)$ as $n \to \infty$ and the Lindeberg-Levy's central limit theorem gives $|\hat{M}_n(\phi) - M(\phi)| = O_p(n^{-1/2})$ for any $\phi \in B_0$, and thus in particular

$$n^{1/2}\sup_{\phi\in B_0}|\hat{M}_n(\phi) - M(\phi)| = O_p(1). \tag{5.A.15}$$

For condition (iv), let us use the derivations obtained earlier and similar arguments than for condition (iii). We have that $\nabla_\phi\hat{c}(\phi) \xrightarrow{p} \nabla_\phi c(\phi)$ and thus $\hat{V}_n(\phi) \xrightarrow{p} V(\phi)$ as $n \to \infty$. Therefore, the condition $||\hat{V}_n(\phi) - V(\phi)|| = o_p(1)$ holds. Applying Theorem 1 in Dominitz and Sherman (2005) yields the desired convergence result.

□

***Proof of Theorem 2**: Normality.* We now make use of Theorem 4 in Dominitz and Sherman (2005) to derive the asymptotic distribution of iOLS. All conditions have been verified in the previous results except that $\sqrt{n}(\hat{M}_n(\beta) - \beta) \xrightarrow{d} Z$ as $n \to \infty$, where Z is a limit distribution. Note that we have

$$\hat{c}(\beta) = n^{-1}\sum_{i=1}^{n}\log(n^{-1}\sum_{j=1}^{n}U_j + U_i) - \log(n^{-1}\sum_{j=1}^{n}U_j) \xrightarrow{p} E[\log(1+U_i)] = c, \tag{5.A.16}$$

as $n \to \infty$, and

$$\hat{\tilde{Y}}_i(\beta) = \log(1+U_i) + X_i'\beta - \hat{c}(\beta), \tag{5.A.17}$$

so that

$$\sqrt{n}[X'X]^{-1}X'\hat{\tilde{Y}}_i(\beta) = \sqrt{n}\left(\beta + [X'X]^{-1}X'(\log(1+U) - \hat{c}(\beta))\right). \tag{5.A.18}$$

Under the iid assumption and the exogeneity condition $E[X_i\log(1+U_i)] = c$, the Lindeberg-Levy's central limit theorem yields

$$\sqrt{n}\left(\hat{M}_n(\beta) - \beta\right) \xrightarrow{d} N(0,\Sigma), \tag{5.A.19}$$

as $n \to \infty$, where Σ is the asymptotic covariance matrix. Remark that it is the asymptotic covariance of the OLS estimator of the regression of $\hat{\tilde{Y}}(\beta)$ onto X. Heteroskedasticity-robust estimators and alike apply exactly as in the standard OLS setting. However, the iOLS estimator has a slightly different asymptotic distribution. Theorem 4 of DS 2005 gives the following result

$$\sqrt{n}\left(\hat{\beta}_{i(n)} - \beta\right) \xrightarrow{d} N(0,\Omega^{-1}), \tag{5.A.20}$$

as $n \to \infty$, where $\Omega = (I_k - V(\beta))^{-1}\Sigma(I_K - V(\beta))$ and the gradient $\nabla_\phi M(\beta) = V(\beta)$ is defined as

$$V(\beta) = E[X_iX_i']^{-1}E[\frac{X_iX_i'}{1+U_i}], \tag{5.A.21}$$

of which each element is strictly below 1. Therefore sandwich-type covariance estimators are changed from the classical expression

$$\hat{\Sigma} = (\frac{1}{n}X'X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'X)^{-1} \tag{5.A.22}$$

to

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{5.A.23}$$

where W is a diagonal weighting matrix with diagonal element $\frac{1}{1+U_i}$, and $\hat{\Sigma}_0$ is an estimator of the covariance of $X_i'(\log(1+U_i)-c)$ across observations. For another $\delta \neq 1$, we would have the weights $\frac{\delta}{\delta+U_i} \in [0,1)$. In layman's terms, the "meat" of HAC-robust estimators is unchanged but the "bread" is modified. As before, the weights become $\frac{\delta}{\delta+U_i}$ when $\delta \neq 1$.

An approximate solution consists in evaluating U_i at its mean so that a simple (though biased) estimator is given by

$$\hat{\Omega} = \frac{1+\delta}{\delta} \times \hat{\Sigma}, \tag{5.A.24}$$

where $\hat{\Sigma}$ is the estimated covariance matrix (robust or not) of the OLS estimator in the last iteration. For instance, this approximation yields standard errors twice as large as those of the OLS procedure for $\delta = 1$.

□

***Proof of Theorem 3**: iOLS$_U$.* This proof is similar to that of the previous theorem, with small modifications.

Let us now consider

$$\begin{aligned}\hat{V}_n(\phi) &= [X'X]^{-1}X'\nabla_\phi\hat{\tilde{Y}}(\phi) \\ &= n^{-1}X'\nabla_\phi\hat{\tilde{Y}}(\phi),\end{aligned} \tag{5.A.25}$$

where $\nabla_\phi\hat{\tilde{Y}}_i(\phi)$ has element (i,k) defined as

$$\left[\nabla_\phi\hat{\tilde{Y}}(\phi)\right]_{i,k} = \frac{\delta\exp(X_i'\phi)X_{ki}}{Y_i+\delta\exp(X_i'\phi)} + \frac{\partial\hat{U}_i(\phi)}{\partial\phi_k}\left(\frac{1}{1+\delta} - \frac{1}{\hat{U}_i(\phi)+\delta}\right). \tag{5.A.26}$$

This expression simplifies, when evaluated at $\phi = \beta$, to

$$\left[\nabla_\beta\hat{\tilde{Y}}(\beta)\right]_{i,k} = X_{ki}\left(1 - \frac{U_i}{1+\delta}\right), \tag{5.A.27}$$

which yields

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} X_{ki}X_{li}\left(1 - \frac{U_i}{1+\delta}\right). \tag{5.A.28}$$

Following the same reasoning as in the previous theorem, a sufficient condition for convergence is that $\frac{U_i}{1+\delta}$ is between 0 and 1 for all i. Therefore, the choice of δ will affect both the speed of convergence and whether the estimator converges at all. An efficient strategy for choosing δ is to start at a relatively small value and increment it if convergence fails – which can be checked by estimating κ as explained above.

The proof of asymptotic normality is also unchanged, except that now the diagonal weighting matrix W in

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{5.A.29}$$

has element $1 - \frac{U_i}{1+\delta}$, and $\hat{\Sigma}_0$ is an estimator of the covariance of $X_i'U_i$ across observations.

□

***Proof of Theorem 3**: iOLS$_\epsilon$).* This proof follows the same lines, with small modifications to the previous one. The gradient $\nabla_\phi \hat{\tilde{Y}}_i(\phi)$ has now element (i, k) defined as

$$\left[\nabla_\phi \hat{\tilde{Y}}(\phi)\right]_{i,k} = \frac{\delta \exp(X_i'\phi)X_{ki}}{Y_i + \delta \exp(X_i'\phi)} - \frac{1}{\hat{U}_i(\phi) + \delta}\frac{\partial \hat{U}_i(\phi)}{\partial \phi_k} + \frac{1}{1+\delta}\frac{\partial(Y_i - \exp(X_i'\phi))}{\partial \phi_k}. \tag{5.A.30}$$

This expression simplifies, when evaluated at $\phi = \beta$, to

$$\left[\nabla_\beta \hat{\tilde{Y}}(\beta)\right]_{i,k} = X_{ki}\left(1 - \frac{\exp(X_i'\beta)}{1+\delta}\right), \tag{5.A.31}$$

which yields

$$\left[\hat{V}_n(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^{n} X_{ki}X_{li}\left(1 - \frac{\exp(X_i'\beta)}{1+\delta}\right). \tag{5.A.32}$$

Following the same reasoning as in the previous theorem, a sufficient condition for convergence is that $\frac{\exp(X_i'\beta)}{1+\delta}$ is between 0 and 1 for all i. We suggest using the same trial and error approach based on estimating κ.

The proof of asymptotic normality is also unchanged, except that now the diagonal weighting matrix W in

$$\tilde{\Sigma} = (\frac{1}{n}X'(I-W)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'(I-W)X)^{-1}, \tag{5.A.33}$$

has element $1 - \frac{\exp(X_i'\beta)}{1+\delta}$, and $\hat{\Sigma}_0$ is an estimator of the covariance of $X_i'\epsilon_i$ across observations.

□

***Proof of Theorem 4** : Instrumental Variables Consistency.* Recall that the parameter $\beta \in \mathbb{R}^K$ is characterized by the fixed-point equation

$$\beta^{IV} = E[\check{X}_i\check{X}_i']^{-1}E\left[\check{X}_i\tilde{Y}_i(\beta)\right], \tag{5.A.34}$$

where $\check{X} = P^Z X$, $P^Z = Z(Z'Z)^{-1}Z'$, $Z \in \mathbb{R}^M$ with $M \geq K$, $E(Z_i'X_i)$ has rank K, and $\tilde{Y}_i(\beta) = \log(Y_i + \exp(X_i'\beta)) - c(\beta)$ is the transformed dependent variable. The mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$ which characterizes the parameter is hence defined $\forall \phi \in \mathbb{R}^K$ as

$$M^{IV}(\phi) = E[\check{X}_i\check{X}_i']^{-1}E\left[\check{X}_i\tilde{Y}_i(\phi)\right]. \tag{5.A.35}$$

The sample counterpart of this mapping is given by

$$\hat{M}_n^{IV}(\phi) = [\check{X}_i'\check{X}_i]^{-1}\check{X}_i'\hat{\tilde{Y}}_i(\phi), \tag{5.A.36}$$

where $\hat{\tilde{Y}}_i(\phi)$ is defined as before.

Our proof is very similar to the one used to show Theorem 2. For condition (i), the first step is to consider that Z is standardized so that $\check{X}$ is prewhitened: $\check{X}'\check{X} = nI_k$. As before, showing condition (i) boils down to showing that the largest eigenvalue of $\nabla_\phi \hat{M}_n^{IV}(\beta) = \hat{V}_n^{IV}(\beta)$ is strictly less that unity as $n \to \infty$. Note that we have

$$\begin{aligned}\hat{V}_n^{IV}(\phi) &= [\check{X}'\check{X}]^{-1}\check{X}'\nabla_\phi \hat{\tilde{Y}}(\phi) \\ &= n^{-1}\check{X}'\nabla_\phi \hat{\tilde{Y}}(\phi),\end{aligned} \tag{5.A.37}$$

where the second equality uses prewhitening on $\check{X}$. Moreover, $\nabla_\phi \hat{\check{Y}}_i(\phi)$ has element (i,k) defined as

$$\left[\nabla_\phi \hat{\check{Y}}(\phi)\right]_{i,k} = \frac{\exp(X_i'\phi)X_{ki}}{Y_i + \exp(X_i'\phi)} - \frac{\partial \hat{c}(\phi)}{\partial \phi_k}. \tag{5.A.38}$$

Let us denote $X_{1i} = 1$ and $Z_{1i} = 1$, for all i as the constant. By prewhitening $\check{X}$, we have $\sum_{j=1}^n \check{X}_{1j} = n$ and $\sum_{j=1}^n \check{X}_{kj} = 0$ for $k > 1$. The derivative of the nuisance parameter estimate writes

$$\frac{\partial \hat{c}(\phi)}{\partial \phi_k} = n^{-1}\sum_{i=1}^n \frac{\exp(X_i^{r\prime}\phi^r + \hat{\phi}^1)(\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki})}{Y_i + \exp(X_i^{r\prime}\phi^r + \hat{\phi}^1)} - n^{-1}\sum_{i=1}^n (\frac{\partial \hat{\phi}^1}{\partial \phi_k} + X_{ki}), \tag{5.A.39}$$

for $k > 1$ and $\frac{\partial \hat{c}(\phi)}{\partial \phi_1} = 0$. As before, this expression simplifies when evaluated at $\phi = \beta$, as shown by

$$\begin{aligned}\frac{\partial \hat{c}(\beta)}{\partial \phi_k} =& n^{-1}\sum_{i=1}^n \frac{X_{ki}}{1+U_i} - n^{-1}\sum_{i=1}^n X_{ki} + O_p(1) \\ =& n^{-1}\sum_{i=1}^n \frac{X_{ki}U_i}{1+U_i} + O_p(1),\end{aligned} \tag{5.A.40}$$

for $k > 1$ because $\hat{\phi}^1(\beta) = \log(n^{-1}\sum_{i=1} Y_i \exp(-X_i^{r\prime}\beta^r)) = \beta_1 + \log(n^{-1}\sum_{i=1} U_i)$, where $\log(n^{-1}\sum_{i=1} U_i) = O_p(1)$ by iid assumption and $E[U_i] = 1$.

Therefore, each element (k,l) of $\hat{V}_n^{IV}(\beta)$ writes

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{\check{X}_{ki}X_{li}}{1+U_i} - (n^{-1}\sum_{i=1}^n \check{X}_{ki})(n^{-1}\sum_{j=1}^n \frac{X_{lj}U_j}{1+U_j}), \tag{5.A.41}$$

for $l > 1$ and

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{\check{X}_{ki}}{1+U_i}, \tag{5.A.42}$$

for $l = 1$. Remark that for $k = 1, \forall l > 1$ we have $\left[V_n^{IV}(\beta)\right]_{1,l} = n^{-1}\sum_{i=1}^n \frac{X_{li}}{1+U_i}$, and for $k = 1, l = 1$, we have $\left[\hat{V}_n^{IV}(\beta)\right]_{1,1} = n^{-1}\sum_{i=1}^n \frac{1}{1+U_i} < 1$. Therefore, all elements (k,l) for $k, l \geq 1$ of this matrix writes

$$\left[\hat{V}_n^{IV}(\beta)\right]_{k,l} = n^{-1}\sum_{i=1}^n \frac{\check{X}_{ki}X_{li}}{1+U_i}. \tag{5.A.43}$$

because of prewhitening. We can write this in matrix form as

$$\left[\hat{V}_n^{IV}(\beta^{IV})\right] = n^{-1}X'P_zWX, \tag{5.A.44}$$

where W is a diagonal matrix with elements (i,i) acting as weights given by $\frac{1}{1+U_i} \in (0,1]$. The projection matrix P_z being symmetric and idempotent, its eigenvalues are equal to either 0 or 1. P_z is hence a positive semi-definite matrix. The product P_zW is thus a positive semi-definite matrix because it is the product of two symmetric positive semi-definite matrices.

Nevertheless P_zW is not necessarily symmetric. For any vector $a \in \mathbb{R}^K$, $a'X'P_zWXa$ and $a'X'\frac{1}{2}(P_zW + W'Pz)Xa$ are the same quadratic forms. We have that $X'\frac{1}{2}(P_zW + W'Pz)X$ is positive semi-definite matrix and all its eigenvalues are nonnegative and corresponds to those of $X'P_zWX$.

We can alternatively write the weight matrix $W = I_n - D$, where D is also a diagonal matrix with elements

$\frac{U_i}{1+U_i} \in [0,1)$. Therefore, we have the alternative expression

$$\begin{aligned}\left[\hat{V}_n^{IV}(\beta)\right] &= n^{-1}X'P_z(I_n - D)X \\ &= X'P_zX - n^{-1}X'P_zDX \\ &= I_K - n^{-1}X'P_zDX,\end{aligned} \tag{5.A.45}$$

where the second equality comes from P_z being idempotent, and prewhitening. It follows that as $n \to \infty$, the maximum eigenvalue is equal to

$$\max_{|a|=1} a'\left[\hat{V}_n^{IV}(\beta)\right]a = \max_{|a|=1} 1 - a'X'\frac{1}{2}(P_zD + D'P_z)Xa. \tag{5.A.46}$$

Assuming the data distribution is non-degenerate, $a'X'\frac{1}{2}(P_zD + D'P_z)Xa$ is positive and bounded away from zero for all unit vectors $a \in R^K$. Thus, as $n \to \infty$, the maximum eigenvalue of $\hat{V}_n^{IV}(\beta)$ is strictly less than unity. This proves the result. The other conditions follow similar derivations as for Theorem 2 which complete the proof.

□

***Proof of Theorem 4**: Instrumental Variables Normality.* We now derive the asymptotic distribution of i2SLS. We must show that $\sqrt{n}(\hat{M}_n^{IV}(\beta) - \beta) \xrightarrow{d} Z$ as $n \to \infty$, where Z is a limit distribution. As before, we have

$$\hat{c}(\beta) \xrightarrow{p} E[\log(1+U_i)] = c, \tag{5.A.47}$$

as $n \to \infty$, and

$$\hat{\tilde{Y}}_i(\beta) = \log(1+U_i) + X_i'\beta - \hat{c}(\beta), \tag{5.A.48}$$

so that

$$\sqrt{n}[\breve{X}'\breve{X}]^{-1}\breve{X}'\hat{\tilde{Y}}_i(\beta) = \sqrt{n}\left(\beta + [\breve{X}'\breve{X}]^{-1}\breve{X}'(\log(1+U) - \hat{c}(\beta))\right). \tag{5.A.49}$$

Under the iid assumption and the exogeneity condition $E[\breve{X}_i(\log(1+U_i) - c)] = 0$, the Lindeberg-Levy's central limit theorem yields

$$\sqrt{n}\left(\hat{M}_n^{IV}(\beta) - \beta\right) \xrightarrow{d} N(0, \Sigma), \tag{5.A.50}$$

as $n \to \infty$, where Σ is the asymptotic covariance matrix. Remark that it is the asymptotic covariance of the 2SLS estimator of the regression of $\hat{\tilde{Y}}(\beta)$ onto X using Z as IV. Heteroskedasticity-robust estimators apply as in the standard setting. However, the i2SLS estimator has a slightly different asymptotic distribution, because the true β is unknown. Using the same reasoning as for iOLS, we obtain

$$\sqrt{n}\left(\hat{\beta}_{i(n)}^{IV} - \beta^{IV}\right) \xrightarrow{d} N(0, [\Omega^{IV}]^{-1}), \tag{5.A.51}$$

as $n \to \infty$, where $\Omega^{IV} = (I_k - V^{IV}(\beta))^{-1}\Sigma(I_K - V^{IV}(\beta))^{-1}$ and the gradient $\nabla_\phi M^{IV}(\beta) = V^{IV}(\beta)$ is defined as

$$V(\beta) = E[\breve{X}_i\breve{X}_i']^{-1}E[\frac{\breve{X}_iX_i'}{1+U_i}]. \tag{5.A.52}$$

Therefore sandwich-type covariance estimators are given by

$$\tilde{\Sigma} = (\frac{1}{n}X'\frac{1}{2}(P_z(I-W) + (I-W)P_z)X)^{-1}\hat{\Sigma}_0(\frac{1}{n}X'\frac{1}{2}(P_z(I-W) + (I-W)P_z)X)^{-1}, \tag{5.A.53}$$

where W is a diagonal weighting matrix with diagonal element $\frac{1}{1+U_i}$, and $\hat{\Sigma}_0$ is an estimator of the covariance of $P_zX'(\log(1+U_i) - c)$ across observations. Symmetrizing the weight matrix, as explained in the proof of the preceding theorem, is required to have a symmetric positive definite matrix, hence invertible.

□

5.2 Model Extensions

5.2.1 Instrumental variables

The estimation of causal relationships is central to social sciences. Yet, doing so is fraught with difficulties. Simultaneity, an omitted variable, or the presence of measurement errors could result in biased estimates. For example, if a researcher is interested in estimating the causal effect of the number of police officers on crime, one may observe that the police is more often deployed in areas where crime is high and conclude that police causes more crime.

A popular solution consists on finding an instrumental variable which affects the outcome only through the endogenous variable. Using variation in the instrument, one can recover the impact of the main variable of interest on the outcome through an estimation procedure known as *Two Stage Least Squares* (2SLS). For example, Worrall and Kovandzic (2010) relies on exogenous variation in federal funding laws to instrument the size of the police force.

Our iterated solution extends directly to this situation and consists, in turn, in running 2SLS iteratively. Let us define Z as a $n \times L$ matrix with $L \geq K$ instrumental variables so that $E[X'Z] \neq 0$. We assume $E(Z'Z) < \infty$ and denote P_z as the projection matrix $Z(Z'Z)^{-1}Z'$. The following algorithm characterizes the i2SLS estimators.

Algorithm 2 (i2SLS estimator)**.** *Let $\hat{\beta}_0$ be an initial estimate, as obtained for example with the 2SLS "popular fix" estimator $\hat{\beta}^{2SLSPF} = [X'P_zX]^{-1}X'P_z\log(Y+\Delta) \in \mathbb{R}^K$, for some $\Delta > 0$. the i2SLS estimator is obtained as follows.*

1. *Initialize t at 0;*
2. *Transform the dependent variable into $\tilde{Y}(\hat{\beta}_t)$;*
3. *Compute the 2SLS estimate $\hat{\beta}_{t+1}^{2SLS} = (X'P_zX)^{-1}(X'P_z\hat{\tilde{Y}}(\hat{\beta}_t))$, and update t to $t+1$;*
4. *Iterate steps 2 and 3 until $\hat{\beta}_t^{2SLS}$ converges.*

This iterative estimator converges under some conditions on $\tilde{Y}(\cdot)$. The same transformations studied earlier apply without further modifications. We prove the consistency of this estimator in the following theorem.

Theorem 4 (Consistency and Asymptotic Normality)**.** *Under the above assumptions, the i2SLS estimator is consistent and achieves the parametric rate of convergence $n^{-1/2}$. Formally, we have*

$$n^{1/2}|\hat{\beta}_{t(n)}^{IV} - \beta| = O_p(1) \tag{5.2.1}$$

as $n \to \infty$ for any $t(n) \geq -\frac{1}{2}\log(n)/\log(\kappa)$, where $\kappa \in [0,1)$ is the modulus of the associated contraction mapping from $\mathbb{R}^K$ to $\mathbb{R}^K$. In addition, the i2SLS estimator is asymptotically normally distributed such that

$$\sqrt{n}\left(\hat{\beta}_{t(n)}^{IV} - \beta\right) \xrightarrow{d} N(0, \Omega^{IV}), \tag{5.2.2}$$

as $n \to \infty$, where Ω^{IV}, as given in the proof, corresponds to the asymptotic covariance of the 2SLS estimator in the last iteration up to minor modifications.

This asymptotic result reveals several desirable properties of our procedure. First, the i2SLS estimators can be obtained easily using available software. Second, this iterative procedure makes non-linear instrumental

variable estimation computationally tractable even when many control variables are included. This is particularly important because current count models are hard to estimate, from a computational standpoint, when identified on the basis of an instrumental variable.[45] Finally, researchers often rely on the control function approach in non-linear models. This method requires the error in the second stage to be an additively separable function of the first-stage error and an independent error term. It also rules out settings where the endogenous variable is not continuous (Wooldridge, 2015). In contrast, 2SLS (and thus i2SLS) does not require such assumptions.

Finally, the specification tests developed for iOLS are easily adapted for situations with endogenous regressors. The main difference is that one must estimate $Pr(Y > 0|Z)$ instead of $Pr(Y > 0|X)$. Further details are provided in Appendix 5.2.8.

5.2.2 Dispensable zeros (iOLS$_S$)

In some circumstances, zeros can be dropped without consequence for identification. However, doing so comes at the cost of a loss of efficiency. The condition for zeros to be "dispensable" in the PPML framework is

$$E(U_i|X_i, U_i > 0) = c, \tag{5.2.3}$$

where c is a constant. This condition holds either when both $E(U_i|X_i)$ and $Pr(U_i > 0|X_i)$ are constant in X_i or when both vary with X_i. In the former case, whether one chooses to discard the zeros before estimation has no consequence for identification but will affect precision. In the (somewhat more realistic) latter case, dropping zeros is required for identification unless the term $Pr(U_i > 0|X_i)$ is explicitly modelled.

We now show that the iOLS estimators can accommodate this latter situation without loss of efficiency even when zeros are dispensable. Without loss of generality, we will focus on iOLS$_U$ and assume $E(U_i|X_i, U_i > 0) = c$, although similar conditions could be considered for iOLS$_\delta$. We propose to keep all observations but introduce a correction such that the conditional expectation $E(U_i|X_i, U_i > 0) = constant$ is respected. Let $\tilde{Y}_i$ denote the transformed dependent variable in (5.3.14), and take the conditional expectation on both sides to obtain

$$E[\tilde{Y}_i|X_i] = X_i'\beta + \frac{1}{1+\delta}\left(E[U_i|X_i] - 1\right). \tag{5.2.4}$$

Substituting $E(U_i|X_i) = cPr(U_i > 0|X_i)$, which holds by definition, yields

$$E[\tilde{Y}_i|X_i] = X_i'\beta + \frac{1}{1+\delta}\left(cPr(U_i > 0|X_i) - 1\right). \tag{5.2.5}$$

Let us further assume that $c = 1/Pr(U_i > 0)$, without loss of generality, and rearrange the above expression into

$$E[\tilde{Y}_i|X_i] = X_i'\beta + \frac{1}{1+\delta}\left(\frac{Pr(U_i > 0|X_i)}{Pr(U_i > 0)} - 1\right), \tag{5.2.6}$$

which is equivalent to

$$E[\tilde{Y}_i|X_i] - \frac{1}{1+\delta}\left(\frac{Pr(U_i > 0|X_i) - Pr(U_i > 0)}{Pr(U_i > 0)}\right) = X_i'\beta. \tag{5.2.7}$$

Therefore, we can define a new transformation of the dependent variable

[45] For example, to our knowledge, there are no packages in Stata which allow one to estimate instrumental variable count models, as in Mullahy (1997), with many categorical control variables.

$$\tilde{Y}_i(\beta) = \log(Y_i + \delta \exp(X_i'\beta)) - c_i(\beta), \tag{5.2.8}$$

where

$$\begin{aligned} c_i(\beta) = &\log(\delta + Y_i \exp(-X_i'\beta)) - \frac{1}{1+\delta}(Y_i \exp(-X_i'\beta) - 1) \\ &- \frac{1}{1+\delta}\left(\frac{Pr(U_i > 0|X_i) - Pr(U_i > 0)}{Pr(U_i > 0)} - 1\right), \end{aligned} \tag{5.2.9}$$

is such that the exogeneity condition holds in the linear model because the conditional expectation of the new transformed dependent variable has the correct mean.

We denote iOLS$_S$ the iOLS estimator based on this transformation. Before applying the iterative procedure, one needs to estimate a probability model to obtain predictions of $Pr(U_i > 0|X_i)$. In our practical implementation, we specify a logistic probability model to remain simple. $Pr(U_i > 0)$ is given by the average across observations.

The asymptotic properties of iOLS$_S$ depends on those of the estimator of $Pr(U_i > 0|X_i)$. Proving the consistency of iOLS$_S$ directly follows from that of iOLS$_U$, where the new added term in c_i does not depend on β but only on $Pr(U_i > 0|X_i)$. Therefore, $\sqrt{n}$-consistency is achieved if one has a $\sqrt{n}$-consistent estimator of that conditional probability. A nonparametric estimator will hence yield a slower convergence rate. Besides, this two-step estimation procedure requires one to correct the estimates' standard errors. A simple yet more computationally demanding approach is to use a bootstrap procedure.

5.2.3 Negative values

It sometimes occur that the dependent variable of interest take negative values in some instances. For example, wholesale hourly electricity prices can be negative for some observations (De Vos, 2015) or firms' profits can turn negative (Draca et al., 2011). This prevents the use of a log-transformation, or requires one to discard observations with negative values.

Our estimator extends to dependent variables taking negative values. However, one needs to specify a model for negative values. For simplicity, we consider model (5.2.1) and assume that U_i can now take both positive and negative values. The "popular fix" counterpart in this context would be to add a constant plus the minimum of Y_i in absolute terms. We consider, instead, the following model

$$\log(Y_i + \rho \exp(X_i'\beta)) = X_i'\beta + \log(\rho + U_i) \tag{5.2.10}$$

where ρ must be chosen such that $Y_i + \rho \exp(X_i'\beta) > 0 \ \forall i$. A necessary identifying restriction is then given by $E[X_i'(\log(\rho + U_i) - c)] = 0$.

This transformation means that the log function's vertical asymptote at zero is shifted leftwise towards the minimum value of Y. Therefore, this approach is fundamentally different from the IHS, which imposes a S-shape transformation around zero.

There are three possibilities to choose ρ. First, the error U_i is known to be bounded below so that $U_i \geq \underline{U}$. One can simply choose $\rho = \delta + |\underline{U}|$, where the choice of δ follows the same argument as in the non-negative setting. The rest of the procedure remains unchanged compared to iOLS$_\delta$.

Second, the error U_i is known to be bounded below, but $\underline{U}$ is unknown. It can be estimated by taking the first-order statistic $\hat{\underline{U}} = \min_i \frac{Y_i}{\exp(X_i'\beta)}$. In this case, $\hat{\rho} = \delta + |\hat{\underline{U}}|$ is estimated from the data. It is akin to the

popular fix for negative data in that it also consists in adding an order statistic to the dependent variable. Here, the convergence rate of $\hat{\underline{U}}$ is crucial to determine that of the iOLS estimator. For instance, if U_i is uniformly distributed, the first-order statistic will converge at rate n^{-1} to the true lower bound and the convergence result of the iOLS estimator will remain unaffected. Reversely, slower convergence rates will prevail if the first-order statistic converges at a rate slower than $n^{-1/2}$.

Finally, the error U_i could be unbounded. It is then unclear what is the appropriate exogeneity restriction. For instance, the first-order statistic of a Gaussian error will go to $-\infty$ at rate $\log(n)$. The parameter ρ will slowly decrease with the sample size and never converge. iOLS consistency would require the identifying restriction to depend on the sample size, which may not be meaningful in empirical applications. This case can be addressed by imposing the same restriction for all sample sizes, say $E[U_i|X_i] = 1$ for instance, and consider the approach detailed in the previous paragraph as a reasonable approximation.

5.2.4 Log-log specifications

In many econometric applications, the main parameter of interest is an elasticity of Y_i with respect to some variable X_i. Elasticities are often estimated using a log-log specification. However, it is common to have both dependent and independent variables that are equal to zero for some observations. Taking the log-transform of either of these variables is impossible. We propose to address this issue as follows.

Let us consider the following data generating process

$$Y_i = X_i^{\beta} U_i, \tag{5.2.11}$$

with $X_i > 0$ and $U_i \geq 0$. The iOLS$_\delta$ estimator directly applies using the transformation

$$\log(Y_i + \delta X_i^{\beta}) = \beta \log(X_i) + \eta_i, \tag{5.2.12}$$

under the exogeneity restriction $E[\log(X_i)\eta_i] = 0$, where $\eta_i = \log(\delta + U_i) - c$ is the mean-zero error term of the linearized model. The only difference with the log-linear setting is that the regressors are also in log form.

A potential issue arises when X_i can take zero values with positive probability. For any independent variable, let us rewrite the above restriction as

$$E[\log(X_i)\eta_i|X_i > 0]Pr(X_i > 0) + \lim_{\epsilon \to 0} E[\log(\epsilon)\eta_i|X_i = 0]Pr(X_i = 0) = 0, \tag{5.2.13}$$

which can be rewritten into

$$E[\log(X_i)\eta_i|X_i > 0]Pr(X_i > 0) + \lim_{\epsilon \to 0} \log(\epsilon) E[1_{(X_i=0)}\eta_i]Pr(X_i = 0) = 0. \tag{5.2.14}$$

A sufficient condition for this equality to hold is to have both $E[\log(X_i)\eta_i|X_i > 0] = 0$ and $E[1_{(X_i=0)}\eta_i] = 0$. The former is the standard exogeneity condition stated for non-negative values of X_i, whereas the latter means that the occurrences of zeros in X_i are exogenous to the errors. In the single covariate setting, one can simply discard observations where $X_i = 0$ and estimate the model based on the condition $E[\log(X_i)\eta_i|X_i > 0] = 0$. In the multivariate case, this approach would lead to discard possibly many observations and greatly dampen statistical power. Instead, one can make use of both restrictions and introduce an extra binary variable in the model,[46] as in

$$\log(Y_i + X_i^{\beta}) = \beta_0 1_{(X_i=0)} + \beta \tilde{X}_i + \eta_i, \tag{5.2.15}$$

[46] See also Battese (1997) for a very similar approach.

where $\tilde{X}_i = \log(X_i)$ for $X_i > 0$ and is equal to 0 otherwise, and the two parameters β_0 and β should be equal in principle. For ease of exposition, we have supposed the existence of a single explanatory variable but this strategy can be used along with an intercept and other covariates.

5.2.5 Incidental parameter problem

In non-linear panel data models, individual fixed-effects are not consistent when the cross-sectional dimension n increases to infinity while the time dimension T remains fixed. This issue is known as the incidental parameters problem. It is a well-known issue with maximum likelihood estimators, but is not a problem for linear estimators because the randomness of individual fixed-effects is "averaged out" and the parameters of interest are hence consistently estimated.

Our estimators do not suffer from this problem as soon as fixed-effects are averaged out at each iteration. A modified version of the iOLS algorithm can be used to accommodate many fixed effects by making use of the Frisch-Waugh-Lovell theorem as follows. Let us decompose the set of regressors $X = [X_0, X_1]$, where X_0 are binary variables capturing all fixed-effects and X_1 the remaining regressors (including the constant term). Define the projection matrix $P_0 = X_0(X_0'X_0)^{-1}X_0'$ and denote the aggregate fixed-effect term by $\Lambda = X_0'\beta_0$.

Algorithm 3 (iOLS estimator with many fixed effects)**.** *Let $\hat{\beta}_0$, and $\hat{\Lambda}_0$ be initial estimates. The iOLS estimator is defined as the following iterative procedure:*

1. *Initialize t at 0;*
2. *Transform the dependent variable into $\tilde{Y}_{iOLS}(\hat{\beta}_t, \hat{\Lambda}_t)$, where the term $X'\hat{\beta}_t$ is replaced by $X_1'\hat{\beta}_t + \hat{\Lambda}_t$;*
3. *Partial out the transformed dependent variable $\acute{\tilde{Y}}_{iOLS}(\hat{\beta}_t, \hat{\Lambda}_t) = (I_n - P_0)\tilde{Y}_{iOLS}(\hat{\beta}_t, \hat{\Lambda}_t)$ and the remaining regressors variable $\acute{X}_1 = (I_n - P_0)X_1$;*
4. *Compute the OLS estimate $\hat{\beta}_{t+1} = [\acute{X}_1{}'\acute{X}_1]^{-1}\acute{X}_1{}'\acute{\tilde{Y}}(\hat{\beta}_t)$, and update t to $t+1$;*
5. *Recover the fixed-effects into the aggregate term $\hat{\Lambda}_t = (\tilde{Y}(\hat{\beta}_t) - X_1'\hat{\beta}_{t+1}) - (\acute{\tilde{Y}}(\hat{\beta}_t) - \acute{X}_1{}'\hat{\beta}_{t+1})$*
6. *Iterate steps 2 to 5 until $\hat{\beta}_t$ converges.*

Note that all matrix inversions in this algorithm can be done only once. The presence of fixed-effects has hence almost no effect on the computation speed of the iterative estimator. Remark further that this approach relates to the Poisson estimator with high-dimensional fixed-effects. Indeed, Correia et al. (2019) transform the PPML estimator into an iteratively reweighted least squares problem, then make use of the Frisch-Waugh-Lovell theorem to fasten computations exactly like above. Their approach bears some similarities with our approach for $iOLS_\epsilon$ (additive poisson), except that ours involves less matrix inversions.

5.2.6 The log of a ratio

Researchers are sometimes willing to estimate equations of the form

$$\log(Y_{i1}/Y_{i2}) = X_i'\beta + \varepsilon_i, \tag{5.2.16}$$

where Y_{i1} and Y_{i2} are two outcomes of interest. It may happen that both outcomes can take zero values, hence not only the log is undefined but also the ratio. The "popular fix" estimator in this case consists in transforming the outcomes and focus on the following model

$$\log((Y_{i1} + \Delta)/(Y_{i2} + \Delta)) = X_i'\beta + \omega_i, \tag{5.2.17}$$

for some $\Delta > 0$.[47] Needless to explain why this simple fix is not satisfactory. Instead, let us consider an alternative solution where the two following equations are estimated jointly

$$\begin{aligned} \log(Y_{i1} + \Delta) &= X_i'\beta_1 + \varepsilon_{1i} \\ \log(Y_{i2} + \Delta) &= X_i'\beta_2 + \varepsilon_{2i}, \end{aligned} \tag{5.2.18}$$

by rewriting the problem as a seemingly unrelated regression problem. Here, we propose to use the popular fix as a starting point, but other methods like iOLS will apply without difficulty. The parameter β of interest corresponds to $\beta_1 - \beta_2$ and inference can be conducted using the delta-method. The advantage of this approach is that one can separately check which model is best to address the log of zero in each equation.

5.2.7 Enforcing the log-linear model's exogeneity condition

An alternative iOLS transformation would consist in letting δ vary across observations. For example, let $\delta_i = \delta(1 - \xi_i)$ where ξ_i takes a zero value when $Y_i = 0$ and is equal to 1 otherwise. Therefore, the iOLS transform becomes

$$\log(Y_i + (1 - \xi_i)\delta \exp(X_i'\beta)) = X'\beta + \log((1 - \xi_i)\delta + U_i). \tag{5.2.19}$$

Let us recall that $U_i = \exp(\varepsilon_i)\xi_i$, thus the error term is $\log((1 - \xi_i)\delta + \exp(\varepsilon_i)\xi_i)$. We now develop its conditional mean into

$$E(\log((1 - \xi_i)\delta + \exp(\varepsilon_i)\xi_i)|X) = E(\varepsilon_i|\xi_i = 1, X)P(X) + \log(\delta)(1 - P(X)) \tag{5.2.20}$$

On the other hand, the exogeneity condition imposed in the log-linear model is about

$$E(\varepsilon_i|X) = E(\varepsilon_i|\xi_i = 1, X)P(X) + E(\varepsilon_i|\xi_i = 0, X)(1 - P(X)). \tag{5.2.21}$$

Therefore, imposing the restriction $E(\log((1 - \xi_i)\delta + \exp(\varepsilon_i)\xi_i)|X) = 0$ under the assumption that $E(\varepsilon_i|X) = 0$ (log-linear) is equivalent to assuming that

$$E(\varepsilon_i|\xi_i = 0, X) = \log(\delta), \tag{5.2.22}$$

where δ can be chosen using the testing procedures presented in the paper.

More generally, $E(\varepsilon_i|X) = 0$ implies that

$$E(\varepsilon_i|\xi_i = 1, X) = -E(\varepsilon_i|\xi_i = 0, X)(1 - P(X))P(X)^{-1}. \tag{5.2.23}$$

We can hence evaluate any assumption about $E(\varepsilon_i|\xi_i = 0, X)$ by considering a function $\delta(\cdot) > 0$ and test whether the following condition holds

$$E(\varepsilon_i|\xi_i = 1, X) = -\log(\delta(X))(1 - P(X))P(X)^{-1}. \tag{5.2.24}$$

This approach can be helpful although the choice of the candidate functions for $\delta(\cdot)$ to be tested is beyond the

[47] Alternatively, in the context of a growth rate, Huber (2018) suggests using the "symmetric growth". This fix consists in using $2\frac{y_t - y_{t-1}}{y_t + y_{t+1}}$ as "a second-order approximation to the ln growth rate. This measure is bounded in the interval [2, 2]. It has become standard in the establishment-level literature because it naturally accommodates zeros in the outcome variable, for example due to zero household debt or firm exit".

scope of this paper. Numerical simulations reveal that the algorithm has similar performance than iOLS$_\delta$.

5.2.8 Testing with endogenous regressors

In this section, we explain how our tests adapt to endogenous regressors.

Testing the Poisson condition. For Poisson models, we have

$$E[U_i|Z_i] = E[U_i|Z_i, U_i > 0]Pr(U_i > 0|Z_i) = E(U_i), \tag{5.2.25}$$

since $E[U_i|Z_i, U_i = 0] = 0$. Following the same step as with exogenous regressors, the error term U_i under the null is such that

$$U_i = \lambda E[U]Pr(U_i > 0|Z_i)^{-1} + \nu_i \tag{5.2.26}$$

for $U_i > 0$ with $\lambda = 1$ and $E[\nu_i|U_i > 0, Z_i] = 0$. There are hence two differences: 1. one needs to estimate $P(U > 0|Z)$ instead of $P(U > 0|X)$, and 2. an IV estimator, like i2SLS, must be used to obtain $\hat{U}$.

Testing the i2SLS restriction. For iOLS$_\delta$, we have $E[\log(\delta + U_i)|Z_i] = c$. The null hypothesis is now

$$H_0 : E[\log(\delta + U_i)|Z_i, U_i > 0] - \log(\delta) = \frac{c - \log(\delta)}{Pr(U_i > 0|Z_i)}, \tag{5.2.27}$$

hence the differences are the same than for Poisson models.

Testing other restrictions. Testing for other restrictions introduces some new steps. Developing the associated exogeneity condition yields

$$E[\omega_i|Z_i, U_i > 0]P(Z_i) + E[\omega_i|Z_i, U_i = 0](1 - P(Z_i)) = 0 \tag{5.2.28}$$

which can be rearranged into

$$E[\omega_i|Z_i, U_i > 0] = -E[\omega_i|Z_i, U_i = 0](1 - P(Z_i))P(Z_i)^{-1}. \tag{5.2.29}$$

For the popular fix estimator, substituting the expression of ω_i on the right-hand-side gives

$$E[\omega_i|Z_i, U_i > 0] = -(\log(\Delta) - E(X'\beta|Z, U > 0))(1 - P(Z_i))P(Z_i)^{-1}, \tag{5.2.30}$$

where the new term $E(X'\beta|Z, U > 0)$ can be obtained from the first-stage estimates of the 2SLS procedure neglecting the zero values. For the IHS estimator, we have the similar form

$$E[\omega_i|Z_i, U_i > 0] = E(X'\beta|Z, U > 0)(1 - P(Z_i))P(Z_i)^{-1}. \tag{5.2.31}$$

5.3 Additional Simulations

There are six DGPs specified as follows:

- DGP 1: (A1) $E[X_i'(\log(1+U_i)-c)]=0$. This DGP is useful to illustrate iOLS$_\delta$. Let us assume that $\log(1+\varepsilon_i)$ is uniformly distributed as $U[\frac{c}{2P(X_i)},\frac{3c}{2P(X_i)}]$ with X_{1i} and X_{1i} also uniformly distributed as $U[-1,2]$. Choosing $c=0.41512$ yields the desired condition $E[X_i'(\log(1+U_i)-c)]=0$ with $E(U_i)=1$.

- DGP 2 (A2): $E[U_i|X_i]=1$. This DGP is aimed at comparing the alternative modelling approaches to PPML. We assume that $(X_{1i},X_{2i})'$ is bivariate normal with mean zero, variance $\sigma^2_{X1}=\sigma^2_{X2}=1$ and covariance $\sigma_{X1X2}=-0.3$. We further assume that ε_i is Gaussian with mean $-\log(P(X_i))-1/2$ and variance 1 so that $\exp(\varepsilon_i)$ is log-normal with conditional mean $1/P(X_i)$.

- DGP 3 (A3): $E[U_i|U_i>0,X_i]=1$. This DGP is such that discarding zeros and using PPML yields consistent estimates. $(X_{1i},X_{2i})'$ is distributed as in DGP 2, but now we assume $\exp(\varepsilon_i)\sim U[1-\min(1,\frac{|X_{1i}+X_{2i}|}{2}),1+\min(1,\frac{|X_{1i}+X_{2i}|}{2})]$. The purpose is to have a multiplicative error with mean 1 and with variance as a function of X_i, as for DGP 1 and 2 but not through $P(X)$.

- DGP 4 (A4): Heckit. This DGP is such that Heckman's model applies. Let $(X_{1i},X_{2i})'$ be distributed as in DGP 2 and 3. In addition, assume that $\xi_i=1$ if $X_i'\gamma+\nu_i>0$ and $\xi_i=0$ otherwise. We further assume $(\varepsilon_i,\nu_i)'$ to be iid joint Gaussian with variances $\sigma^2_\varepsilon=\sigma^2_\nu=3$ and covariance $\sigma^2_{\varepsilon\nu}=-2.7$.

- DGP 5 (A5): Inverse Hyperbolic Sine. This DGP is designed so that applying the IHS transform yields consistent OLS estimates. Let $(X_{1i},X_{2i})'$ be iid uniform draws in $[-0.5,0.5]$ and ϵ_i be iid uniform in $[-X_i'\beta,X_i'\beta+2X_i'\beta(1-P(X))P(X)^{-1}]$. The model is $\log(Y_i+\sqrt{Y_i^2+1})=X_i'\beta+\omega_i$, with $\omega_i=\xi_i\epsilon_i-(1-\xi_i)X_i'\beta$.

- DGP 6 (IV): $E[U_i|X_i]\neq 1$ but $E[U_i|Z_i]=1$. We finally turn to IV regression. Let us assume that $Pr(\xi_i=0|Z_i)=P(Z_i)=\frac{1}{1+\exp(\gamma_0+\gamma_1Z_{1i}+\gamma_2Z_{2i})}$, with the same parameters. The instrumental variables Z_{1i} and Z_{2i} are iid normal with mean 1 and variance $\sigma^2_{Z1}=\sigma^2_{Z2}=1$. We further assume that ε_i is Gaussian with mean $-\log(P(Z_i))-1/2$ and variance 1 so that $\exp(\varepsilon_i)$ is log-normal with conditional mean $1/P(Z_i)$. Finally the endogenous regressors are such that $X_{ik}=0.8Z_{ik}+0.2\varepsilon_i$, for $k=1,2$.

Table 5.3.1 reports the results for DGP 3, where we can drop zeros because $E[U_i|U_i>0,X_i]=1$ is the right identifying restriction. We first observe in this case that only iOLS$_S$ and PPML0 provide consistent estimates of β_1 and β_2. Second, we note that iOLS$_S$ provides standard errors which are several times smaller than those corresponding to PPML0. Third, we report that OLS without zeros and the Heckman correction provide estimates with some bias. This bias is more accentuated for the popular fix, the inverse hyperbolic sine transformation and also applies to the remaining models. These results suggest that ignoring to drop zeros when they are dispensable can lead to biased estimates.

Table 5.3.2 reports the results for DGP 4 which relies on the joint normality of the error terms in the selection and outcome equations. As expected, the Heckman model provides the right estimates. The standard errors of

the associated parameters are large nonetheless. We also observe that other models provide biased estimate, suggesting that ignoring the selection process which governs the zeros can lead to misleading conclusions.

Finally, Table 5.3.4 reports the results for DGP 5, where the regressors are endogenous and requires the use of instrumental variables to achieve identification under the assumption that $E[U_i|Z_i] = 1$. First, we observe that only i2SLS$_U$ provides consistent estimates. Second, however and as expected, i2SLS$_{\delta=100}$ provides similar results to i2SLS$_U$ with a slight bias. This bias is more accentuated in i2SLS$_{\delta=1}$ and is even more severe for the other models. In particular, the instrumental variable popular fix reports the wrong sign for β_2 further demonstrating its invalidity.

Table 5.3.1: Simulations: DGP 3 (A3)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	0.10	1.51	0.41	0.11	1.51	0.41
		(0.15)	(0.10)	(0.10)	(0.05)	(0.03)	(0.03)
	$iOLS_{\delta=100}$	−0.04	1.32	0.68	−0.04	1.31	0.68
		(0.10)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)
(A2)	$iOLS_U$	−0.04	1.31	0.69	−0.04	1.30	0.69
		(0.10)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)
	$iOLS_\varepsilon$	−0.02	1.26	0.72	−0.03	1.27	0.73
		(0.30)	(0.11)	(0.12)	(0.11)	(0.04)	(0.04)
	$PPML$	−0.02	1.26	0.72	−0.03	1.27	0.73
		(0.30)	(0.11)	(0.12)	(0.11)	(0.04)	(0.04)
(A3)	$iOLS_S$	0.06	1.00	1.00	0.06	1.00	1.00
		(0.05)	(0.03)	(0.02)	(0.02)	(0.01)	(0.01)
	OLS	0.98	0.91	0.91	0.98	0.91	0.91
		(0.05)	(0.04)	(0.04)	(0.02)	(0.01)	(0.01)
	$PPML0$	1.00	1.00	1.00	1.00	1.00	1.00
		(0.17)	(0.07)	(0.07)	(0.07)	(0.03)	(0.03)
(A4)	$Heckman$	0.97	0.91	0.91	0.98	0.91	0.91
		(1.38)	(0.26)	(0.26)	(0.41)	(0.08)	(0.08)
Others	$IHST$	0.67	0.72	−0.02	0.67	0.72	−0.02
		(0.09)	(0.06)	(0.06)	(0.03)	(0.02)	(0.02)
	PF	0.24	0.68	−0.00	0.24	0.68	−0.00
		(0.08)	(0.06)	(0.05)	(0.03)	(0.02)	(0.02)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP3, as described in Section 5.3. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard Errors are presented in between parentheses and are calculated using pairs bootstrap based on 10,000 simulations.

Table 5.3.2: Simulations: DGP 4 (A4)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	-0.98	1.86	0.14	-0.97	1.85	0.15
		(0.19)	(0.11)	(0.11)	(0.06)	(0.04)	(0.03)
	$iOLS_{\delta=100}$	-1.03	1.67	0.33	-1.02	1.66	0.34
		(0.16)	(0.09)	(0.09)	(0.05)	(0.03)	(0.03)
(A2)	$iOLS_U$	-1.03	1.66	0.35	-1.02	1.65	0.35
		(0.16)	(0.09)	(0.09)	(0.05)	(0.03)	(0.03)
	$iOLS_\varepsilon$	-0.93	1.53	0.44	-0.94	1.54	0.45
		(0.47)	(0.19)	(0.16)	(0.19)	(0.08)	(0.06)
	$PPML$	-0.93	1.53	0.44	-0.94	1.54	0.45
		(0.47)	(0.19)	(0.16)	(0.20)	(0.08)	(0.06)
(A3)	$iOLS_S$	-0.92	1.34	0.66	-0.92	1.35	0.65
		(0.13)	(0.07)	(0.07)	(0.04)	(0.02)	(0.02)
	OLS	-0.57	1.29	0.71	-0.56	1.29	0.71
		(0.12)	(0.06)	(0.07)	(0.04)	(0.02)	(0.02)
	$PPML0$	0.04	1.31	0.66	0.03	1.33	0.67
		(0.41)	(0.18)	(0.14)	(0.18)	(0.07)	(0.05)
(A4)	$Heckman$	1.02	1.00	1.00	1.00	1.00	1.00
		(2.41)	(0.45)	(0.45)	(0.73)	(0.14)	(0.14)
Others	$IHST$	0.29	0.68	0.00	0.29	0.68	0.00
		(0.07)	(0.05)	(0.04)	(0.02)	(0.02)	(0.01)
	PF	-0.08	0.63	0.00	-0.08	0.63	0.00
		(0.06)	(0.05)	(0.04)	(0.02)	(0.02)	(0.01)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP4, as described in Section 5.3. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard Errors are presented in between parentheses and are calculated using pairs bootstrap based on 10,000 simulations.

Table 5.3.3: Simulations: DGP 5 (A5)

		n=1000			n=10,000		
Cond.	Estim.	β_0	β_1	β_2	β_0	β_1	β_2
(A1)	$iOLS_{\delta=1}$	2.12	3.23	4.33	2.13	3.23	4.35
		(0.09)	(0.41)	(0.43)	(0.03)	(0.13)	(0.14)
	$iOLS_{\delta=100}$	2.06	3.34	5.42	2.07	3.34	5.45
		(0.09)	(0.31)	(0.32)	(0.03)	(0.09)	(0.10)
(A2)	$iOLS_U$	2.06	3.35	5.49	2.07	3.35	5.52
		(0.09)	(0.30)	(0.31)	(0.03)	(0.09)	(0.10)
	$iOLS_\varepsilon$	1.89	2.96	6.24	1.93	2.99	6.30
		(0.37)	(0.87)	(1.26)	(0.11)	(0.27)	(0.39)
	$PPML$	1.89	2.96	6.24	1.93	2.99	6.29
		(0.37)	(0.87)	(1.26)	(0.11)	(0.27)	(0.40)
(A3)	$iOLS_S$	2.06	3.02	5.75	2.07	3.03	5.78
		(0.09)	(0.25)	(0.27)	(0.03)	(0.08)	(0.08)
	OLS	1.78	2.08	3.81	1.78	2.08	3.81
		(0.10)	(0.34)	(0.37)	(0.03)	(0.11)	(0.12)
	$PPML0$	2.88	2.65	6.55	2.92	2.67	6.60
		(0.34)	(0.81)	(1.17)	(0.10)	(0.25)	(0.37)
(A4)	$Heckman$	-32.86	7.64	-4.78	-39.49	10.59	-4.94
		(436.74)	(18.21)	(18.86)	(21.66)	(4.00)	(4.54)
Others	$IHST$	1.00	1.00	1.00	1.00	1.00	1.00
		(0.05)	(0.19)	(0.20)	(0.02)	(0.06)	(0.06)
	PF	0.58	0.93	0.94	0.58	0.93	0.94
		(0.05)	(0.18)	(0.19)	(0.02)	(0.06)	(0.06)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP5, as described in Section 5.3. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard Errors are presented in between parentheses and are calculated using pairs bootstrap based on 10,000 simulations.

Table 5.3.4: Simulations: DGP 5 (IV-A2)

	n=1000			n=10,000		
Estim.	β_0	β_1	β_2	β_0	β_1	β_2
$i2SLS_{\delta=1}$	1.07	1.32	0.68	1.07	1.31	0.69
	(0.25)	(0.16)	(0.13)	(0.08)	(0.05)	(0.04)
$i2SLS_{\delta=100}$	0.99	1.03	0.97	0.99	1.02	0.98
	(0.22)	(0.13)	(0.11)	(0.06)	(0.04)	(0.03)
$i2SLS_U$	0.99	1.01	0.98	1.00	1.00	1.00
	(0.22)	(0.13)	(0.11)	(0.07)	(0.04)	(0.04)
$i2SLS_\epsilon$	1.42	0.99	0.95	1.33	1.02	1.01
	(0.58)	(0.27)	(0.27)	(0.32)	(0.12)	(0.15)
$PPML$	0.29	1.42	1.26	0.33	1.39	1.27
	(0.67)	(0.27)	(0.29)	(0.34)	(0.13)	(0.15)
OLS	0.83	1.10	1.66	0.91	1.05	1.63
	(0.10)	(0.06)	(0.06)	(0.09)	(0.06)	(0.04)
$Heckman$	-1.66	1.69	1.07	-1.65	1.67	1.01
	(1.73)	(0.40)	(0.40)	(0.58)	(0.14)	(0.16)
$2SLS$	1.53	0.66	1.33	1.52	0.66	1.34
	(0.11)	(0.07)	(0.06)	(0.03)	(0.02)	(0.02)
$IVPF$	-2.10	1.25	-0.47	-2.10	1.26	-0.48
	(0.23)	(0.16)	(0.13)	(0.07)	(0.05)	(0.04)

Notes: This table shows the parameter estimates and standard errors calculated on data simulated according to DGP5, as described in Section 5.3. The column "Cond." identifies the family of identifying condition on which the models in column "Estim." rely. The estimates are reported based on a sample of size $n = 1000$ or of $n = 10,000$. Standard Errors are presented in between parentheses and are calculated using pairs bootstrap based on 10,000 simulations.

Table 5.3.5: Simulations: Data-driven selection of δ (iOLS$_\delta$, kNN)

		δ						
n	DGP	0.1	0.5	1	5	10	50	100
1000	1	6%	19%	21%	21%	10%	7%	16%
	2	0%	2%	4%	9%	8%	11%	65%
10,000	1	0%	17%	62%	21%	1%	0%	0%
	2	0%	0%	0%	2%	7%	12%	79%

Notes: This table shows the relative frequency with which a given δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ was chosen on the basis of the 10,000 simulations. These simulations vary by sample size n and by DGP. Interpretation: when the sample size is n=10,000 and the data was generated using DGP1, $\delta = 1$ was selected 50% of the time.

Table 5.3.6: Simulations: Data-driven selection of δ (iOLS$_\delta$, Probit)

					δ			
n	DGP	0.1	0.5	1	5	10	50	100
1000	1	19%	17%	20%	19%	11%	6%	9%
	2	8%	7%	7%	8%	6%	5%	58%
10,000	1	0%	11%	57%	30%	1%	0%	0%
	2	0%	0%	0%	1%	3%	7%	89%

Notes: This table shows the relative frequency with which a given δ in the set $\{0.1, 0.5, 1, 5, 10, 50, 100\}$ was chosen on the basis of the 10,000 simulations. These simulations vary by sample size n and by DGP. Interpretation: when the sample size is n=10,000 and the data was generated using DGP1, $\delta = 1$ was selected 50% of the time.

Table 5.3.7: Simulations: Specification testing (Probit)

n	DGP	$t_{\delta=1}$	$t_{\delta=100}$	t_U	t_ϵ	t_{PPML}	t_{PF}	t_{IHST}	t_{HECK}	t_{PPML0}
1000	1	6%	23%	21%	63%	63%	21%	21%	83%	57%
	2	9%	6%	6%	4%	4%	11%	10%	32%	8%
	3	11%	8%	9%	7%	7%	12%	11%	5%	6%
	4	9%	7%	8%	10%	10%	29%	29%	10%	7%
	5	6%	7%	7%	19%	19%	5%	5%	0%	0%
10,000	1	13%	100%	100%	100%	100%	91%	92%	100%	100%
	2	33%	8%	7%	4%	4%	54%	49%	100%	41%
	3	51%	27%	30%	34%	34%	40%	36%	5%	5%
	4	46%	25%	28%	23%	23%	99%	99%	56%	11%
	5	6%	22%	25%	94%	94%	5%	5%	43%	7%

Notes: This table shows the relative rejection frequency of each null hypothesis for 10,000 simulations. These simulations vary by sample size (as reported by the column "n") and by Data Generating Process (as reported in the column "DGP"). Interpretation: when the sample size is n=1000 and the data was generated using DGP1, $t_{\delta=1}$ was rejected 6% of the time.

Table 5.3.8: Simulations: Model selection (kNN)

n	DGP	$(A1)$	$(A2)$	$(A3)$	$(A4)$	$(A5)$
1000	1	33%	41%	0%	0%	26%
	2	7%	62%	5%	0%	25%
	3	1%	48%	34%	0%	17%
	4	2%	65%	28%	0%	4%
	5	15%	39%	24%	0%	22%
10,000	1	62%	5%	0%	0%	32%
	2	0%	94%	0%	1%	5%
	3	0%	21%	74%	5%	0%
	4	0%	68%	19%	13%	0%
	5	22%	46%	2%	0%	30%

Notes: This table shows the selection frequency of each identifying restriction for 10,000 simulations. These simulations vary by sample size (as reported by column "n") and by Data Generating Process (as reported in column "DGP"). Interpretation: when the sample size is n=1000 and generated by DGP1, a model with moments (A1) is chosen 51% of the time.

Table 5.3.9: Simulations: Model selection (Probit)

n	DGP	$(A1)$	$(A2)$	$(A3)$	$(A4)$	$(A5)$
1000	1	54%	26%	0%	0%	19%
	2	18%	70%	5%	0%	6%
	3	11%	66%	18%	0%	5%
	4	21%	61%	17%	0%	0%
	5	5%	6%	87%	0%	2%
10,000	1	87%	0%	0%	12%	1%
	2	1%	98%	0%	1%	0%
	3	0%	37%	61%	1%	1%
	4	0%	79%	18%	3%	0%
	5	43%	14%	3%	0%	41%

Notes: This table shows the selection frequency of each identifying restriction for 10,000 simulations. These simulations vary by sample size (as reported by column "n") and by Data Generating Process (as reported in column "DGP"). Interpretation: when the sample size is n=1000 and generated by DGP1, a model with moments (A1) is chosen 51% of the time.

5.4 Data Appendix

5.4.1 American Economic Review (2016-2020)

Table 5.4.1: Solutions to the Log of Zero in the AER (2016-2020)

Log of Zero	log(Δ + Y_i)	PPML	Drop	IHS
48	23 (48%)	17 (35%)	15 (31%)	7 (15%)

Notes: This table reports the number of articles published in the American Economic Review from 2016 to 2020 where the issue of the log of zero was encountered. "Log of Zero" is the number of publications where at least one regression had to address this issue. "log(Δ + Y_i)" refers to the common fix of adding some discretionary constant to the dependent variable before taking the logarithmic transformation. "PPML" refers to Pseudo-Poisson Maximum Likelihood or Negative Binomial regression. "Drop" refers to cases where the problematic observations are discarded. "IHS" refers to the Inverse Hyperbolic Sine Transformation of the dependent variable. Some articles used several solutions, as robustness checks, which explains why the sum of solutions is different larger than 48.

Table 5.4.2: American Economic Review Cases per Year

Year	Emp. Pub.	log(Y_i)	log(Δ+Y_i)	PPML	Drop	IHS
2016	69	27	2	4	7	1
2017	71	28	5	2	4	1
2018	69	32	4	4	2	1
2019	79	27	6	6	2	3
2020	53	19	6	1	0	1

Notes: This table displays the frequency of solutions observed in American Economic Review. The sample extends over the period Jan. 2016 to Oct. 2020. *Emp. Pub.* is the number of empirical papers (includes "data" section). The column $\log(Y_i)$ counts cases where the dependent variable was in logarithmic form or in which a fix (such as $\log(\Delta + Y_i)$, PPML, Drop, or IHS) is used. It excludes cases where the author openly states that a logarithmic specification was preferred but rejected due to the existence of non-positive observations. $\log(\Delta + Y_i)$ is the popular fix. *PPML* refers to Poisson and Negative Binomial regression. *Drop* refers to cases where the author dropped the problematic observations. *IHS* is the Inverse Hyperbolic Transformation.

5.4.2 ResearchGate

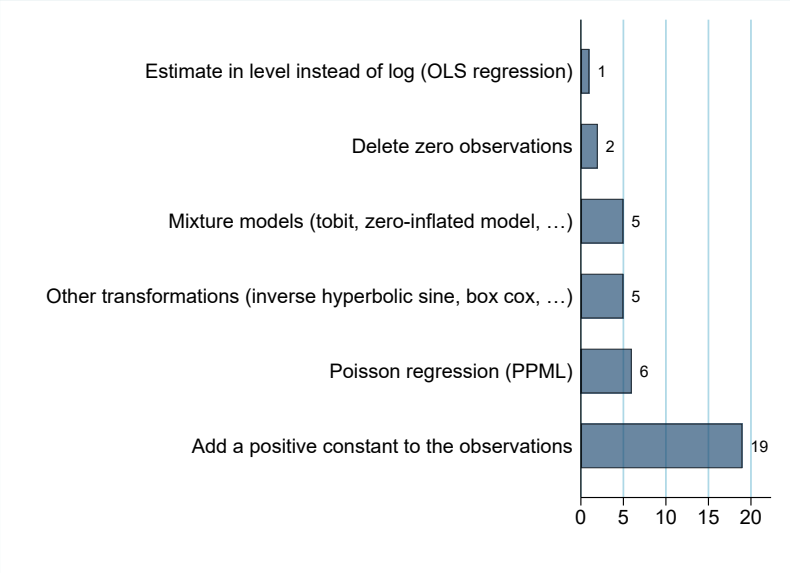


Figure 5.4.1: Proposed solutions by category on ResearchGate (November 2018)

5.4.3 Wooclap Survey

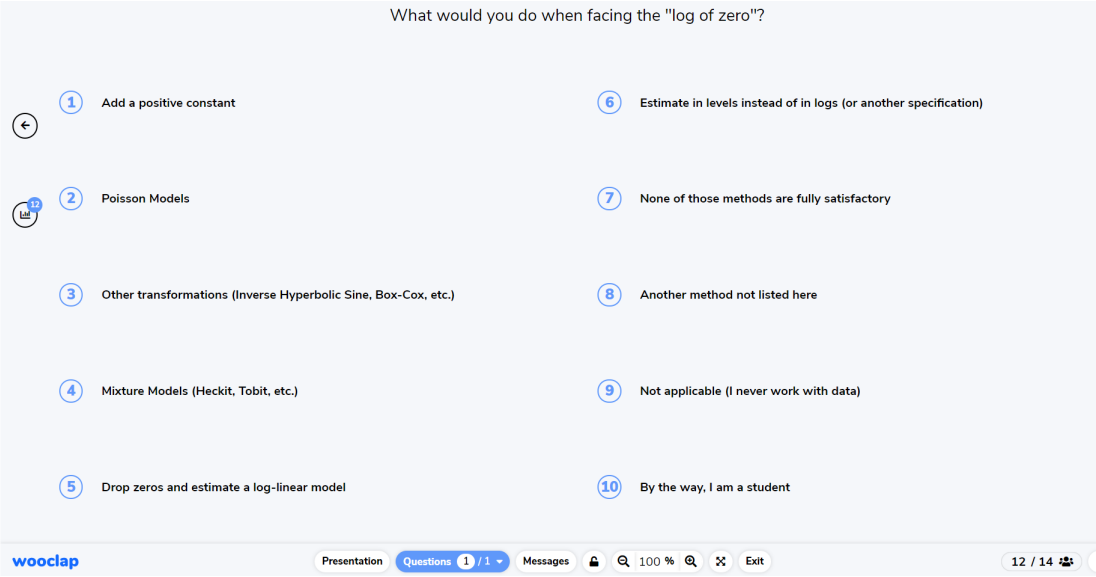


Figure 5.4.2: Wooclap Survey

Description. The survey was implemented during 3 seminars (CREST, HEC Montréal, and University of Montréal) in 2021, before the speaker clarified the different approaches. The attendees could provide multiple answers to the questions displayed in Figure 5.4.2 and were invited to indicate if they were a student. Results are presented in Table 5.4.3.

Table 5.4.3: Wooclap Survey Results

	Frequency
Popular fix	42,8 %
Poisson	17,8 %
Other transformation	17,8 %
Mixture	35,7 %
Drop zeros	17,8 %
Levels instead of logs	17,8 %
Another method	3,5 %
None satisfactory	25 %
Not applicable	3,5 %
Student	21,4 %
Nb. Respondents	28

Notes: This table displays relative frequency of answers to the Wooclap Survey. Intrepretation: 42.8% of respondents would use the popular fix (but not necessarily exclusively).

5.4.4 Santos Silva and Tenreyro (2006)

Table 5.4.4: Tests for Santos Silva and Tenreyro (2006)'s Table 3

	PF	IHST	$t_{\delta=100}$	t_{PPML}	t_{iOLS_U}	t_{HECK}	t_{PPML0}
Logit Model							
$\hat{\lambda}$	0.04	0.03	0.46	1.26	0.44	0.81	-0.22
(s.e)	(0.02)	(0.02)	(0.06)	(0.39)	(0.06)	(0.09)	(0.37)
t-Stat.	[-53.45]	[-57.90]	[-9.82]	[0.68]	[-9.24]	[8.75]	[-0.59]
KNN Model							
$\hat{\lambda}$	0.27	0.21	0.93	1.72	0.92	0.81	0.75
(s.e)	(0.01)	(0.01)	(0.05)	(0.25)	(0.05)	(0.10)	(0.37)
t-Stat.	[-49.85]	[-66.82]	[-1.55]	[2.86]	[-1.51]	[8.47]	[2.00]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for several models of trade gravity presented in Table 5.7. $iOLS_{\delta=100}$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. $\hat{\lambda}$ and t-tests are defined in Section 5.4.1.

Table 5.4.5: Logit Model for Santos Silva and Tenreyro (2006)'s Table 3

	$1(Trade > 0)$	
	Coef.	*s.e.*
Log-GDP (Exp.)	0.82	(0.02)
Log-GDP (Imp.)	0.59	(0.01)
Log-GDP per Capita (Exp.)	0.18	(0.02)
Log-GDP per Capita (Imp.)	0.20	(0.02)
Log(Distance)	-0.84	(0.04)
Contiguity	-0.84	(0.18)
Language	0.63	(0.07)
Colonial	0.26	(0.07)
LandLocked (Exp.)	0.10	(0.06)
LandLocker (Imp.)	-0.13	(0.06)
Remote (Exp.)	0.24	(0.09)
Remote (Imp.)	-0.12	(0.09)
Free Trade Agreement	2.27	(0.26)
Openness	0.49	(0.05)

Notes: This table displays the logit estimates and standard errors (s.e) based on 300 pairs bootstrap used to calculate the various t-statistics of Tables 5.7 and 5.4.4.

5.4.5 Michalopoulos and Papaioannou (2013)

Table 5.4.6: Tests for Michalopoulos and Papaioannou (2013)'s Table 2 (Logit)

	PF	iOLS$_{\delta=0.1}$	iOLS$_{\delta=0.5}$	iOLS$_{\delta=100}$	PPML	iOLS$_U$	HECK	PPML0
No Controls								
$\hat{\lambda}$	-2.95	1.00	1.00	1.01	1.04	1.00	65.16	30.07
(s.e)	(0.33)	(0.00)	(0.00)	(0.00)	(0.02)	(0.00)	(16782)	(7049)
t-Stat.	[-11.98]	[1.19]	[1.17]	[2.43]	[2.11]	[0.75]	[0.00]	[0.00]
Pop. Controls								
$\hat{\lambda}$	-0.88	1.00	1.01	1.08	2.62	1.09	19.38	11.12
(s.e)	(0.29)	(0.02)	(0.02)	(0.06)	(0.57)	(0.07)	(5.50)	(4.64)
t-Stat.	[-6.40]	[0.03]	[0.67]	[1.36]	[2.85]	[1.30]	[3.34]	[2.18]
Pop. & Loc. Controls								
$\hat{\lambda}$	-0.56	0.98	1.00	1.03	3.68	1.03	7.02	0.51
(s.e)	(0.24)	(0.05)	(0.04)	(0.05)	(1.24)	(0.05)	(1.91)	(2.45)
t-Stat.	[-6.58]	[-0.41]	[-0.04]	[0.58]	[2.16]	[0.55]	[3.15]	[-0.20]
Pop. & Loc. & Geo. Controls								
$\hat{\lambda}$	-0.18	0.80	0.82	0.85	1.94	0.85	2.84	0.80
(s.e)	(0.12)	(0.19)	(0.19)	(0.21)	(0.91)	(0.22)	(0.55)	(0.65)
t-Stat.	[-9.46]	[-1.06]	[-0.91]	[-0.69]	[1.04]	[-0.68]	[3.35]	[-0.30]
Pop. & Loc. & Geo. Controls with Rule of Law Index								
$\hat{\lambda}$	-0.10	0.60	0.62	0.62	1.96	0.62	2.56	0.47
(s.e)	(0.12)	(0.22)	(0.22)	(0.24)	(1.11)	(0.24)	(0.56)	(0.66)
t-Stat.	[-9.13]	[-1.83]	[-1.72]	[-1.57]	[0.86]	[-1.56]	[2.78]	[-0.81]
Pop. & Loc. & Geo. Controls with Log(GDP/Capita)								
$\hat{\lambda}$	-0.06	0.40	0.42	0.45	1.76	0.46	2.37	0.41
(s.e)	(0.10)	(0.26)	(0.27)	(0.28)	(1.54)	(0.28)	(0.47)	(0.73)
t-Stat.	[-10.50]	[-2.29]	[-2.17]	[-1.95]	[0.49]	[-1.93]	[2.90]	[-0.81]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night, and presented in Tables 5.8 and 5.4.8. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 5.4.1.This table relies on the logit procedure for these tests. Six specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). The last two controls for, respectively, the quality of the legal system (in 2007) and GDP per Capita (in 2007).

Table 5.4.7: Tests for Michalopoulos and Papaioannou (2013)'s Table 2 (KNN)

	PF	iOLS$_{\delta=0.1}$	iOLS$_{\delta=0.5}$	iOLS$_{\delta=100}$	PPML	iOLS$_U$	HECK	PPML0
No Controls								
$\hat{\lambda}$	-3.61	1.23	1.24	1.26	2.30	1.31	65.16	0.00
(s.e)	(1.22)	(0.20)	(0.21)	(0.30)	(0.53)	(0.32)	(190.65)	(39.73)
t-Stat.	[-3.78]	[1.16]	[1.13]	[0.86]	[2.46]	[0.96]	[0.34]	[-0.03]
Pop. Controls								
$\hat{\lambda}$	-2.14	0.91	0.89	0.81	0.02	0.79	19.38	1.52
(s.e)	(0.33)	(0.04)	(0.05)	(0.11)	(0.72)	(0.12)	(5.58)	(1.42)
t-Stat.	[-9.39]	[-2.23]	[-2.10]	[-1.73]	[-1.36]	[-1.82]	[3.29]	[0.37]
Pop. & Loc. Controls								
$\hat{\lambda}$	-2.10	0.94	0.92	0.87	1.62	0.87	7.02	-1.81
(s.e)	(0.27)	(0.03)	(0.04)	(0.07)	(0.70)	(0.07)	(2.14)	(3.01)
t-Stat.	[-11.34]	[-1.96]	[-2.02]	[-1.84]	[0.88]	[-1.85]	[2.82]	[-0.93]
Pop. & Loc. & Geo. Controls								
$\hat{\lambda}$	-1.83	1.08	1.05	1.01	1.56	1.01	2.84	-1.55
(s.e)	(0.30)	(0.04)	(0.04)	(0.06)	(0.24)	(0.06)	(0.64)	(1.57)
t-Stat.	[-9.43]	[1.91]	[1.19]	[0.12]	[2.33]	[0.13]	[2.86]	[-1.63]
Pop. & Loc. & Geo. Controls with Rule of Law Index								
$\hat{\lambda}$	-1.24	1.08	1.06	1.02	1.73	1.02	2.56	-0.46
(s.e)	(0.28)	(0.06)	(0.07)	(0.09)	(0.57)	(0.09)	(0.56)	(0.87)
t-Stat.	[-7.98]	[1.39]	[0.89]	[0.23]	[1.29]	[0.25]	[2.77]	[-1.68]
Pop. & Loc. & Geo. Controls with Log(GDP/Capita)								
$\hat{\lambda}$	-1.61	1.17	1.14	1.11	2.15	1.11	2.37	0.61
(s.e)	(0.26)	(0.08)	(0.09)	(0.13)	(0.97)	(0.13)	(0.45)	(0.70)
t-Stat.	[-9.96]	[1.97]	[1.51]	[0.83]	[1.18]	[0.84]	[3.02]	[-0.56]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night, and presented in Tables 5.8 and 5.4.8. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 5.4.1. This table relies on the KNN procedure for these tests. Six specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). The last two controls for, respectively, the quality of the legal system (in 2007) and GDP per Capita (in 2007).

Table 5.4.8: Estimates from Michalopoulos and Papaioannou (2013)'s Table 2

Coefficient estimate on *Jurisdictional Hierarchy*

	PF	iOLS$_{\delta=0.1}$	iOLS$_{\delta=0.5}$	iOLS$_{\delta=100}$	PPML	iOLS$_U$	HECK	PPML0
No Controls								
β	0.41	0.66	0.66	0.44	0.50	0.38	5.65	0.43
(s.e)	(0.07)	(0.11)	(0.10)	(0.18)	(0.21)	(0.20)	(23.06)	(0.21)
t-Stat.	[-11.98]	[1.19]	[1.17]	[2.43]	[2.11]	[0.75]	[0.00]	[0.00]
Pop. Controls								
β	0.35	0.53	0.53	0.44	0.29	0.41	1.59	0.30
(s.e)	(0.07)	(0.11)	(0.11)	(0.13)	(0.12)	(0.14)	(0.42)	(0.11)
t-Stat.	[-6.40]	[0.03]	[0.67]	[1.36]	[2.85]	[1.30]	[3.52]	[2.39]
Pop. & Loc. Controls								
β	0.32	0.42	0.40	0.36	0.14	0.35	0.72	0.12
(s.e)	(0.06)	(0.09)	(0.08)	(0.09)	(0.11)	(0.10)	(0.18)	(0.11)
t-Stat.	[-6.58]	[-0.41]	[-0.04]	[0.58]	[2.16]	[0.55]	[3.67]	[0.21]
Pop. & Loc. & Geo. Controls								
β	0.19	0.05	0.09	0.11	0.00	0.10	0.18	0.01
(s.e)	(0.05)	(0.11)	(0.10)	(0.09)	(0.10)	(0.09)	(0.09)	(0.10)
t-Stat.	[-9.46]	[-1.06]	[-0.91]	[-0.69]	[1.04]	[-0.68]	[5.16]	[1.23]
Pop. & Loc. & Geo. Controls with Rule of Law Index								
β	0.16	0.01	0.05	0.07	-0.04	0.07	0.12	-0.02
(s.e)	(0.06)	(0.12)	(0.11)	(0.10)	(0.11)	(0.10)	(0.09)	(0.11)
t-Stat.	[-9.13]	[-1.83]	[-1.72]	[-1.57]	[0.86]	[-1.56]	[4.57]	[0.71]
Pop. & Loc. & Geo. Controls with Log(GDP/Capita)								
β	0.20	0.01	0.04	0.04	-0.11	0.04	0.16	-0.09
(s.e)	(0.05)	(0.12)	(0.12)	(0.10)	(0.10)	(0.10)	(0.08)	(0.10)
t-Stat.	[-10.50]	[-2.29]	[-2.17]	[-1.95]	[0.49]	[-1.93]	[5.01]	[0.56]

Notes: This table displays the coefficient associated with jurisdictional hierarchy, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. The t-Stats rely on the logit probability model and procedure. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). Six specifications are presented, controlling cumulatively for population density (Pop.), Location (Loc.), and Geography (Geo.). The last two controls for, respectively, the quality of the legal system (in 2007) and GDP per Capita (in 2007).

Table 5.4.9: Tests for Michalopoulos and Papaioannou (2013)'s Table 3 (Logit)
(Panel A, columns (1)-(4))

	PF	$\text{iOLS}_{\delta=100}$	PPML	iOLS_U	HECK	PPML0
Country Fixed Effects Only						
$\hat{\lambda}$	-1.50	1.00	1.26	1.00	5.56	6.47
(s.e)	(0.22)	(0.01)	(0.14)	(0.01)	(3.47)	(4.29)
t-Stat.	[-11.42]	[0.31]	[1.86]	[0.27]	[1.31]	[1.27]
with Loc. & Geo. Controls						
$\hat{\lambda}$	-0.11	0.75	1.39	0.76	1.32	1.12
(s.e)	(0.16)	(0.22)	(1.02)	(0.22)	(0.62)	(0.66)
t-Stat.	[-6.86]	[-1.12]	[0.39]	[-1.08]	[0.52]	[0.18]
with Pop. Controls						
$\hat{\lambda}$	-0.05	0.55	4.60	0.55	0.65	-7.17
(s.e)	(0.21)	(0.26)	(1.78)	(0.26)	(1.41)	(3.26)
t-Stat.	[-5.07]	[-1.74]	[2.03]	[-1.72]	[-0.25]	[-2.51]
with Pop. & Loc. & Geo. Controls						
$\hat{\lambda}$	-0.01	0.16	0.22	0.16	0.96	-0.86
(s.e)	(0.14)	(0.36)	(1.75)	(0.36)	(0.52)	(0.95)
t-Stat.	[-7.10]	[-2.32]	[-0.44]	[-2.31]	[-0.08]	[-1.97]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 5.4.1. This table relies on the Logit procedure for these tests. Four specifications are presented, controlling for different combinations of population density (Pop.), Location (Loc.), and Geography (Geo.) along with country fixed effects.

Table 5.4.10: Tests for Michalopoulos and Papaioannou (2013)'s Table 3 (KNN) (Panel A, columns (1)-(4))

	PF	iOLS$_{\delta=100}$	PPML	iOLS$_U$	HECK	PPML0
Country Fixed Effects Only						
$\hat{\lambda}$	-2.66	1.02	1.14	1.02	5.56	-5.85
(s.e)	(0.59)	(0.07)	(0.17)	(0.07)	(3.91)	(4.93)
t-Stat.	[-6.25]	[0.30]	[0.81]	[0.29]	[1.17]	[-1.39]
with Loc. & Geo. Controls						
$\hat{\lambda}$	-1.70	0.95	0.99	0.95	1.32	-0.38
(s.e)	(0.40)	(0.10)	(0.43)	(0.10)	(0.63)	(1.49)
t-Stat.	[-6.80]	[-0.52]	[-0.01]	[-0.53]	[0.51]	[-0.93]
with Pop. Controls						
$\hat{\lambda}$	-2.07	0.95	1.70	0.95	0.65	0.58
(s.e)	(0.42)	(0.06)	(0.72)	(0.06)	(1.29)	(2.57)
t-Stat.	[-7.25]	[-0.86]	[0.97]	[-0.85]	[-0.27]	[-0.16]
with Pop. & Loc. & Geo. Controls						
$\hat{\lambda}$	-1.81	1.03	1.68	1.03	0.96	-0.70
(s.e)	(0.34)	(0.07)	(0.53)	(0.07)	(0.52)	(0.95)
t-Stat.	[-8.33]	[0.46]	[1.28]	[0.45]	[-0.08]	[-1.80]

Notes: This table displays the $\hat{\lambda}$-parameter, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). $\hat{\lambda}$ and t-tests are defined in Section 5.4.1. This table relies on the KNN procedure for these tests. Four specifications are presented, controlling for different combinations of population density (Pop.), Location (Loc.), and Geography (Geo.) along with country fixed effects.

Table 5.4.11: Estimates from Michalopoulos and Papaioannou (2013)'s Table 3 (Panel A, columns (1)-(4))

Coefficient estimate on *Jurisdictional Hierarchy*

	PF	iOLS$_{\delta=100}$	PPML	iOLS$_U$	HECK	PPML0
Country Fixed Effects Only						
β	0.33	0.38	0.34	0.38	0.71	0.29
(s.e)	(0.07)	(0.09)	(0.19)	(0.09)	(3.47)	(0.20)
t-Stat.	[-11.42]	[0.31]	[1.86]	[0.27]	[1.60]	[1.51]
with Loc. & Geo. Controls						
β	0.28	0.43	0.03	0.44	0.38	0.02
(s.e)	(0.07)	(0.12)	(0.18)	(0.13)	(0.62)	(0.17)
t-Stat.	[-6.86]	[-1.12]	[0.39]	[-1.08]	[2.14]	[1.70]
with Pop. Controls						
β	0.21	0.25	-0.02	0.25	0.19	-0.03
(s.e)	(0.05)	(0.06)	(0.11)	(0.06)	(1.41)	(0.11)
t-Stat.	[-5.07]	[-1.74]	[2.03]	[-1.72]	[0.46]	[-2.20]
with Pop. & Loc. & Geo. Controls						
β	0.18	0.15	-0.11	0.15	0.14	-0.09
(s.e)	(0.04)	(0.08)	(0.09)	(0.08)	(0.52)	(0.10)
t-Stat.	[-7.10]	[-2.32]	[-0.44]	[-2.31]	[1.86]	[-0.91]

Notes: This table displays the coefficient associated with jurisdictional hierarchy, standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of economic activity in African regions, proxied by light intensity at night. The t-Stats rely on the logit probability model and procedure. $iOLS_\delta$, $iOLS_U$, and PPML0 are defined in Section 5.3 and 5.4.1. PF is the baseline relying on the popular fix ($\Delta = 0.01$). Four specifications are presented, controlling for different combinations of population density (Pop.), Location (Loc.), and Geography (Geo.) along with country fixed effects.

Table 5.4.12: Logit Estimates for Michalopoulos and Papaioannou (2013)'s Table 2

	$1(Light > 0)$					
	(1)	(2)	(3)	(4)	(5)	(6)
Jurisdictional Hierarchy	0.85	0.68	0.68	0.62	0.55	0.41
	(0.16)	(0.16)	(0.20)	(0.23)	(0.23)	(0.26)
Population Density	No	Yes	Yes	Yes	Yes	Yes
Location Controls	No	No	Yes	Yes	Yes	Yes
Geographic Controls	No	No	No	Yes	Yes	Yes
Rule of Law Controls	No	No	No	No	Yes	No
Log(GDP per capita (2007))	No	No	No	No	No	Yes

Notes: This table displays the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 5.8, 5.4.6, and 5.4.8.

Table 5.4.13: Logit Estimates for Michalopoulos and Papaioannou (2013)'s Table 3 (Panel A, columns (1)-(4))

	$1(Light > 0)$			
	(1)	(2)	(3)	(4)
Jurisdictional Hierarchy	0.85	0.74	0.68	0.62
	(0.16)	(0.20)	(0.16)	(0.23)
Country Fixed Effects	Yes	Yes	Yes	Yes
Population Density	No	No	Yes	Yes
Geographic Controls	No	Yes	No	Yes
Location Controls	No	Yes	No	Yes

Notes: This table displays the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 5.4.10 and 5.4.11.

5.4.6 Card and DellaVigna (2020)

Table 5.4.14: Tests for Card and DellaVigna (2020) (KNN)

	IHS	$\text{iOLS}_{\delta=50}$	PPML	iOLS_{U}
No correction for Endogeneity				
$\hat{\lambda}$	0.64	0.97	0.89	0.97
(s.e)	(0.01)	(0.01)	(0.02)	(0.01)
t-Stat.	[-27.35]	[-4.41]	[-4.50]	[-5.10]
Control Function				
$\hat{\lambda}$	0.60	0.96	0.87	0.95
(s.e)	(0.01)	(0.01)	(0.03)	(0.01)
t-Stat.	[-28.58]	[-6.12]	[-4.22]	[-6.61]
Instrumental Variables				
$\hat{\lambda}$	0.57	0.96	0.91	0.96
(s.e)	(0.02)	(0.08)	(0.04)	(0.01)
t-Stat.	[-25.79]	[-0.48]	[-2.09]	[-3.20]

Notes: This table displays the coefficient associated with an invitation to revise & resubmit (R&R), standard errors (s.e) using 300 pairs bootstrap, and t-statistics (t-Stat.) for various models of citations based on the KNN procedure. $iOLS_{\delta}$, $iOLS_{U}$, and PPML0 are defined in Section 5.3 and 5.4.1. Three specifications are presented: no correction for endogeneity (OLS) contrasts with control function (CF) and instrumental variables (IV) which rely on the Editor leave-out mean R&R rate for identification.

Table 5.4.15: First-Stage Estimates Based on Card and DellaVigna (2020)

	Revise & Resubmit	
	Coef.	*s.e*
Editor leave-out-mean R&R rate	0.38	(0.08)
Fractions of referee recommendations		
Reject	-0.00	(0.01)
No Recommendation	0.21	(0.02)
Weak R&R	0.29	(0.01)
R&R	0.69	(0.02)
Strong R&R	0.96	(0.03)
Accept	0.91	(0.03)
Author Publications in 35 high-impact journal		
One Publication	-0.00	(0.01)
Two Publications	0.01	(0.01)
Three Publications	0.02	(0.01)
Four or Five Publications	0.03	(0.01)
Six or More Publications	0.05	(0.01)
Number of authors		
Two Author	-0.01	(0.01)
Three Authors	-0.00	(0.01)
Four Authors	0.01	(0.01)

Notes: This table provides the first stage estimates used for the instrumental variable estimates provided in Table 5.9 and based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable before transformation is a dummy equal to one if the authors were invited to resubmit. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. The editor leave-out-mean R&R rate is the main variable of interest and is considered as an exogenous instrument, measuring the proclivity with which an editor invites other authors to revise and resubmit their research. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.

Table 5.4.16: Logit Estimates based on Card and DellaVigna (2020) (OLS)

	$1(Citations > 0)$	
	Coef.	*s.e*
Revise & Resubmit	0.40	(0.10)
Fractions of referee recommendations		
Reject	0.76	(0.08)
No Recommendation	0.77	(0.16)
Weak R&R	1.44	(0.14)
R&R	1.86	(0.16)
Strong R&R	2.00	(0.25)
Accept	2.24	(0.28)
Author Publications in 35 high-impact journal		
One Publication	0.30	(0.06)
Two Publications	0.56	(0.08)
Three Publications	0.75	(0.08)
Four or Five Publications	1.00	(0.09)
Six or More Publications	0.88	(0.09)
Number of Authors		
Two Authors	0.27	(0.05)
Three Authors	0.36	(0.07)
Four Authors	0.56	(0.13)

Notes: This table provides the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 5.9 (in the OLS case), based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable is a dummy equal to one if the authors obtained at least one citation. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.

Table 5.4.17: Logit Estimates Based on Card and DellaVigna (2020) (Control Function)

	$1(Citations > 0)$	
	Coef.	*s.e*
Revise & Resubmit	-0.15	(0.26)
Control Function	0.35	(0.16)
Fractions of referee recommendations		
Reject	0.76	(0.08)
No Recommendation	0.88	(0.16)
Weak R&R	1.57	(0.15)
R&R	2.18	(0.22)
Strong R&R	2.48	(0.33)
Accept	2.69	(0.34)
Author Publications in 35 high-impact journal		
One Publication	0.31	(0.06)
Two Publications	0.56	(0.08)
Three Publications	0.76	(0.09)
Four or Five Publications	1.01	(0.09)
Six or More Publications	0.90	(0.09)
Number of Authors		
Two Authors	0.27	(0.05)
Three Authors	0.36	(0.07)
Four Authors	0.56	(0.13)

Notes: This table provides the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 5.9 (in the Control Function (CF) case), based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable is a dummy equal to one if the authors obtained at least one citation. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. The editor leave-out-mean R&R rate is used to form a control function for the invitation to revise and resubmit the manuscript. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.

Table 5.4.18: Logit Estimates Based on Card and DellaVigna (2020) (Instrumental Variable)

	$1(Citations > 0)$	
	Coef.	*s.e*
Editor leave-out mean R&R rate	-1.06	(0.71)
Fractions of referee recommendations		
Reject	0.76	(0.08)
No Recommendation	0.86	(0.16)
Weak R&R	1.52	(0.14)
R&R	2.08	(0.15)
Strong R&R	2.33	(0.23)
Accept	2.54	(0.27)
Author Publications in 35 high-impact journal		
One Publication	0.31	(0.06)
Two Publications	0.56	(0.08)
Three Publications	0.76	(0.08)
Four or Five Publications	1.01	(0.09)
Six or More Publications	0.90	(0.09)
Number of authors		
Two Author	0.27	(0.05)
Three Authors	0.36	(0.07)
Four Authors	0.56	(0.13)

Notes: This table provides the logit estimates and standard errors (s.e) used to calculate the various t-statistics of Tables 5.9 (in the Instrumental Variable (IV) case), based on the research of Card and DellaVigna (2020). Each observation of the data is at the submission level. The dependent variable is a dummy equal to one if the authors obtained at least one citation. Each row of the table reports this estimate for a different control variable. This specification includes year fixed effects and publication field fixed effects. Standard errors are provided in between parenthesis and were calculated on the basis of 300 pairs bootstraps. The editor leave-out-mean R&R rate is the main variable of interest and is considered as an exogenous instrument, measuring the proclivity with which an editor invites other authors to revise and resubmit their research. Variables ending with *Fract.* measure the fraction of referee reports which were, respectively, negative, neutral, weakly positive, very positive and pushing for acceptance of the article. Variables ending in *Pub.* refer to the number of publications published in the top 35 journals by the submitting authors. Variables ending with *Authors* refer to the number of authors submitting their article for publication to the journal.

Bibliography

Abel, W., S. Tenreyro, and G. Thwaites (2018, October). Monopsony in the UK. Technical Report 1827, Centre for Macroeconomics (CFM).

Abowd, J. M., F. Kramarz, and D. N. Margolis (1999). High wage workers and high wage firms. *Econometrica 67*(2), 251–333.

Adams, B. and K. R. Williams (2019). Zone pricing in retail oligopoly. *American Economic Journal: Microeconomics 11*(1), 124–56.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician 46*(3), 175–185.

Anderson, J. E. and E. van Wincoop (2003, March). Gravity with gravitas: A solution to the border puzzle. *American Economic Review 93*(1), 170–192.

Arnold, D. (2019, October). Mergers and Acquisitions, Local Labor Market Concentration, and Worker Outcomes. SSRN Scholarly Paper ID 3476369, Social Science Research Network, Rochester, NY.

Aryal, G. and M. F. Gabrielli (2020). An empirical analysis of competitive nonlinear pricing. *International Journal of Industrial Organization 68*, 102538.

Ashenfelter, O. and D. Hyslop (2001). Measuring the effect of arbitration on wage levels: The case of police officers. *Industrial and Labor Relations Review 54*(2), 316–328.

Assad, S., R. Clark, D. Ershov, and L. Xu (2020). Algorithmic pricing and competition: Empirical evidence from the german retail gasoline market.

Azar, J., S. Berry, and I. E. Marinescu (2019, Sep). Estimating labor market power. *Social Science Research Network (SSRN)*.

Azar, J., I. Marinescu, and M. Steinbaum (2019, May). Measuring Labor Market Power Two Ways. *AEA Papers and Proceedings 109*, 317–321.

Azar, J., I. Marinescu, and M. I. Steinbaum (2017a). Labor Market Concentration. Working Paper 24147, National Bureau of Economic Research.

Azar, J., I. Marinescu, and M. I. Steinbaum (2017b, December). Labor market concentration. Working Paper 24147, National Bureau of Economic Research.

Azar, J. A., I. Marinescu, M. I. Steinbaum, and B. Taska (2018, March). Concentration in US Labor Markets: Evidence From Online Vacancy Data. Working Paper 24395, National Bureau of Economic Research.

Balasubramanian, N., J. W. Chang, M. Sakakibara, J. Sivadasan, and E. Starr (2017, Jan). Locked in? the enforceability of covenants not to compete and the careers of high-tech workers. *Social Science Research Network (SSRN)*.

Banner, S. (2015). *The baseball trust: a history of baseballs antitrust exemption*. Oxford University Press.

Barth, E., A. Bryson, and H. Dale-Olsen (2017, October). Union Density, Productivity, and Wages. DoQSS Working Papers 17-11, Department of Quantitative Social Science - UCL Institute of Education, University College London.

Bassanini, A., C. Batut, and E. Caroli (2020, July). Labor Market Concentration and Stayers' Wages: Evidence from France. SSRN Scholarly Paper ID 3506243, Social Science Research Network, Rochester, NY.

Battese, G. E. (1997). A note on the estimation of cobb-douglas production functions when some explanatory variables have zero values. *Journal of Agricultural Economics 48*(1-3), 250–252.

Bellemare, M. F. and C. J. Wichman (2020). Elasticities and the inverse hyperbolic sine transformation. *Oxford Bulletin of Economics and Statistics 82*(1), 50–61.

Bellégo, C., D. Benatia, and L. D. Pape (2021, Sep). Dealing with the log of zero in regression models. *Social Science Research Network (SSRN)*.

Benmelech, E., N. K. Bergman, and H. Kim (2018). Strong Employers and Weak Employees: How Does Employer Concentration Affect Wages? pp. 44.

Berri, D. J. and A. C. Krautmann (2019). How much did baseball's antitrust exemption cost bob gibson? *The Antitrust Bulletin 64*(4), 566–583.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, 841–890.

Bhat, C. R. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. *Transportation Research Part B: Methodological 37*(9), 837–855.

Binder, J. J. and M. Findlay (2012). The effects of the bosman ruling on national and club teams in europe. *Journal of Sports Economics 13*(2), 107–129.

Blackburn, M. (2007, 02). Estimating wage differentials without logarithms. *Labour Economics 14*, 73–98.

Blundell, R., D. Kristensen, and R. L. Matzkin (2013). Control functions and simultaneous equations methods. *American Economic Review 103*(3), 563–69.

Blundell, R. and J. M. Robin (1999, May-June). Estimation in Large and Disaggregated Demand Systems: An Estimator for Conditionally Linear Systems. *Journal of Applied Econometrics 14*(3), 209–232.

Blundell, R. W. and J. L. Powell (2004). Endogeneity in semiparametric binary response models. *The Review of Economic Studies 71*(3), 655–679.

Boal, W. M. and M. R. Ransom (1997). Monopsony in the Labor Market. *Journal of Economic Literature 35*(1), 86–112.

Bradbury, J. C. (2013). What is right with scully estimates of a player's marginal revenue product. *Journal of Sports Economics 14*(1), 87–96.

Brams, S. J. and S. Merrill (1983). Equilibrium strategies for final-offer arbitration: There is no median convergence. *Management Science 29*(8), 927–941.

Bresnahan, T. F., J. P. Davis, and P.-L. Yin (2015). *Economic Value Creation in Mobile Applications*. University of Chicago Press.

Brown, C. and J. L. Medoff (1987, June). The impact of firm acquisitions on labor. Working Paper 2273, National Bureau of Economic Research.

Brown, Z. Y. and A. MacKay (2021). Competition in pricing algorithms. Technical report, National Bureau of Economic Research.

Burbidge, J. B., L. Magee, and A. L. Robb (1988). Alternative transformations to handle extreme values of the dependent variable. *Journal of the American Statistical Association 83*(401), 123–127.

Calonico, S. (2017). rdrobust: Software for regression-discontinuity designs. *Stata Journal 17*(2), 372–404(33).

Calonico, S., M. D. Cattaneo, and M. H. Farrell (2019, 11). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*. utz022.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica 82*(6), 2295–2326.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2015). Optimal data-driven regression discontinuity plots. *Journal of the American Statistical Association 110*(512), 1753–1769.

Calvano, E., G. Calzolari, V. Denicolo, and S. Pastorello (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review 110*(10), 3267–97.

Cameron, A. and P. Trivedi (2005). *Microeconometrics*. Cambridge University Press.

Carare, O. (2012). The impact of bestseller rank on demand: Evidence from the app market. *International Economic Review 53*(3), 717–742.

Card, D. and S. DellaVigna (2020). What do editors maximize? evidence from four economics journals. *The Review of Economics and Statistics 102*(1), 195–217.

Carroll, R. J. and D. Ruppert (1984). Power transformations when fitting theoretical models to data. *Journal of the American Statistical Association 79*(386), 321–328.

Cattaneo, M. D., M. Jansson, and X. Ma (2019). Simple local polynomial density estimators. *Journal of the American Statistical Association 0*(0), 1–7.

Cestone, A., C. Fumagalli, F. Kramarz, and G. Pica (2017). Insurance between firms: The role of internal labor market. *CREST Working Papers*.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of econometrics 34*(3), 305–334.

Cho, S. and J. Rust (2010). The flat rental puzzle. *The Review of Economic Studies 77*(2), 560–594.

Chu, C. S., L. Phillip, and A. Sorensen (2011). Bundle-size pricing as an approximation to mixed bundling. *American Economic Review 101*, 263–303.

Ciani, E. and P. Fisher (2018). Dif-in-dif estimators of multiplicative treatment effects. *Journal of Econometric Methods 8*(1), 20160011.

Cohen, P., R. Hahn, J. Hall, S. Levitt, and R. Metcalfe (2016). Using big data to estimate consumer surplus: The case of uber. Technical report, National Bureau of Economic Research.

Conyon, M., S. Girma, S. Thompson, and P. Wright (2001). Do hostile mergers destroy jobs? *Journal of Economic Behavior Organization 45*(4), 427–440.

Correia, S., P. Guimarães, and T. Zylkin (2019). ppmlhdfe: Fast poisson estimation with high-dimensional fixed effects. *arXiv e-prints*.

Council of Economic Advisors (2015). Big data and differential pricing – executive office of the president of the united states, council of economic advisors. Technical report, Executive Office of the President of the United States, Washington.

Council of Economic Advisors (2016, oct). Labor market monopsony: Trends, consequences, and policy responses. *Obama White House Archives*.

Crawford, G. S. and M. Shum (2007). Monopoly quality degradation in the cable television industry. *The Journal of Law and Economics 50*(1), 181–219.

Davidson, R. and J. G. MacKinnon (1981). Several tests for model specification in the presence of alternative hypotheses. *Econometrica: Journal of the Econometric Society*, 781–793.

De Vos, K. (2015). Negative wholesale electricity prices in the german, french and belgian day-ahead, intra-day and real-time markets. *The Electricity Journal 28*(4), 36–50.

Debeauvais, T. and C. V. Lopes (2015). Gate Me If You Can: The Impact of Gating Mechanics on Retention and Revenues in Jelly Splash. *in Proc Int. Foundations of Digital Games Conf. (J. P. Zagal, E. MacCallumStewart, and J. Togelius, eds.), Society for the Advancement of the Science of Digital Games*.

DellaVigna, S. and M. Gentzkow (2019). Uniform pricing in us retail chains. *Quarterly Journal of Economics 134*(4), 2011–2084.

Dembinski, H., M. Schmelling, and R. Waldi (2019, Oct). Application of the iterated weighted least-squares fit to counting experiments. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 940*, 135–141.

Dodini, S., M. Lovenheim, K. G. Salvanes, and A. Willén (2020, October). Monopsony, Skills, and Labor Market Concentration. SSRN Scholarly Paper ID 3723636, Social Science Research Network, Rochester, NY.

Dominitz, J. and R. P. Sherman (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory 21*(4), 838–863.

Dong, Y. (2019). Regression discontinuity designs with sample selection. *Journal of Business & Economic Statistics 37*(1), 171–186.

Draca, M., S. Machin, and J. Van Reenen (2011, January). Minimum wages and firm profitability. *American Economic Journal: Applied Economics 3*(1), 129–51.

Dube, A., J. Jacobs, S. Naidu, and S. Suri (2020, March). Monopsony in online labor markets. *American Economic Review: Insights 2*(1), 33–46.

Dube, J.-P. and S. Misra (2019). Personalized pricing and customer welfare. Working paper.

Dubois, P., R. Griffith, and M. O'Connell (2017). The Effects of Banning Advertising in Junk Food Markets. *The Review of Economic Studies 85*(1), 396–436.

Eaton, J. and A. Tamura (1994). Bilateralism and regionalism in japanese and u.s. trade and direct foreign investment patterns. *Journal of the Japanese and International Economies 8*(4), 478 – 510.

Einav, L. and J. Levin (2010). Empirical industrial organization: A progress report. *Journal of Economic Perspectives 24*(2), 145–62.

Ershov, D. (2018). Competing with superstars in the mobile app market. *Available at SSRN 3265662*.

Farber, H. S. (1980). An analysis of final-offer arbitration. *The Journal of Conflict Resolution 24*(4), 683–705.

Fioretti, M. (2020). Caring or pretending to care? social impact, firms' objectives and welfare. Technical report, Tech. rep., Mimeo, Sciences Po.

Frandsen, B. R., M. Frölich, and B. Melly (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics 168*(2), 382 – 395.

Fudenberg, D. and M. J. Villas-Boas (2006). Behavior based price discrimination and customer recognition. In T. J. Hendershott (Ed.), *Handbook on Economics and Information Systems*, pp. 377–436.

Fudenberg, D. and M. J. Villas-Boas (2012). Price discrimination in the digital economy. In M. Peitz and J. Waldfogel (Eds.), *The Oxford Handbook of the Digital Economy*.

Gandhi, A. and A. Nevo (2021). Empirical models of demand and supply in differentiated products industries. Technical report, National Bureau of Economic Research.

Garbarino, E. and S. Maxwell (2010). Consumer response to norm-breaking pricing events in e-commerce. *Journal of Business Research 63*(9), 1066 – 1072. Advances in Internet Consumer Behavior & Marketing Strategy.

Ghose, A. and S. P. Han (2014a). Estimating demand for mobile applications in the new economy. *Management Science 60*(6), 1470–1488.

Ghose, A. and S. P. Han (2014b). Estimating demand for mobile applications in the new economy. *Management Science 60*(6), 1470–1488.

Gokhale, J., E. L. Groshen, and D. Neumark (1995). Do hostile takeovers reduce extramarginal wage payments? *The Review of Economics and Statistics 77*(3), 470–485.

Goldberger, A. S. (1968). The interpretation and estimation of cobb-douglas functions. *Econometrica 36*(3/4), 464–472.

Goldfarb, A. and C. Tucker (2019, March). Digital Economics. *Journal of Economic Literature 57*(1), 3–43.

Gourieroux, C., A. Monfort, and A. Trognon (1984). Pseudo maximum likelihood methods: Applications to poisson models. *Econometrica 52*(3), 701–20.

Gugler, K. and B. Yurtoglu (2004). The effects of mergers on company employment in the usa and europe. *International Journal of Industrial Organization 22*(4), 481 – 502.

Hakes, J. K. and R. D. Sauer (2006). An economic evaluation of the moneyball hypothesis. *The Journal of Economic Perspectives 20*(3), 173–186.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *Prototype Methods and Nearest-Neighbors*, pp. 459–483. New York, NY: Springer New York.

Head, K. and T. Mayer (2014). Gravity Equations: Workhorse,Toolkit, and Cookbook. In G. Gopinath, . Helpman, and K. Rogoff (Eds.), *Handbook of International Economics*, Volume 4 of *Handbook of International Economics*, Chapter 0, pp. 131–195. Elsevier.

Head, K. and T. Mayer (2019). Brands in motion, how frictions shape multinational production. *American Economic Review forthcoming*.

Heckman, J. J. (1979a). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Heckman, J. J. (1979b). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Helpman, E., M. Melitz, and Y. Rubinstein (2008, 05). Estimating Trade Flows: Trading Partners and Trading Volumes*. *The Quarterly Journal of Economics 123*(2), 441–487.

Hershbein, B., C. Macaluso, and C. Yeh (2018). Concentration in U.S. local labor markets: evidence from vacancy and employment data. pp. 31.

Hortaçsu, A., O. R. Natan, H. Parsley, T. Schwieg, and K. R. Williams (2021). Organizational structure and pricing: Evidence from a large us airline.

Huang, Y., P. B. Ellickson, and M. J. Lovett (2020). Learning to set prices. *Available at SSRN 3267701*.

Huber, K. (2018, March). Disentangling the effects of a banking crisis: Evidence from german firms and counties. *American Economic Review 108*(3), 868–98.

Hylton, J. G. (1999, Spring). Why baseball's antitrust exemption still survives. *Marquette Sports Law Review 9*(2), 1–13.

Iaria, A. and A. Wang (2021). An empirical model of quantity discounts with large choice sets.

Jacobi, L. and M. Sovinsky (2016, August). Marijuana on main street? estimating demand in markets with limited access. *American Economic Review 106*(8), 2009–45.

Jarosch, G., J. S. Nimczik, and I. Sorkin (2019). Granular search, market structure, and wages. Working paper.

Johnson, N. L. (1949, 06). Systems of Frequency Curves Generated by Methods of Translation. *Biometrika 36*(1-2), 149–176.

Johnson, S. R. and G. C. Rausser (1971). Effects of misspecifications of log-linear functions when sample values are zero or negative. *American Journal of Agricultural Economics 53*(1), 120–124.

Kahn, L. M. (2000, September). The sports business as a labor market laboratory. *Journal of Economic Perspectives 14*(3), 75–94.

Karaca-Mandic, P. and K. Train (2003). Standard error correction in two-stage estimation with nested samples. *The Econometrics Journal 6*(2), 401–407.

Krautmann, A. C. (1999). What's wrong with the scully-estimates of player's marginal revenue product. *Economic Inquiry 37*(2), 369–381.

Lehto, E. and P. Böckerman (2008). Analysing the employment effects of mergers and acquisitions. *Journal of Economic Behavior Organization 68*(1), 112 – 124.

Levitt, S. D. and J. A. List (2009). Field experiments in economics: The past, the present, and the future. *European Economic Review 53*(1), 1–18.

Levitt, S. D., J. A. List, S. Neckermann, and D. Nelson (2016). Quantity discounts on a virtual good: The results of a massive pricing experiment at king digital entertainment. *Proceedings of the National Academy of Sciences 113*(27), 7323–7328.

Li, K. J. and S. Jain (2016). Behavior-based pricing: An analysis of the impact of peer-induced fairness. *Management Science 62*(9), 2705–2721.

Lipsius, B. (2018, November). Labor Market Concentration does not Explain the Falling Labor Share. Technical Report pli1202, Job Market Papers.

List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *The Quarterly Journal of Economics 119*(1), 49–89.

MacKinnon, J. G. and L. Magee (1990). Transforming the dependent variable in regression models. *International Economic Review*, 315–339.

MaCurdy, T. E. and J. H. Pencavel (1986). Testing between competing models of wage and employment determination in unionized markets. *Journal of Political Economy 94*(3), S3–S39.

Manning, A. (2011). Imperfect competition in the labor market. *Handbook of labor economics 4*, 973–1041.

Margolis, D. N. (2006, Aug). Should employment authorities worry about mergers and acquisitions? *Portuguese Economic Journal 5*(2), 167–194.

Marinescu, I. and H. J. Hovenkamp (2018). Anticompetitive mergers in labor markets. *Faculty Scholarship at Penn Law*.

Marinescu, I. and E. N. Posner (2019). Why has antitrust law failed workers? *Forthcoming, Cornell Law Review*.

Martins, P. S. (2018, October). Making their own weather? Estimating employer labour-market power and its wage effects. Technical Report 95, Queen Mary, University of London, School of Business and Management, Centre for Globalisation Research.

Mas, A. (2006). Pay, reference points, and police performance. *The Quarterly Journal of Economics 121*(3), 783–821.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698 – 714. The regression discontinuity design: Theory and applications.

McCrary, J. and H. Royer (2011, February). The effect of female education on fertility and infant health: Evidence from school entry policies using exact date of birth. *American Economic Review 101*(1), 158–95.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *P. Zarembka (ed.), FRONTIERS IN ECONOMETRICS. Academic Press: New York*, 105–142.

McFadden, D. and K. Train (2000). Mixed mnl models for discrete response. *Journal of applied Econometrics 15*(5), 447–470.

McManus, B. (2007, Summer). Nonlinear pricing in an oligopoly market: The case of specialty coffee. *Rand Journal of Economics 38*, 512–532.

Michalopoulos, S. and E. Papaioannou (2013). Pre-colonial ethnic institutions and contemporary african development. *Econometrica 81*(1), 113–152.

Michalopoulos, S. and E. Papaioannou (2014). National Institutions and Subnational Development in Africa. *The Quarterly Journal of Economics 129*(1), 151–213.

Miklós-Thal, J. and C. Tucker (2019). Collusion by algorithm: Does better demand prediction facilitate coordination between sellers? *Management Science 65*(4), 1552–1561.

Mitchell, T. M. (1997). *Machine Learning* (1 ed.). USA: McGraw-Hill, Inc.

Monhait, J. (2010). Baseball arbitration: An adr success. *Harvard Journal of Sports Entertainment Law*, 105–821.

Montornès, J. and J. B. Sauner Leroy (2009). Wage-setting behavior in france: additional evidence from an ad-hoc survey.

Mullahy, J. (1997). Instrumental-variable estimation of count data models: Applications to models of cigarette smoking behavior. *The Review of Economics and Statistics 79*(4), 586–593.

Mullin, M. and R. Sukthankar (2000). Complete cross-validation for nearest neighbor classifiers. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, San Francisco, CA, USA, pp. 639–646. Morgan Kaufmann Publishers Inc.

Naidu, S., E. A. Posner, and E. G. Weyl (2018a). Antritrust remedies for labor market power. *Harvard Law Review*.

Naidu, S., E. A. Posner, and G. E. Weyl (2018b, Feb). Antitrust remedies for labor market power. *Social Science Research Network (SSRN)*.

Nevo, A. (2001a). Measuring market power in the ready-to-eat cereal industry. *Econometrica 69*(2), 307–342.

Nevo, A. (2001b). Measuring market power in the ready-to-eat cereal industry. *Econometrica 69*(2), 307–342.

Orbach, B. Y. and L. Einav (2007). Uniform prices for differentiated goods: The case of the movie-theater industry. *International Review of Law and Economics 27*(2), 129–153.

Papps, K. L. (2010, Aug). Productivity under large pay increases: Evidence from professional baseball. *Social Science Research Network (SSRN)*.

Petrin, A. and K. Train (2010). A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research 47*(1), 3–13.

Pissarides, C. A. (2009). The unemployment volatility puzzle: Is wage stickiness the answer? *Econometrica 77*(5), 1339–1369.

Pons, V. and C. Tricaud (2018). Expressive voting and its cost: Evidence from runoffs with two or three candidates. *Econometrica 86*(5), 1621–1649.

Posner, E. A. (2021). *How antitrust failed workers*. Oxford University Press.

Prager, E. and M. Schmitt (2018). Employer Consolidation and Wages: Evidence from Hospitals. Technical report.

Prospectus, B. (2021). Cot's baseball contracts.

Qiu, Y. and A. Sojourner (2019, January). Labor-Market Concentration and Labor Compensation. SSRN Scholarly Paper ID 3312197, Social Science Research Network, Rochester, NY.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological) 31*(2), 350–371.

Ravallion, M. (2017). A concave log-like transformation allowing non-positive values. *Economics Letters 161*, 130 – 132.

Rinz, K. (2018). Labor Market Concentration, Earnings Inequality, and Earnings Mobility. Working Paper 2018-10, Center for Economic Studies, US Census Bureau.

Robinson, J. (1969). *The Economics of Imperfect Competition, 2nd Edition*. Palgrave Macmillan.

Rogerson, R., R. Shimer, and R. Wright (2005, December). Search-Theoretic Models of the Labor Market: A Survey. *Journal of Economic Literature 43*(4).

Rosencrance, L. (2000 (accessed October 3, 2019)). *Consumers balk at variable DVD pricing*. ComputerWorld, 34 (2000), p. 4. https://www.computerworld.com/article/2597065/customers-balk-at---variable-dvd-pricing.html.

Rossi, P. E., R. E. McCulloch, and G. M. Allenby (1996). The value of purchase history data in target marketing. *Marketing Science 15*(4), 321–340.

Rottenberg, S. (1956). The baseball players' labor market. *Journal of Political Economy 64*.

Santos Silva, J. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and Statistics 88*(4), 641–658.

Santos Silva, J. and S. Tenreyro (2010). On the existence of the maximum likelihood estimates in poisson regression. *Economics Letters 107*(2), 310 – 312.

Santos Silva, J. and S. Tenreyro (2011). poisson: Some convergence issues. *The Stata Journal 11*(2), 207 – 212.

Santos Silva, J. M. C., S. Tenreyro, and F. Windmeijer (2015). Testing competing models for non-negative data with many zeros. *Journal of Econometric Methods 4*(1), 29 – 46.

Schubert, G., A. Stansbury, and B. Taska (2020). Employer Concentration and Outside Options. pp. 150.

Scully, G. W. (1974). Pay and performance in major league baseball. *The American Economic Review 64*(6), 915–930.

Shiller, B. and J. Waldfogel (2011). Music for a song: An empirical look at uniform pricing and its alternatives. *The Journal of Industrial Economics 59*(4), 630–660.

Shiller, B. R. (2015). First degree price discrimination using big data. *Working Paper*.

Shleifer, A. and L. Summers (1988). *Breach of Trust in Hostile Takeovers*. National Bureau of Economic Research, Inc.

Siegel, D. S. and K. L. Simons (2010). Assessing the effects of mergers and acquisitions on firm performance, plant productivity, and workers : new evidence from matched employer - employee data. *Strategic Management Journal 31*(8), 903–916.

Starr, E., J. Prescott, and N. Bishara (2015, Jul). Noncompetes in the u.s. labor force. *Social Science Research Network (SSRN)*.

Statista (2021). Digital media report 2021 – video games. Technical report, Statista.

Tirole, J. (1988). *The Theory of Industrial Organisation*. The MIT Press.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica 26*(1), 24–36.

Todd, P. E. and K. I. Wolpin (2020). The best of both worlds: Combining rcts with structural modeling. *Journal of Economic Literature*.

Train, K. E. (2009). *Discrete choice methods with simulation*. Cambridge university press.

Varian, H. (1989). Price discrimination. In R. Schmalensee and R. Willig (Eds.), *Handbook of Industrial Organization* (1 ed.), Volume 1, Chapter 10, pp. 597–654. Elsevier.

Vernimmen, P., P. Quiry, M. Dallocchio, Y. L. Fur, and A. Salvi (2018). *Chapter 8: How to perform a financial analysis*. John Wiley Sons, Inc.

Wagner, S. and J. Runge (2018, September/October). Nothing is free: Data-driven optimisation unlocks freemium business models' real potential. *European Business Review*, 47–50.

Waldfogel, J. (2015). Price discrimination goes to school. *The Journal of Industrial Economics 63*(4), 569–597.

Wen, W. and F. Zhu (2019). Threat of platform-owner entry and complementor responses: Evidence from the mobile app market. *Strategic Management Journal 40*(9), 1336–1367.

Wooldridge, J. (1997). Quasi-likelihood methods for count data. *Handbook of applied econometrics 2*, 352–406.

Wooldridge, J. M. (2015). Control Function Methods in Applied Econometrics. *Journal of Human Resources 50*(2), 420–445.

Worrall, J. L. and T. V. Kovandzic (2010). Police levels and crime rates: An instrumental variables approach. *Social Science Research 39*(3), 506–516.

Yi, J., Y. Lee, and S.-H. Kim (2019). Determinants of growth and decline in mobile game diffusion. *Journal of Business Research 99*, 363–372.

Yin, P.-L., J. P. Davis, and Y. Muzyrya (2014). Entrepreneurial innovation: Killer apps in the iphone ecosystem. *American Economic Review 104*(5), 255–59.

Young, K. H. and L. Y. Young (1975). Estimation of regressions involving logarithmic transformation of zero values in the dependent variable. *The American Statistician 29*(3), 118–120.

Yuan, H. (2020). Competing for time: A study of mobile applications. Technical report, Working Paper.

ECOLE DOCTORALE

Titre: Essais sur la politique de concurrence et l' économétrie appliquée

Mots clés: Concurrence, Monopsone, Log Zéro, Poisson

Résumé: A travers des données françaises, le premier article mesure la concentration des emplois dans les firmes françaises et trouve un effet négatif de celle-ci sur les salaires et le nombre de recrutements. Une deuxième étude sur les joueurs de Baseball américains trouve que la politique antitrust permet à certains de ces joueurs d'augmenter leurs salaires d'au moins 30% en mesurant la différence de salaire entre ceux ne pouvant pas changer d'employeur et d'autres aléatoirement choisis pour profiter d'un marché du travail concurrentiel. La troisième contribution porte la capacité d'une firme à gagner en revenus à travers des prix différenciés dans un jeu vidéo pour téléphone portable. En modélisant la demande des consommateurs, nous trouvons que la firme ne peut pas exploiter l'hétérogénéité de la demande pour augmenter ses profits ; suggérant l'absence d'un besoin de régulation. Finalement, une contribution méthodologique est faite dans le contexte de l'estimation d'élasticités ou de semi-élasticités où la variable dépendante inclut un zéro.

Title: Essays on competition policy and applied econometrics

Keywords: Competition, Monopsony, Log Zero, Poisson

Abstract: The first article measures labor market concentration in France, and finds it to have a negative association with wages and new hires. The second study focuses on the wages of Baseball players and finds that antitrust laws can increase wages by at least 30%. The third article quantifies the capacity to price discriminate in an online mobile game and finds that the firm cannot exploit heterogeneity in willingness to pay for the purpose of increasing her profits. Finally, a fourth article contributes to the correct estimation of elasticities and semi-elasticities through regression analysis.

Institut Polytechnique de Paris
91120 Palaiseau, France